Speech Analysis Synthesis and Perception

Third Edition

James L. Flanagan
Jont B. Allen
Mark A. Hasegawa-Johnson

March 30, 2011

Preface to the Third Edition

The 1972 edition of *Speech Analysis, Synthesis, and Perception* defined the highest possible standards of scientific precision in the field of speech engineering, and especially in the foundational scientific disciplines upon which speech engineering is based: acoustics, audition, probability, and signal processing. Treatment of speech acoustics in the second edition continues to be unmatched by any other text. The other three foundational disciplines have moved on, leaving the 1972 edition behind. The goal of this third edition of *Speech Analysis, Synthesis, and Perception* is to teach twenty-first century acoustics, audition, probability, and signal processing with the same high standards that Flanagan applied in 1972.

James Flanagan, Warren Township, New Jersey
Jont Allen, Mahomet, Illinois
Mark Hasegawa-Johnson, Champaign, Illinois

Preface to the Second Edition

The first edition of this book has enjoyed a gratifying existence. Issued in 1965, it found its intended place as a research reference and as a graduate-level text. Research laboratories and universities reported broad use. Published reviews—some twenty-five in number—were universally kind. Subsequently the book was translated and published in Russian (Svyaz; Moscow, 1968) and Spanish (Gredos, S.A.; Madrid, 1972).

Copies of the first edition have been exhausted for several years, but demand for the material continues. At the behest of the publisher, and with the encouragement of numerous colleagues, a second edition was begun in 1970. The aim was to retain the original format, but to expand the content, especially in the areas of digital communications and computer techniques for speech signal processing. As before, the intended audience is the graduate-level engineer and physicist, but the psychophysicist, phonetician, speech scientist and linguist should find material of interest.

Preparation of the second edition could not have advanced except for discussions, suggestions and advice from many colleagues. In particular, professors and scientists who have used the book in their university lectures, both here and abroad, provided valuable comment about organization and addition of new material. Also, research colleagues, especially my associates in the Acoustics Research Department at Bell Laboratories, provided critical assessment of technical data and views about emphasis. To list individually all who influenced these factors would require inordinate space. Rather, I commend to you their many scientific contributions described among the following pages. Naturally, any shortcomings in exposition or interpretation rest solely with me.

The task of examining page proofs was shared, with notable enthusiasm, among several associates. I owe special thanks to Doctors L. R. Rabiner, R. W. Schafer, N. S. Jayant, A. E. Rosenberg, J. L. Hall, R. C. Lummis, J. M. Kelly and J. R. Haskew for this assistance. Further, I am indebted to my company, Bell Laboratories, for supporting the work and making its facilities available for typing and drafting. My secretary, Mrs. B. Masaitis, bore the brunt of this work and deserves special praise. As earlier, the efficient staff of Springer, through the organization of Dr. H. Mayer-Kaupp, shielded me from many details in actualizing the printed volume. Finally, again, to my wife and sons I express warm thanks for their contribution of weekends which might have been spent otherwise.

James Flanagan
Warren Township, New Jersey
January 15, 1972

Preface to the First Edition

This book has its origin in a letter. In November of 1959, the late Prof. Dr. Werner Meyer-Eppler wrote to me, asking if I would contribute to a series he was planning on Communication. His book "Grundlagen und Anwendungen der Informationstheorie" was to serve as the initial volume of the series.

After protracted consideration, I agreed to undertake the job provided it could be done outside my regular duties at the Bell Telephone Laboratories. Shortly afterwards, I received additional responsibilities in my research organization, and felt that I could not conveniently pursue the manuscript. Consequently, except for the preparation of a detailed outline, the writing was delayed for about a year and a half. In the interim, Professor Meyer-Eppler suffered a fatal illness, and Professors H. Wolter and W. D. Keidel assumed the editorial responsibilities for the book series.

The main body of this material was therefore written as a leisuretime project in the years 1962 and 1963. The complete draft of the manuscript was duplicated and circulated to colleagues in three parts during 1963. Valuable comments and criticisms were obtained, revisions made, and the manuscript submitted to the publisher in March of 1964. The mechanics of printing have filled the remaining time.

If the reader finds merit in the work, it will be owing in great measure to the people with whom I have had the good fortune to be associated. In earlier days at the M.I.T. Acoustics Laboratory, my association with Professor K. N. Stevens, Dr. A. S. House, and Dr. J. M. Heinz was a great privilege. During this same time, and on two separate occasions, Dr. G. Fant was a guest researcher at the M.I.T. laboratory. Later, during a summer, I had the priviledge of working as a guest in Dr. Fant's laboratory in Stockholm. On all occasions I profited from his views and opinion.

In more recent times, my associates at Bell Laboratories have been a constant stimulus and encouragement. Beginning with Dr. J. R. Pierce, under whose direction research in speech and hearing has taken on renewed vigor, Doctors E. E. David, Jr., M. R. Schroeder, M. V. Mathews, J. L. Kelly, Jr., N. Guttman, P. B. Denes, G. G. Harris, and many, many others have provided sage advice, valuable collaboration and a stimulating research atmosphere. I am certain that this collection of technical talent is duplicated at no other place in the world.

I am greatly in the debt of numerous colleagues for valuable criticism and comment of the draft material. Their appraisals have aided materially in the revisions. Besides several of those already named, Professor G. E. Peterson and Dr. H. K. Dunn, and a number of their associates at the University of Michigan, provided a wealth of valuable suggestions. Professor Osamu Fujimura of the University of Electro-Communications, Tokyo, supplied many penetrating remarks, particularly on points relating to vocal-tract acoustics. Dr. W. A. Van Bergeijk of Bell Laboratories reviewed Chapter IV in detail. Messrs. A. M. Noll, J. L. Sullivan, and H. R. Silbiger, also of the Laboratories, studied the entire manuscript and supplied numerous helpful comments.

It is with deep regret that I conclude this effort without the counsel of Professor

Meyer-Eppler. I sincerely hope that it fulfills his original concept of the volume. I wish to express my appreciation to Professor Wolter and to Professor Keidel for their continued support during the preparation. Also, the many details could not have been surmounted without the help of Dr. H. Mayer-Kaupp of Springer.

Finally, to my wife and family I express my deep appreciation for their contribution of my time.

James Flanagan
Warren Township, New Jersey
July 29, 1964

Contents

4 Speech Production 75

5 Techniques for Speech Analysis 135

9 Human Speech Recognition 275

10 Automatic Speech Recognition 319

Chapter 1

Voice Communication

> "Nature, as we often say, makes nothing in vain, and man is the only animal whom she has endowed with the gift of speech. And whereas mere voice is but an indication of pleasure or pain, and is therefore found in other animals, the power of speech is intended to set forth the expedient and inexpedient, and therefore likewise the just and the unjust. And it is a characteristic of man that he alone has any sense of good and evil, of just and unjust, and the like, and the association of living beings who have this sense makes a family and a state."
> ARISTOTLE, Politics

While there is some serious doubt today about what other animals might have the gift of speech (bees and birds included), humans seem to be unique in our ability to transmit complex information via both by voice and the written word. Of the myriad varieties of life sharing our world, only humans have developed the vocal means for coding and conveying information beyond a rudimentary stage and at a high rate. Where would be be without this amazing process of speech? It is more to our credit that we have developed the facility from apparatus designed to subserve other, more vital purposes. The true extent of human's higher order is to be determined, but by any measure it is impressive.

Because humans evolved in an atmosphere, it is not unnatural that we should learn to communicate by causing air molecules to collide. In sustaining longitudinal vibrations, the atmosphere provides a medium. At the acoustic level, speech signals consist of rapid and significantly erratic fluctuations in air pressure. These sound pressures are generated and radiated by the vocal apparatus. At a different level of coding, the same speech information is contained in the neural signals which actuate the vocal muscles and manipulate the vocal tract. Speech sounds radiated into the air are detected by the ear and apprehended by the brain. The mechanical motions of the middle and inner ear, and the electrical pulses traversing the auditory nerve, may be thought of as still different codings of the speech information.

Acoustic transmission and reception of speech works fine, but only over very limited distances. The reasons are several. At the frequencies used by the vocal tract and ear, radiated acoustic energy spreads spatially and diminishes rapidly in intensity. Even if the source could produce great amounts of acoustic power, the medium can support only limited variations in pressure without distorting the signal. The sensitivity of the receiver—the ear—is limited by the acoustic noise of the environment and by the physiological noises of the body. The acoustic wave is not, therefore, a good means for distant transmission.

Through the ages men have striven to communicate at distances. They are, in fact, still striving. The ancient Greeks are known to have used intricate systems of signal fires which they placed on judiciously selected mountains for relaying messages between cities. One enterprising Greek, Aeneas Tacitus by name, is credited with a substantial improvement upon the discrete bonfire message. He placed water-filled earthen jars at the signal points. A rod, notched along its length and supported on a cork float, protruded from each jar. At the first signal light, water was started draining from the jar. At the second it was stopped. The notch on the rod at that level represented a previously agreed upon message. (In terms of present day information theory, the system must have had an annoyingly low channel capacity, and an irritatingly high equivocation and vulnerability to jamming!)

History records other efforts to overcome the disadvantages of acoustic transmission. In the sixth century B.C., Cyrus the Great of Persia is supposed to have established lines of signal towers on high hilltops, radiating in several directions from his capital. On these vantage points he stationed leather-lunged men who shouted messages along, one to the other. Similar "voice towers" reportedly were used by Julius Caesar in Gaul. (Anyone who has played the party game of vocally transmitting a story from one person to another around a circle of guests cannot help but reflect upon the corruption which a message must have suffered after several miles of such transmission.)

Despite the desires and motivations to accomplish communication at distances, it was not until humans learned to generate, control and convey electrical current, that telephony could be brought within the realm of possibility. As history goes, this has been exceedingly recent.

Telephony did not develop as an isolated advance. Rather, it occurred in the context of evolving understanding of electrical phenomena. It built upon fundamental knowledge established by early researchers, such as Oersted, Faraday, Ampere, Maxwell, Heaviside,[1] and many others. The nineteenth century was a remarkable time for invention, application, innovation and newbusiness creation, and communication over ever greater distances was a constant goal. In particular, electromagnetic behavior was key to the introduction of telegraphy—one of the earliest means of electrical communication over distances. Telegraph communication emerged as a

---

[1] Heaviside was profoundly influenced by Maxwell's work, and his uncle, Sir Charles Wheatstone (1802-1875), was the a co-inventor of the telegraph c1830." (paraphrased from Wikipedia, 10/12/08).

commercial endeavor in 1844.

A leading contributor to telegraphy was Samuel Finley Breese Morse, a successful portrait painter who turned inventor in mid-age, and who exploited on-off transmission of electrical pulses over conducting wires. His original receiver was an electromechanical stylus to mark long (dash) and short (dot) pulses on a paper tape. Unique sequences of dots and dashes indicated letters of text. Morse and his associate (Vail) are credited with the Morse code, whose characters are based upon the distribution of type pieces in a printer's box, the most probable letter being assigned the briefest code character (for example, the letter ”e” being one dot). It was soon observed that the ‘click/clack’ sound of the electromagnet receiver could be apprehended auditorily, and the mechanical indentation of tape was supplanted by a ”sounder” and human hearing. A skilled telegrapher could send and receive text in the order of 20 words per minute. Speech is nearly three times this speed (i.e., ≈ 1 wd/s).

Morse recognized that the serial resistance and shunt capacitance of a wire-pair attenuated the transmitted pulses to a level too feeble to actuate the receiver. He consequently conceived the electrical relay (whereby a weak current could control a stronger one) which permitted exact, noise-free regeneration of the signal before it became unusable or contaminated by noise. (This concept of regeneration of on-off (binary) signals is central to digital transmission today, and makes received quality independent of distance.) Continued improvements in transmission enabled transcontinental telegraphy in 1861. And, the ambition for greater range spurred efforts to span the Atlantic ocean with undersea cable. After initial failures, a practical cable was established in 1866. But, the great (unrepeatered) length provided very narrow bandwidth and telegraphic transmission at only about two words per minute. While quite slow, this rate still bettered the five days required for a ocean voyage. The later invention of inductive loading and improvements in automatic send/receive extended into the twentieth century, eventually providing about 400 words per minute. Even so, the transmission bandwidth remained limited to several hundred hertz (a future stimulus for ”Vocoder” techniques.)

While telegraphy developed rapidly and served widely, communication was slow and the conversion and re-conversion of text in terms of dots and dashes required operational skill not found in the typical customer. Faster communication by end-to-end users, requiring no specialized training, focused continued interest on the natural mode of information exchange—human speech.

Quantify the following statements:

Little more than a hundred years have passed since the first practical telephone was put into operation (c1880); there are now more telephones than people on planet Earth.

Many early inventors and scientists labored on electrical telephones and laid foundations which facilitated the development of commercial telephony. Their biographies make interesting and humbling reading for today's communication engineer comfortably ensconced in a well equipped laboratory.

Mention AGB's father?

Among the pioneers, A. G. Bell was somewhat unique for his background in physiology and phonetics. His comprehension of the mechanisms of speech and hearing was crucial to his electrical experimentations. Similar understanding is equally important to today's telephone researcher. It was perhaps his training that influenced Bell—according to his assistant Watson—to summarize the telephony problem by saying "If I could make a current of electricity vary in intensity precisely as the air varies in density during the production of a speech sound, I should be able to transmit speech telegraphically." And this is what he accomplished, culminating in his famous patent of March 3, 1876, perhaps one of the most commercially-valuable in history.

Bell's basic notion—namely, preservation of acoustic waveform—clearly proved to be an effective means for speech transmission, and has supported voice communication for more than a century. Improvements on Bell's first implementations followed rapidly. In particular, Thomas Edison in 1878 used compressed lamp black to create the variable-resistance carbon button microphone (transmitter), an immediate and major advance over the original stylus/liquid transducer. Its robustness and power amplifying properties have served telephone systems for more than a hundred years, and is only now being supplanted by the electro-statically biased, solid-dielectric element (electret). Waveform coding was the most widely used form of telephony until approximately the year 2000, when the number of digital cellular telephones outnumbered the analog handsets.

Although this "waveform preserving principle" is exceedingly satisfactory and has long endured, it is not the most efficient means for voice transmission. As we shall see, digital telephony does not preserve the waveform, in that that only perceptually significant speech components are preserved.

Communication engineers have recognized for many years that a substantial mismatch exists between the information capacity of human perception and the capacity of the "waveform" channel. Specifically, the channel is capable of transmitting information at rates much higher than those the human can assimilate. As an example, if one listens to modem modulation used to send data over the acoustic channel, it sounds like noise. This is because it is formated in a way that cannot be effectively decoded by the human ear.

Modern developments in communication theory have established techniques for quantifying the information in a signal and the rate at which information can be signalled over a given channel. These analytical tools have accentuated the desirability of matching the transmission channel to the information source. From their application, conventional telephony has become a much-used example of disparate source rate and channel capacity. This disparity—expressed in numbers—has provided much of the impetus toward investigating more efficient means for speech coding and for reducing the bandwidth and channel capacity used to transmit speech.

Figure 1.1: Conversation over lunch: Renoir's *Luncheon of the Boating Party*, 1881. (Phillips Collection, Washington D.C.)

1.1 Speech as a Communication Channel

We speak to establish social bonds, and to develop complex trains of thought. The natural environment for speaking is noisy and complicated, with a continuously changing visual and auditory channel, as depicted, for example, in Fig. 1.1. In this famous painting, friends relax on a Sunday afternoon at the restaurant *Maison Fournaise*. The image provides examples of many different kinds of conversations: flirtations, expositions, relaxed subdued conversations, and even a conversation between a woman (Aline Charigot, who would later marry Renoir) and her dog.

Before (and during) speaking, every talker conceives a message: a sequence of words, possibly annotated with subtle hints of nuance and opinion (Levelt [1989]). Unlike mona Lisa's famous smile, a spoken message is 'symbolic,' meaning a sequence of "discrete" well defined symbols. In most cases, however, we find it pleasant to encode the message in an analog medium, by an expression or by configuring the speech articulators (the lips, jaw, tongue, soft palate, larynx, and lungs), in order to generate an acoustic waveform. Each listener "hears" the acoustic signal with their cochlea, or inner ear, which converts the sound into a neural code in the auditory nerve, which then passes through a series of tuned circuits, designed to detect the speech units. Eventually the listener decodes the intended symbolic linguistic message.

The subject of this book is the encoding and decoding of the messages conveyed by speech: the digital-to-analog and analog-to-digital transformations used by humans and machines to produce and understand ordinary conversation. Before considering the analog channel in more detail, however, it's worthwhile to evaluate the end-to-end channel performance.

The mathematical theory of information (Shannon and Weaver [1949]) provides a useful mechanism for analyzing the end-to-end performance of any communications channel, independent of the details of its implementation. Figure 1.2 shows the

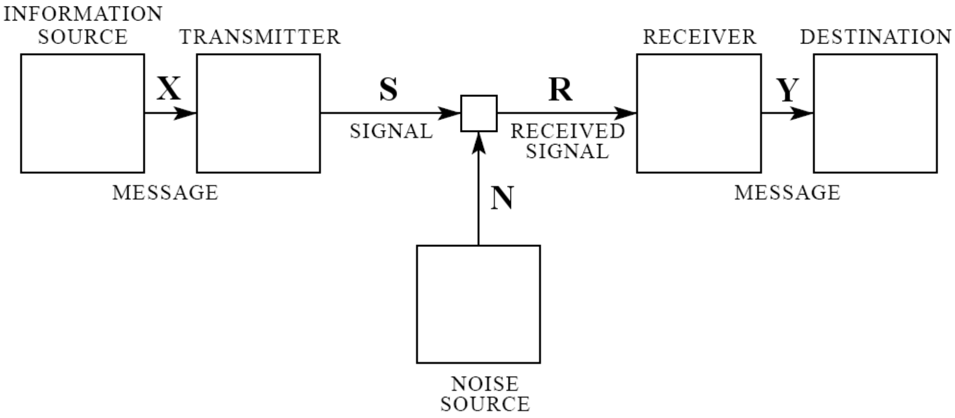


Figure 1.2: Schematic diagram of a Shannon's channel (communication model). The basic definitions are in terms of the siz boxes: X =source message, Y =received message, S =transmitted signal, R =received signal, N =noise (After Shannon and Weaver, 1949). Here it is assumed that the message X is a discrete sequence of symbols, whereas the signal S is an analog message, to which noise N is added, producing the signal R which is received and decoded by the receiver to produce the symbolic received message Y.

schematic of an abstract communication channel. There are six boxes in this figure. The boxes marked "information source" and "noise source" each draw a message or a noise signal, at random, from some probability distribution.

The definition of the ensemble of messages are the unique sequences of symbols. Each message has an entropy $\mathcal{H}$ defined as the average number of possible messages $N = 2^{\mathcal{H}}$, typically measured in *bits*. The purpose of using discrete symbols is to greatly reduce the channel error rate. The goal of the box marked "transmitter" is to encode the message, and that of the "receiver" is to decode the message, so that the received message will be as similar as possible to the transmitted message. As we shall see, the average information rate of any speech source is remarkably low. There are apparently two reasons for the low information rate of speech. First, there is evidence that human listeners are unable to process information at a rate much higher than that of the speech message; in this respect, humans appear to be less effective than machines. Second, low information rate allows speech transmission over extremely noisy acoustic channels. Human listeners (but not machines) are able to correctly understand meaningful linguistic messages transmitted at signal to noise ratios (SNR) as low as -20dB; in this respect, making humans appear to be more effective than machines. The low information rate of speech, and its remarkable noise robustness, are best understood as an adaptation to noisy natural environments like the outdoor lunch sweet cacophony exampled in Fig. 1.1.

1.2 Entropy

Shannon's theory provided a paradygm shift [Brillouin, 1962]. His elementary relations of information theory define the information associated with the selection of a discrete message from a specified ensemble. If the messages of the set are x_i, are independent, and have probability of occurrence $P(x_i)$, the partial, or *specific information* (i.e., information density), associated with a selection is $\mathcal{I}_i = \log_2 (1/P(x_i))$ bits.[2] The average information associated with selections from the set is the ensemble average

$$\mathcal{H}(X) \equiv \mathcal{E}\{\mathcal{I}\} = -\sum_i P(x_i) \log_2 P(x_i)$$

bits, or the source entropy. The log() transformation converts reciprocal probability into entropy, and with a base of 2, having units of *bits.*

A first important concept here is information being reciprocal probability; thus a rare event (small probability) is defined as having large information. The concept of probability, and thus of information, requires a *set of outcomes.* The vector of probabilities $[p_n]$ requires an index n labeling N possible outcomes, while element p_n measures the relative frequency (parts of a whole) of these outcomes, which obey the condition $\sum_n p_n = 1$.

Entropy is the average amount of information, as computed by taking a weighted average of the information density. When all the outcomes are equal (i.e., $p_n = 1/N$) the entropy $\mathcal{H}$ is maximum, and the information is minimum Cover and Thomas [1991]. The maximum in $\mathcal{H}(p_1, p_2, \cdots, p_N)$ is a measure of the maximum *size* of the message space, which numerically is $N = 2^{\mathcal{H}}$.

Discuss the conservation of probability, and the resulting variational principle (minimazation) used when defning entropy.

1.2.1 Entropy of the Speech Source

As an example consider a phonemic transcription of speech; that is, the written equivalent of the meaningfully distinctive sounds of speech. Take English for example. Table 1.2 shows a list of 42 English phones, including vowels, diphthongs and consonants, and their relative frequencies of occurrence in prose (Dewey [1923]). If the phonemes are selected for utterance with equal probability [i.e., $P(x_i) = 1/42$] the average information per phoneme would be approximately $H(X) = 5.4$ bits. If the phonemes are selected independently, but with probabilities equal to the relative frequencies shown in Table 1.2, then $H(X)$ falls to 4.9 bits. The sequential constraints imposed upon the selection of speech sounds by a given language reduce this average information still further.[3] In conversational speech about 10 phonemes are uttered

---

[2] The base-2 logarithm is used to compute information in bits. A base-10 logarithm computes information in "digits;" a natural logarithm computes information in "nats."

[3] Make this an exercize: Related data exist for the letters of printed English. Conditional constraints imposed by the language are likewise evident here. If the 26 English letters are considered equiprobable, the average information per letter is 4.7 bits. If the relative frequencies of the letters

Table 1.1: LDC unvoiced consonants, voiced consonants and vowels. The DARPA-bet (Dbet) and *International phonetic alphabet* (IPA) symbols are in the second and third columns. Dbet sympols use standard simple roman characters to represent standard english IPA symbols.

Unvoiced Consonants			**Voiced Consonants**			**Vowels**		
Example	Dbet	IPA	Example	Dbet	IPA	Example	Dbet	IPA
/ch/urch	C	ʧ	/th/is	D	ð	h/a/t	@	æ
/sh/e	S	ʃ	/b/ee	b	b	h/u/t	A	ʌ
/th/ink	T	θ	/d/og	d	d	b/e/t	E	ɛ
/f/ish	f	f	/g/ab	g	g	h/i/t	I	ɪ
/h/e	h	h	/j/udge	J	ʤ	b/oy/	O	ɔɪ
/c/at	k	k	/l/ook	l	l	b/ir/d	R	ɝ
/p/en	p	p	/m/an	m	m	p/u/t	U	ʊ
/s/ee	s	s	/n/ap	n	n	h/ow/	W	aʊ
ca/t/	t	t	/r/eal	r	r	wh/y/	Y	aɪ
			plea/s/ure	Z	ʒ	c/au/ght	c	ɔ
			si/ng/	G	ŋ	c/o/t	a	ɑ
			/v/ow	v	v	b/ai/t	e	e / eɪ
			/w/in	w	w	b/ee/	i	i
			/y/ou	y	j	b/oa/t	o	o / oʊ
			/z/oo	z	z	b/oo/	u	u
						th/e/		ə

per second. The written equivalent of the information generated is therefore less than 50 bits/sec.

1.3 Conditional Entropy and context

Because of noise, the speech signal arriving at the receiver may be different from the signal generated by the transmitter. If the decoding algorithm is not sufficiently robust, noise in the acoustic signal may lead to errors in the received message. Per-

---

are used as estimates of $P(x_i)$, the average information per letter is 4.1 bits. If digram frequencies are considered, the information per letter, when the previous letter is known, is 3.6 bits. Taking account of trigram frequencies lowers this figure to 3.3 bits. By a limit-taking procedure, the long range statistical effects can be estimated. For sequences up to 100 letters in literary English the average information per letter is estimated to be on the order of one bit. This figure suggests a redundancy of about 75 per cent. If statistical effects extending over longer "meaning" units such as paragraphs or chapters are considered, the redundancy will be still higher, leading to more robust information transfer due to the greater context [Shannon and Weaver, 1949].

Table 1.2: Relative frequencies of English speech sounds in standard prose, and the information content of each symbol $\mathcal{I}(x) = \log_2(1/p(x))$ (Data from Dewey, 1923).

Vowels and diphthongs			Consonants		
IPA	% relative frequence	$\mathcal{I}$	IPA	% relative frequence	$\mathcal{I}$
ɪ	8.53	3.55	n	7.24	3.79
a	4.63	4.43	t	7.13	3.81
æ	3.95	4.66	r	6.88	3.86
ɛ	3.44	4.86	s	4.55	4.46
ɒ	2.81	5.15	d	4.31	4.54
ʌ	2.33	5.42	l	3.74	4.74
i	2.12	5.56	ð	3.43	0.1669
e, eɪ	1.84	5.76	z	2.97	5.07
u	1.60	5.97	m	2.78	5.17
aɪ	1.59	5.97	k	2.71	5.21
oʊ	1.30	6.27	v	2.28	5.45
ɔ	1.26	6.31	w	2.08	5.59
ʊ	0.69	7.18	p	2.04	5.62
aʊ	0.59	7.41	f	1.84	5.76
ɑ	0.49	7.67	h	1.81	5.79
o	0.33	8.24	b	1.81	6.7
ju	0.31	8.33	ŋ	0.96	6.7
ɔɪ	0.09	10.12	ʃ	0.82	6.93
			ɡ	0.74	7.08
			j	0.60	7.38
			tʃ	0.52	7.59
			dʒ	0.44	7.83
			θ	0.37	8.08
			ʒ	0.05	10.97
Totals	38			62	

$H(X) = -\sum_i P(x_i)\log_2 P(x_i) = 4.9$ bits. If all phonemes were equiprobable, then $\mathcal{H}_{\max}(X) = \log_2 42 = 5.4$ bits, and the *maximum entropy* condition would be achieved.

ceptual errors can be characterized by the conditional probability that the receiver decodes *heard* symbol $h \in X$, given that the transmitter encoded *spoken* symbol $s \in Y$. This probability may be writtten as $P_{h|s}(SNR)$, in order to emphasize that it is also a function of several channel characteristics, including the encoding system used by the transmitter and receiver (A), the bandwidth of the channel (B), and the SNR ($\gamma = S/N$, where S is the power of the signal coming out of the transmitter, and N is the power of the noise signal). For example, an error-free communication system is characterized by the conditional probability distribution

$$P_{h|s}(SNR) = \delta_{ij} \equiv \begin{cases} 1 & h = s \\ 0 & \text{otherwise} \end{cases} \tag{1.1}$$

If $P_{h|s}(SNR) \neq \delta_{ij}$, then one may say that the communication system is increasing the entropy of the received signal. This is an undesirable behavior, because the entropy generated by the communication channel is independent of the information generated at the source, and is equivalent to a noise. The average rate at which the communication channel introduces errors into a transmitted signal is called the *equivocation* or *conditional entropy* of Y given X, and is defined to be

$$\mathcal{H}_{h|s}(SNR) \equiv \mathcal{E}\{\mathcal{I}(h|s)\} = -\sum_s \sum_h P_{h|s}(SNR) \log_2 P_{h|s}(SNR) \tag{1.2}$$

$$= -\sum_s P_s(SNR) \sum_h P_{h|s}(SNR) \log_2 P_{h|s}(SNR)$$

The average amount of information successfully transmitted over the channel is equal to the information rate of the source, $\mathcal{H}(X)$, minus the rate at which errors are introduced by the channel, $\mathcal{H}_{Y|X}(SNR)$. This rate is called the *mutual information* between the transmitted message and the received message:

$$\mathcal{I}_{s,h}(SNR) = \mathcal{H}_h - \mathcal{H}_{h|s}(SNR) \tag{1.3}$$

$$= \sum_s \sum_h P_s P_{h|s}(SNR) \left(\frac{P_{h|s}(SNR)}{P_s} \right)$$

Make this proof a assignment.

Human speech production is a coding algorithm, and may be evaluated just like any other coding algorithm: by computing the mutual information $\mathcal{I}_{AB\gamma}$ that it achieves over any particular acoustic channel. Fletcher 1922 found that, for SNRs of at least 30dB, phones in maximum entropy (e.g., nonsense) syllables are perceived correctly about 98.5% of the time, corresponding to an equivocation of roughly[4]

$$\mathcal{H}(Y|X) \approx 0.985\log_2(1/0.985) + 0.015\log_2(1/0.015) = 0.11 \quad \text{[bits/symbol]}. \quad (1.4)$$

In order to force listeners to make perceptual errors, Fletcher was forced to distort the acoustic channel by introducing additive noise and/or linear filtering (lowpass, highpass, or bandpass filters applied to the acoustic channel).

Eq. (1.4) is only an approximation of the speech channel equivocation: in order to calculate the equivocation exactly, it is necessary to know the probability $P_{AB\gamma}(y_j|x_i)$ for every (i, j) combination. Miller and Nicely 1955 measured conditional probability tables under fifteen different channel conditions for a subset of the English language: specifically, for the subset $x_i \in \{$p,b,t,d,k,g,f,v,θ,ð,s,z,ʃ,ʒ,m,n$\}$, and y_j drawn from the same set. Each consonant was produced in a consonant vowel (CV) syllable, and the vowel was always /ɑ/. In order to cause perceptual errors, Miller and Nicely limited the bandwidth of the acoustic channel (9 conditions), or the SNR (5 conditions). After several thousand trials, the perceptual effect of each channel was summarized in the form of a *confusion matrix*, like the one shown in Fig. 1.3. In an *articulation count matrix*, entry $\mathcal{C}(s, h)$ lists the number of times that spoken phoneme s was heard as phoneme h. The conditional probability $P_{h|s}$ may be estimated by taking a row frequency (count) and normalizing it by the row-sum of the count

$$P_{h|s} \approx \frac{C(s,h)}{\sum_h C(s,h)} \quad (1.5)$$

Representations of the confusion matrix (CM): Figure 1.3 shows a typical MN55 consonant-vowel (CV) *confusion matrix* or *count matrix* CM for wideband speech (0.2-6.5 kHz), at a *speech-to-noise ratio* (SNR) of -6 dB [Miller and Nicely, 1955, Table III]. The 16 consonants were presented along with the vowel /a/ as in father (i.e., the first three sounds were [/pa/, /ta/, /ka/]). After hearing one of the 16 CV sounds as labeled by the first column, the consonant that was reported is given as labeled along the top row. This array of numbers form the basic CM, denoted $\mathcal{C}_{s,h}$, where integer indices s and h (i.e., "spoken" and "heard") each run between 1 and 16. For example, /pa/ was spoken 230 times (the sum of the counts in the first row), and was reported heard 80 times ($\mathcal{C}_{1,1}$), while /ta/ was reported 43 times ($\mathcal{C}_{1,2}$). For Table III the mean row count was 250, with a standard deviation of 21 counts.

---

[4] HSJ: we need to make the nature of this approx clear. Why not start from the actual formula rather than assume the probs are uniform? This approx never really holds (probs are not uniform, as shown in the table of $\mathcal{I}$. This approximation results from the assumption that only two events matter: the phone is either correctly or incorrectly recognized. The actual equivocation of a 42-phone communication system with a 1.5% error rate could be anywhere between 0.02 and 0.19 bits/symbol, depending on the error rates of each individual phoneme, and the distribution of errors across the various possible substitutions.

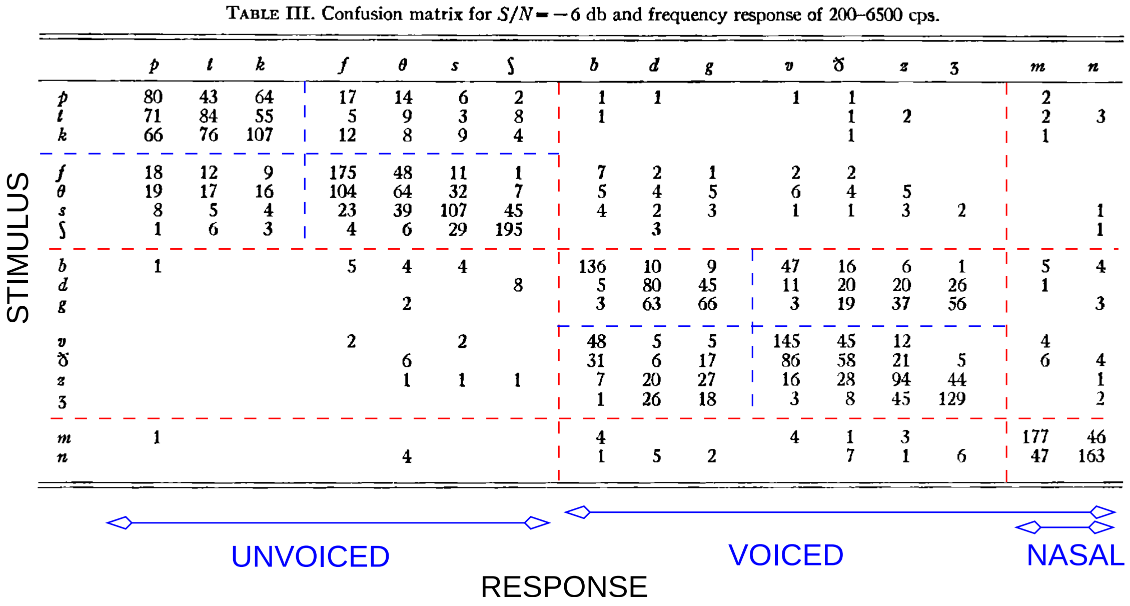


TABLE III. Confusion matrix for $S/N = -6$ db and frequency response of 200–6500 cps.

	p	*t*	*k*	*f*	*θ*	*s*	*ʃ*	*b*	*d*	*g*	*v*	*ð*	*z*	*ʒ*	*m*	*n*
p	80	43	64	17	14	6	2	1	1		1	1			2	
t	71	84	55	5	9	3	8	1				1	2		2	3
k	66	76	107	12	8	9	4					1			1	
f	18	12	9	175	48	11	1	7	2	1	2	2				
θ	19	17	16	104	64	32	7	5	4	5	6	4	5			
s	8	5	4	23	39	107	45	4	2	3	1	1	3	2		1
ʃ	1	6	3	4	6	29	195		3							1
b	1			5	4	4		136	10	9	47	16	6	1	5	4
d							8	5	80	45	11	20	20	26	1	
g					2			3	63	66	3	19	37	56		3
v				2		2		48	5	5	145	45	12		4	
ð					6			31	6	17	86	58	21	5	6	4
z					1	1	1	7	20	27	16	28	94	44		1
ʒ								1	26	18	3	8	45	129		2
m	1							4			4	1	3		177	46
n					4			1	5	2		7	1	6	47	163

Figure 1.3: *Typical Miller-Nicely confusion (or count) matrix (CM) $\mathcal{C}$, from Table III at -6dB SNR, 6.3 kHz bandwidth. Each entry in the matrix $\mathcal{C}_{s,h} \equiv P_c(h|s)$ is the subject response count. The rows correspond to the* spoken *CVs, each row representing a different consonant, from $s = 1, \ldots, 16$. The columns correspond to the* heard *CVs, each column representing a different consonant, from $h = 1, \ldots, 16$. The common vowel /a/, as in father, was used throughout. When the 16 consonants are ordered as shown, the count matrix shows a "block-symmetric" partitioning in the consonant confusions. In this matrix there are three main blocks delineated by the dashed lines, corresponding to UNVOICED, VOICED and NASAL. Within the VOICED and UNVOICED subgroups, there are two additional symmetric blocks, corresponding to AFFRICATION and DURATION, also delineated with dashed lines. The presentation probabilities were uniform in this experiment, not the natural probabilities, given by Table 1.2. fix ref to table*

When the sounds are ordered as shown in Fig. 1.3, they form groups, identified in terms of hierarchical clusters of *articulatory features*. For example, the first group of sounds 1-7 correspond to UNVOICED, group 8-14 are VOICED, and [15,16] are NASAL and VOICED.

At an SNR of -6 dB, the intra-confusions (within a group) are much greater than the inter-confusions (between groups). For example, members of the group 1-7 (the UNVOICED sounds) are much more likely to be confused among themselves, than between the NONNASAL-VOICED sounds (8-14), or the NASAL sounds (15,16). The NASAL are confused with each other, but rarely with any of the other sounds 1-14. Two of the groups [sounds (1-7) and (8-14)] form sub-clusters.

Using the approximation in Eq. 1.5, the equivocation of the speech communication system, at -6 dB SNR, is 2.176 bits. Since each syllable is chosen uniformly from $2^4 = 16$ possible syllables, the source entropy is $H(X) = \log_2 16 = 4$ bits. The amount

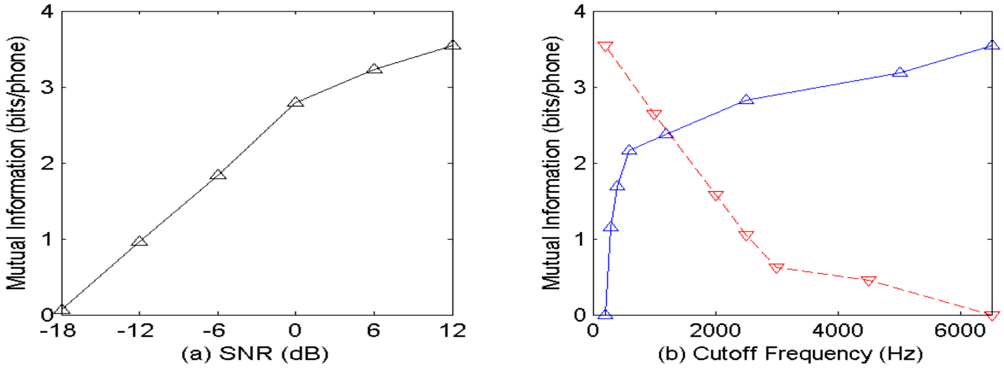


Figure 1.4: (a) Mutual information between spoken and perceived consonant labels, as a function of SNR, over an acoustic channel with 6.3 [kHz] bandwidth (0.2-6.5 [kHz]). (b) Mutual information between spoken and perceived consonant labels, at 12dB SNR, over lowpass and highpass acoustic channels with the specified cutoff frequencies. The lowpass channel contains information between 0.2 [kHz] and the cutoff; bit rate is shown with a solid line. The highpass channel contains information between the cutoff and 6.5 [kHz]; bit rate is shown with a dashed line. (After Miller and Nicely, 1955)

of information successfully transmitted from talker to listener, therefore, is $4-2.176 = 1.834$ bits. Figure 1.4(a) shows the information transmitted from talker to listener, over the wideband acoustic channel, as a function of SNR. Mutual information is greater than one bit per consonant at -12dB, and the information rate only drops to zero below -18dB SNR. Figure 1.4(b) shows the information transmitted over the lowpass and highpass filtered channels, as a function of the cutoff frequency.

1.4 Capacity of the Acoustic Channel

Mutual information is a summary of the efficiency with which algorithm A transmits information over a channel with bandwidth B and noise statistics N. Shannon has demonstrated 1949 that no algorithm can transmit more information than

$$\mathcal{I}(X,Y) \leq \mathcal{C}\left(B, \frac{S}{N}\right), \tag{1.6}$$

where B is the bandwidth of the channel, S/N is the power signal to noise ratio, and $\mathcal{C}(B, S/N)$ is called the *channel capacity*. Shannon has shown that the channel capacity of a channel with additive Gaussian noise is given by

$$C(B,S/N) = \int_0^B \log_2\left(\frac{\sigma_{s+n}^2(f)}{\sigma_s^2(f)}\right) df \approx \int_0^B \log_2\left(1 + \frac{\sigma_s^2(f)}{N(f)}\right) df \quad \left[\frac{\text{bits}}{\text{second}}\right] \tag{1.7}$$

where $\sigma_s^2(f)$ and $\sigma_n^2(f)$ are the power spectra of the speech and noise, respectively. Speech is transmitted over an acoustic channel with bandwidths varying between about 3.0 [kHz] (telephone transmission) to 20kHz (the audible frequency range, usable during face-to-face communication). Under very noisy listening conditions (e.g., at an SNR of -12dB or $\sigma_s^2/\sigma_n^2 = 1/16$), the capacity of a telephone-band acoustic channel is 188 bits/second–far greater than the information transmitted from a human talker to a human listener. In a quiet room (at an SNR of about 30 [dB], or $\sigma_s^2/\sigma_n^2 \approx 1000$), the channel capacity of a 20 [kHz] channel is 20 [kbits/second] – 400 times greater than the information rate achieved by a human conversationalist.

Why is speech limited to a rate of 50 bits/second? Phrased another way: why don't people talk more quickly under quiet listening conditions, or more clearly, in order to communicate at a bit rate higher than 50 bps? Is the extra information already present, in the form of subtle nuances of intonation? Is the time waveform simply an inefficient code, incapable of carrying more than 50bps? Is the human incapable of processing information at rates much higher than 50 bits/sec? Does the receiver discard much of the transmitted information? Chapter 9 will consider these questions in much greater detail; for now, let us consider some experimental studies that have tried to answer this question.

A number of experimental efforts have been made to assess the informational capacity of human listeners. The experiments necessarily concern specific, idealized perceptual tasks. In most cases it is difficult to generalize or to extrapolate the results to more complex and applied communication tasks. Even so, the results do provide quantitative indications which might reasonably be taken as order-of-magnitude estimates for human communication in general.

In one response task, for example, subjects were required to echo verbally, as fast as possible, stimuli presented visually (Licklider *et al.* [1954]). The stimuli consisted of random sequences of binary digits, decimal digits, letters and words. The maximal rates achieved in this processing of information were on the order of 30 bits/sec. When the response mode was changed to manual pointing, the rate fell to about 15 bits/sec.

The same study considered the possibility for increasing the rate by using more than a single response mode, namely, by permitting manual and vocal responses. For this two-channel processing, the total rate was found to be approximately the sum of the rates for the individual response modes, namely about 45 bits/sec. In the experience of the authors this was a record figure for the unambiguous transmission of information through a human channel.

Another experiment required subjects to read lists of common monosyllables aloud (Pierce and Karlin [1957]). Highest rates attained in these tests were 42 to 43 bits/sec. It was found that prose could be read faster than randomized lists of words. The limitation on the rate of reading was therefore concluded to be mental rather than muscular. When the task was changed to reading and tracking simultaneously, the rates decreased.

A different experiment measured the amount of information subjects could assimilate from audible tones coded in several stimulus dimensions (Pollack and Ficks [1954]). The coding was in terms of tone frequency, loudness, interruption rate, spatial direction of source, total duration of presentation and ratio of on-off time. In this task subjects were found capable of processing 5.3 bits per stimulus presentation. Because presentation times varied, with some as great as 17 sec, it is not possible to deduce rates from these data.

A later experiment attempted to determine the rate at which binaural auditory information could be processed (Webster [1961]). Listeners were required to make binary discriminations in several dimensions: specifically, vowel sound; sex of speaker; ear in which heard; and, rising or falling inflection. In this task, the best subject could receive correctly just under 6 bits/sec. Group performance was a little less than this figure.

As indicated earlier, these measures are determined according to particular tasks and criteria of performance. They consequently have significance only within the scopes of the experiments. Whether the figures are representative of the rates at which humans can perceive and apprehend speech can only be conjectured. Probably they are. None of the experiments show the human to be capable of processing information at rates greater than the order of 50 bits/sec.

Assuming this figure does in fact represent a rough upper limit to man's ability to ingest information, he might allot his capacity in various ways. For example, if a speaker were rapidly uttering random equiprobable phones, a listener might require all of his processing ability to receive correctly the written equivalent of the distinctive speech sounds. Little capacity might remain for perceiving other features of the speech such as stress, inflection, nasality, timing and other attributes of the particular voice. On the other hand, if the speech were idle social conversation, with far-reaching statistical constraints and high redundancy, the listener could direct more of his capacity to analyzing personal characteristics and articulatory peculiarities.

1.4.1 Human speech recognition

In order to succinctly summarize our knowledge of human speech recognition (HSR), it is important to develop a model. Our model is presented in Fig. 1.5 which shows the structural relations between the various quantitative probabilistic measures of recognition. In the model, *all* of the recognition errors in HSR are a result of event extraction labeling errors, as depicted by the second box, and modeled by the articulation-band errors e_k. In other words, *sound recognition errors are modeled as a noise in the event conversion from analog to discrete "objects"*. I will argue that much of this front-end event processing is implemented as *parallel processing*, which is equivalent to assuming that the recognition of events is independent, on average, across cochlear frequency bands.

As shown in Fig. 1.5, the input speech signal is continuous while the output stream

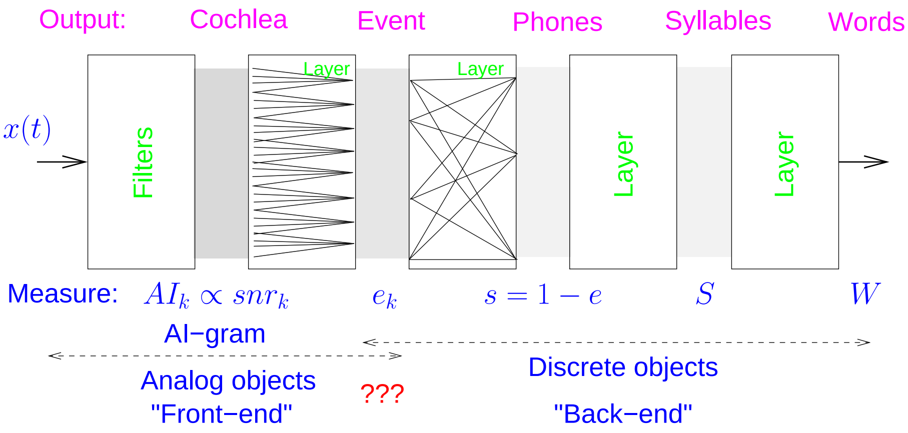


Figure 1.5: Model block diagram summary of speech recognition by humans. At the top of each block is a label that attempts to identify the physical operation, or a unit being recognized. The labels below the boxes indicate the probability measure defined at that level. See the text for the discussion of objects, at the very bottom. The speech $s(t)$ enters on the left and is processed by the cochlea (first block), breaking the signal into a filtered continuum of band-passed responses. The output of the cochlea is characterized by the specific articulation index measure AI_k, a normalized SNR, expressed in dB units Allen [1994], Regnier and Allen [2008], Li and Allen [2008]. The second box represents the work of the early auditory brain, which is responsible for the identification of events in the speech signal, such as onset transients and the detection of basic measures. The third block puts these basic features together defining phones. The remaining blocks account for context processing.

is discrete. Somewhere within the auditory brain discrete decisions must be made. A critical aspect of our understanding is to identify at what point and at what level this conversion from continuous to discrete takes place. I will argue that this conversion is early, at the event level. Once these decisions have been made, the processing is modeled as a *noiseless state machine* (i.e., a state machine having no stochastic elements).

When testing either HSR or ASR systems, *it is critical to control for language context effects*. This was one of the first lessons learned by Fletcher *et al.*, that context is a powerful effect, since the score is strongly affected by context.

The HSR model of Fig. 1.5 is a "bottom–up," divide and conquer strategy. Humans recognize speech based on a hierarchy of context layers. Humans have an intrinsic robustness to noise and filtering. In fact, the experimental evidence suggests that this *robustness* does not seem to interact with semantic context (language), as reflected by the absences of feedback in the model block diagram.

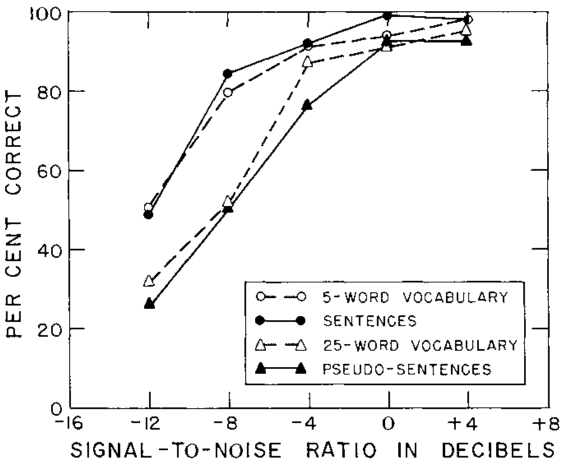


Figure 1.6: This figure, from Miller (1962), summarizes the results of a 4-way experiment, performed as a function of the signal to noise ratio. Test 1 (open circles, dashed line) shows $P_c(SNR)$ for 5 word vocabularies, with no context. In test 2 (closed circles, solid line) 5 word sentences were made from the 5, 5 word lists. As an example "Don brought his black socks." The word "Don" was one of the 5 possibilities [Don, He, Red, Slim, Who]. For tests 1 and 2, $P_c(SNR)$ is the same. Test 3 (open triangles, dashed line) was to test using the larger corpus of one of the 25 words, spoken in isolation. Test 4 (closed triangles, solid line) was to generate "pseudo-sentences" by reversing the order of the sentences of test 3. Going from 5 to 25 isolated words (test 1 to 3) causes a 4 dB SNR reduction in performance at the 50% correct level. Presenting the 25 words as pseudo-sentences, that make no sense (test 4), has no effect on $P_c(SNR)$. However adding a grammar (test 2) to a 25 word test returns the score to the 5 word test. In summary, increasing the test size from 5 to 25 words reduces performance by 4 dB. Making 5 word grammatically correct sentences out of the 25 words restores the performance to the 5 word low entropy case.

1.4.2 Context

A detailed example of the utility of context in HSR was demonstrated by Miller [1962] This example stands out because of the early use of ideas from information theory to control for the entropy of the source, with the goal of modulating human performance via context. The experiment was simple, yet it provides an insight into the workings of context in HSR.

In this experiment 5 groups of 5 words each make up the test set. This is a *closed–set*[5] listening task with the number of words and the signal to noise ratio varied. There are 4 conditions. For test condition 1 the subjects are shown 1 of the 5 lists, and they hear a word from that list. For the other 3 conditions the subjects

---

[5] A *closed–set* test is one with a limited number of outcomes that are known *a priori* to the subjects.

are shown 1 list of all the 25 words. The probability correct $P_c(SNR)$ was measured for each of the 4 conditions:

1. 5 words
2. 5 word grammatically correct sentences, chosen from the 25 words
3. 25 words
4. non grammatical sentences chosen from the 25 words.

As described in the caption of Fig. 1.6, in condition (1) 5 word lists are used in each block of trials. The lists are randomized. The subject hears 1 of 5 words, degraded by noise, and is asked to pick the word from the list. In condition (3) the number of words is increased from 5 to 25, causing a reduction of 4 dB in performance (at the 50% level). These two conditions (1 and 3) were previously studied in a classic paper Miller *et al.* [1951] which observed that the size of the set of CVCs has a large impact on the score, namely $P_c(SNR, \mathcal{H})$ depends on the entropy $\mathcal{H}$ of the task. In condition (2), the effect of a grammar context is measured. By placing the 25 words in a context having a grammar, the scores returned to the 5 isolated word level (condition 1). When sentences having no grammar (pseudo-sentences) were used (condition 4), generated by reversing the meaningful sentences of condition 2, the score remains equal to the 25 isolated word case of condition 3.

Thus the grammar in experiment (2) improves the score to the isolated word level (1), but not beyond. It probably does this by providing an improved framework for remembering the words. Without the grammatical framework, the subjects become confused and treat the pseudo-sentences as 25 random words Miller and Isard [1963].

1.5 Organization of this Book

The goal of this book is to teach the science and technology of speech analysis, synthesis, and perception. The book is loosely divided into a "science" half and a "technology" half. The science and technology are unified by an information-theoretic view of speech communication, based on the theory and terminology developed by Shannon.

The first half of the book (chapters 1-5) addresses the science of speech communication. The science of speech, in our view, is the study of the speech behaviors of human beings, and includes a mathematically sophisticated treatment of ideas from both physics and psychology. Like all other communication channels, the speech communication channel is best studied by methodically elucidating the characteristics of the message, the transmitter, the receiver, and the channel. Chapter 2 describes the characteristics of the message: the alphabet of phones and suprasegmental speech gestures, and the probabilistic rules that govern their combination. Chapter 3 describes the speech transmitter, with a particular emphasis on the physical acoustic principles of speech production. Chapter 4 describes the speech receiver, including the results of both physiological and psychological experiments studying the transductive processes of the ear. Finally, chapter 5 describes characteristics of the channel and the receiver that relate to the perception and understanding of speech.

The second half of the book (chapters 6-9) describes technological methods that have been used to analyze, replace or augment each component of the speech communication system. Chapter 6 describes fundamental signal analysis methods that are common to the algorithms of all succeeding chapters. Once the reader has understood chapter 6, the remainder of the book need not be read in order; each of chapters 7-9 may be studied independently as a self-contained introduction to the technology it describes. Chapter 7 describes algorithms that replace the speech transmitter by converting a text message into a natural-sounding acoustic speech signal. Chapter 8 describes algorithms that replace the speech receiver, in the sense that they automatically convert an acoustic speech signal into a written sequence of phonemes or words. Finally, chapter 9 describes algorithms that replace the acoustic channel with a low-bit-rate digital channel, for purposes of secure, cellular, or internet telephony. All three of these areas are the subjects of active ongoing research; the goal of this book is to present fundamental concepts and derivations underlying the most effective solutions available today.

Problem set: What is the information rate of a sinewave, or a square wave? How about a sine wave with a 1 Hz bandwidth modulation?

Chapter 2

The Mechanism of Speech Production

2.1 Physiology of the Vocal Apparatus

Speech is the acoustic end product of voluntary, formalized motions of the respiratory and masticatory apparatus. It is a motor behavior which must be learned. It is developed, controlled and maintained by the acoustic feedback of the hearing mechanism and by the kinesthetic feedback of the speech musculature. Information from these senses is organized and coordinated by the central nervous system and used to direct the speech function. Impairment of either control mechanism usually degrades the performance of the vocal apparatus[1].

The speech apparatus also subserves the more fundamental processes of breathing and eating. It has been conjectured that speech evolved when ancient peoples discovered that they could supplement their communicative hand signals with related "gestures" of the vocal tract. Sir Richard Paget sums up this speculation quite neatly. "What drove man to the invention of speech was, as I imagine, not so much the need of expressing his thoughts (for that might have been done quite satisfactorily by bodily gesture) as the difficulty of 'talking with his hands full.' It was the continual use of man's hands for craftsmanship, the chase, and the beginnings of art and agriculture, that drove him to find other methods of expressing his ideas–namely, by a specialized pantomime of the tongue and lips (Paget [1930])."

The machinery involved in speech production is shown schematically in Fig. Fig. 2.1. The diagram represents a mid-sagittal section through the vocal tract of an adult. The primary function of inhalation is accomplished by expanding the rib cage, reducing the air pressure in the lungs, and drawing air into the lungs via nostrils, nasal cavity, velum port and trachea (windpipe). Air is normally expelled by the same

---

[1] Most of us are aware of the difficulties that partially or totally deaf persons have in producing adequate speech. Even more familiar, perhaps, are the temporary difficulties in articulation experienced after the dentist desensitizes a large mouth area by an injection of anesthetic.

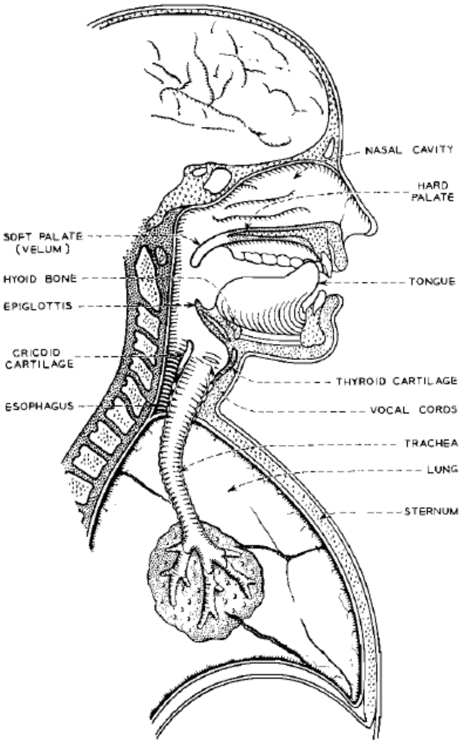


Figure 2.1: Schematic diagram of the human vocal mechanism

route. In eating, mastication takes place in the oral cavity. When food is swallowed the structures at the entrance to the trachea are drawn up under the epiglottis. The latter shields the opening at the vocal cords and prevents food from going into the windpipe. The esophagus, which normally lies collapsed against the back wall of the throat, is at the same time drawn open to provide a passage to the stomach.

The vocal tract proper is an acoustical tube which is nonuniform in cross-sectional area. It is terminated by the lips at one end and by the vocal cord constriction at the top of the trachea at the other end. In an adult male the vocal tube is about 17cm long and is deformed in crosssectional area by movement of the articulators; namely, the lips, jaw, longue and velum. The cross-sectional area in the forward portion of the tract can be varied from zero (i.e. complete closure) to upwards of 20 cc.

The nasal tract constitutes an ancillary path for sound transmission. It begins at the velum and terminates at the nostrils. In the adult male the cavity has a length of about 12 cm and a volume on the order of 60 cc, It is partitioned over part of its front-to-back extent by the nasal septum. Acoustic coupling between the nasal and vocal tracts is controlled by the size of the opening at the velum. In Fig. Fig. 2.1 the velum is shown widely open. In such a case, sound may be radiated from both the mouth and nostrils. In general, nasal coupling can substantially influence the character of sound radiated from the mouth. For the production of non-nasal sounds the velum is drawn tightly up and effectively seals off the entrance to the nasal cavity. In an adult male the area of the velar opening can range from zero to around 5 cc.

The source of energy for speech production lies in the thoracic and abdominal musculatures. Air is drawn into the lungs by enlarging the chest cavity and lowering the diaphragm. It is expelled by contracting the rib cage and increasing the lung pressure. Production of vowel sounds at the softest possible level requires a lung pressure of the order of 4 cm H_20. For very loud, high-pitched sounds, on the other hand, pressures of about 20 cm H_20 or more are not uncommon. During speaking the lung pressure is maintained by a steady, slow contraction of the rib cage.

As air is forced from the lungs it passes through the trachea into the pharynx, or throat cavity. The top of the trachea is surmounted by a structure which is shown in additional detail in Fig. Fig. 2.2. This is the larynx. The cartilaginous frame houses two lips of ligament and muscle. These are the vocal cords and are denoted VC. The slit-like orifice between the cords is called the glottis. The knobby structures, protruding upward posterior to the cords, are the arytenoid cartilages, and are labelled AC. These cartilages support the fleshy cords and facilitate adjustment of tension. The principal outside cartilages of the larynx “box” are the anterior thyroid (labelled TC in Fig. Fig. 2.2) and the posterior cricoid. Both of these can be identified in Fig. Fig. 2.1.

The voiced sounds of speech are produced by vibratory action of the vocal cords. Production of sound in this manner is called phonation. Qualitatively, the action of the vocal folds is very similar to the flapping of a flag, or the vibration of the reed in a woodwind instrument. Like a flag flapping in the wind, the vocal folds must have

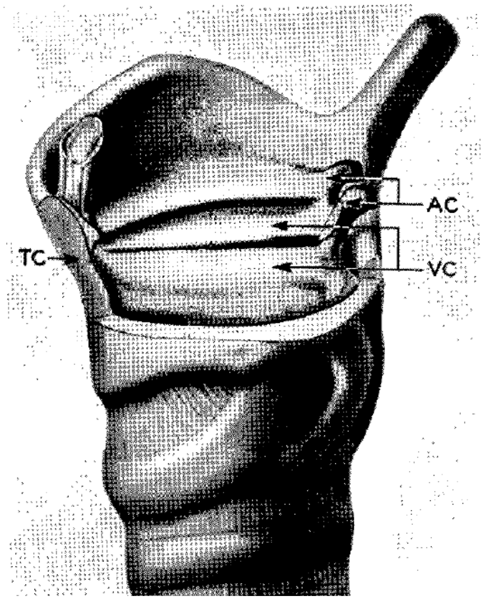


Figure 2.2: Cut-away view of the human larynx. (After Farnsworth.) VC-vocal cords; AC-arytenoid cartilages; TC-thyroid cartilage

at least two regions that are out of phase with one another. Like the jet of air passing over the surface of a flag, the jet of air passing through the glottis has two regimes: a laminar regime, and a turbulent regime. In the laminar regime, Bernoulli's equation holds, so air pressure is inversely proportional to the square of air jet velocity. In the turbulent regime, differences in velocity are absorbed by the creation of vortices, so that air pressure remains low and constant throughout the turbulent regime. The glottis flaps from bottom to top: the lower vocal folds separate first, followed by the upper folds. While the lower folds are wider than the upper folds, air flow within the glottis is laminar, and therefore the pressure within the glottis is high, driving the folds open. When the upper folds flap open to a position wider than the lower folds, air within the glottis becomes turbulent, and therefore the pressure within the glottis drops to a low constant value. At this point, the stiffness of the vocal folds forces them back together again, and the cycle repeats. Notice that it is not necessary for the vocal folds to completely close at any point in the cycle. The "breathy voice" employed to great effect by some singers and actresses is apparently a form of phonation in which the glottis never completely closes. The mass and compliance of the cords, and the subglottal pressure, essentially determine the period of the oscillation. This period is generally shorter than the natural period of the cords; that is, the cords are driven in a forced oscillation.

The variable area orifice produced by the vibrating cords permits quasi-periodic pulses of air to excite the acoustic system above the vocal cords. The mechanism is somewhat similar to blowing a tone on a brass instrument, where the vibrating lips

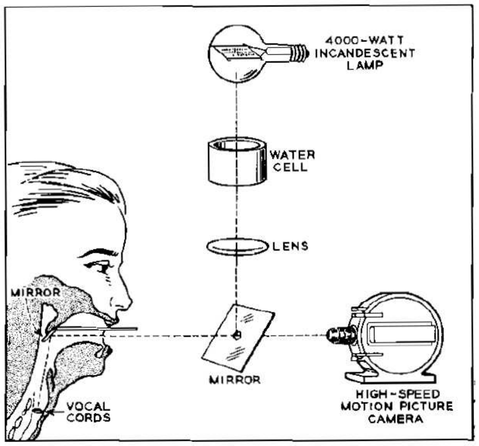


Figure 2.3: Technique for high-speed motion picture photography of the vocal cords. (After Farnsworth)

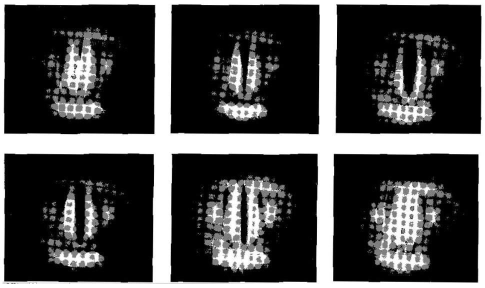


Figure 2.4: Successive phases in one cycle of vocal cord vibration. The total elapsed time is approximately 8 msec

permit quasiperiodic pulses of air to excite the resonances of the flared horn. Over the past years the vibratory action of the vocal cords has been studied in considerable detail. Direct observations can be made by positioning a 45-degree mirror toward the back of the mouth, near the naso-pharynx. Stroboscopic illumination at the proper frequency slows or "stops" the vibratory pattern and permits detailed scrutiny.

Still more revealing and more informative is the technique of high-speed photography, pioneered by Farnsworth(Farnsworth [1940]), in which moving pictures are taken at a rate of 4000 frames/sec, or higher. The technique is illustrated in Fig. Fig. 2.3. The cords are illuminated by an intense light source via the arrangement of lenses and mirrors shown in the diagram. Photographs are taken through an aperture in the large front mirror to avoid obstructing the illumination. The result of such photography is illustrated in Fig. Fig. 2.4. The figure shows six selected frames in one cycle of vibration of the cords of an adult male. In this case the fundamental frequency of vibration, or voice "pitch," is 125Hz.

The volume flow of air through the glottis as a function of time is similar to (though not exactly proportional to) the area of the glottal opening. For a normal voice effort and pitch, the waveform can be roughly triangular in shape and exhibit duty factors (i.e., ratios of open time to total period) commonly of the order of 0.3 to 0.7. The glottal volume current therefore has a frequency spectrum relatively rich in overtones or harmonics. Because of the approximately triangular waveform, the higher frequency components diminish in amplitude at about 12db/octave.

The waveform of the glottal volume flow for a given individual can vary widely. In particular, it depends upon sound pitch and intensity. For low-intensity, low-pitched sounds, the subglottal pressure is low, the vocal cord duty factor high, and the amplitude of volume flow low. For high-intensity, high-pitched sounds, the subglottal pressure is large, the duty factor small and the amplitude of volume flow great. The amplitude of lateral displacement of the vocal cords, and hence the maximum glottal area, is correlated with voice intensity to a surprisingly small extent (Fletcher(Fletcher [1950])). For an adult male, common peak values of glottal area are of the order of 15 mm^2.

Because of its relatively small opening, the acoustic impedance of the glottal source is generally large compared to the acoustic impedance looking into the vocal tract, at least when the tractis not tightly constricted. Under these conditions changes in tract configuration have relatively small (but not negligible) influence upon the glottal volume flow. For tight constriction of the tract, the acoustic interaction between the tract and the vocal-cord oscillator can be pronounced.

Another source of vocal excitation is produced by a turbulent flow of air created at some point of stricture in the tract. An acoustic noise is thereby generated and provides an incoherent excitation for the vocal system. The unvoiced continuant sounds are formed from this source. Indirect measurements and theory suggest that the spectrum of the noise, at its point or region of generation, is relatively broad and uniform. The vocal cavities forward of the constriction usually are the most influential in spectrally shaping the sound.

A third source of excitation is created by a pressure buildup at some point of closure. An abrupt release of the pressure provides a transient excitation of the vocal tract. To a crude approximation the aperiodic excitation is a step function of pressure, and might be considered to huvc a spectrum which falls inversely with frequency. The closure can he effected at various positions toward the front of the tract; for example, at labial, dental, and palatal positions. The transient excitation can be used with or without vocal cord vibration to produce voiced or unvoiced plosive sounds.

Whispered speech is produced by substituting a noise source for the normally vibrating vocal cords. The source may by produced by turbulent flow at the partially closed glottis, or at some other constricted place in the tract.

2.2 The Sounds of Speech

To be a practicable medium for the transmission of information, a language must consist of a finite number of distinguishable, mutuallyexclusive sounds. That is, the language must be constructed of basic linguistic units which have the property that if one replaces another in an utterance, the meaning is changed. The acoustic manifestations of a basic unit may vary widely. All such variations, however–when heard by a listener skilled in the language–signify the same linguistic element. This basic linguistic element is called a phoneme (Bloch and Trager [1942]). Its manifold acoustic variations are called allophones.

The phonemes might therefore be looked upon as a code uniquely related to the articulatory gestures of a given language. The allophones of a given phoneme might be considered representative of the acoustic freedom permissible in specifying a code symbol. This freedom is not only dependent upon the phoneme, but also upon its position in an utterance.

The set of code symbols used in speech, and their statistical properties, depend upon the language and dialect of the communicators. When a linguist initially studies an unknown language, his first step is to make a phonetic transcription in which every perceptually-distinct sound is given a symbol. He then attempts to relate this transcription to behavior, and to determine which acoustically-distinguishable sounds belong to the same phoneme. That is, he groups together those sounds which are not distinct from each other in meaning. The sounds of each group differ in pronunciation, but this difference is not important to meaning. Their difference is merely a convention of the spoken language.

Features of speech which may be phonemically distinct in one language may not be phonemic in another. For example, in many East Asian and Western African languages, changing the pitch of a vowel changes the meaning of the word. In European and Middle Eastern languages, this generally is not the case. Other striking examples are the Bantu languages of southern Africa, such as Zulu, in which tongue clicks and lip smacks are phonemes.

The preceding implications are that speech is, in some sense, discrete. Yet an oscillographic representation of the sound pressure wave emanating from a speaker producing connected speech shows surprisingly few gaps or pause intervals. Connected speech is coupled with a near continuous motion of the vocal apparatus from sound to sound. This motion involves changes in the configuration of the vocal tract as well as in its modes of excitation. In continuous articulation the vocal tract dwells only momentarily in a state appropriate to a given phoneme.

The statistical constraints of the language greatly influence the precision with which a phoneme needs to be articulated. In some cases it is merely sufficient to make a vocal gesture in the direction of the normal configuration to signal the phoneme. Too, the relations between speech sounds and vocal motions are far from unique, although normal speakers operate with gross similarity. Notable examples of the “many–valuedness” of speech production are the compensatory articulation of ventriloquists and the mimicry of parrots and myna birds.

Despite the mutability of the vocal apparatus in connected speech, and the continuous nature of the speech wave, humans can subjectively segment speech into phonemes. Phoneticians are able to make written transcriptions of connected speech events, and phonetic alphabets have been devised for the purpose. It has been argued that the concept of a phonetic alphabet was invented only once in human history, by the Phoenicians of Lebanon in the early first millenium B.C., but the uniqueness of this invention is obscured by the rapidity with which it was adopted worldwide. By 300 B.C., the Indus river scholar Panini had organized the phonemes of his language into a rank-three array, with dimensions specifying the manner of articulation (vowel, glide, nasal, fricative, stop), place of articulation (lips, teeth, alveolar ridge, hard palate, soft palate, uvula, pharynx), and glottal features (voiced vs. unvoiced, aspirated vs. unaspirated). Panini’s organization remains the foundation of all modern phonetic alphabets, including the international standard alphabet developed by the International Phonetic Association (IPA). The international phonetic alphabet (also abbreviated IPA: the meaning of the acronym is usually apparent from context) provides symbols for representing the speech sounds of most of the major languages of the world.

Linguists transcribe speech at several different levels of precision. As specified previously, two phonemes are different only if it is possible to change the meaning of a word by interchanging the two. A transcription in terms of phonemes is called “phonemic,” and is conventionally enclosed in virgules // (Fairbanks [1940]). On the other hand, the IPA provides notation for many subtle acoustic distinctions that are never used, in any given language, to change the meaning of a word; a transcription that specifies any of these allophonic or sub-phonemic distinctions is called “phonetic,” and is conventionally enclosed in brackets []. In the remainder of this book, most transcriptions will be phonemic, but we will occsionally also make use of phonetic transcription.

Classification of speech sounds is customarily accomplished according to their manner and place of production. Phoneticans have found this method convenient to indicate the gross characteristics of sounds. For example, the articulation of vowel sounds is generally described by the position of the tongue hump along the vocal tract (which is often, but not always, the place of greatest constriction) and the degree of the constriction. This classification method will be employed in the following discussion of speech sounds. The examples extend to the sounds of English speech of General American (GA) dialect.

2.2.1 Vowels

Vowels are speech sounds with no narrow constriction in the vocal tract. They are usually voiced (produced with vocal fold excitation), though they may of course

Table 2.1: Vowels

Degree of constriction	Tongue hump position					
	front		central		back	
High	/i/	eve	/ɜ^r/	bird	/u/	boot
	/ɪ/	it	/ə^r/	lover (unstressed)	/ʊ/	foot
Medium	/e/	hate*	/ʌ/	up	/o/	obey*
	/ɛ/	met	/ə/	ado (unstressed)	/ɔ/	all
Low	/æ/	at			/ɑ/	father

*These two sounds usually exist as diphthongs in GA dialect. They are included in the vowel table because they form the nuclei of related diphthongs. See Section 2.27 for further discussion. (See also (Lehiste and Peterson [1961]).)

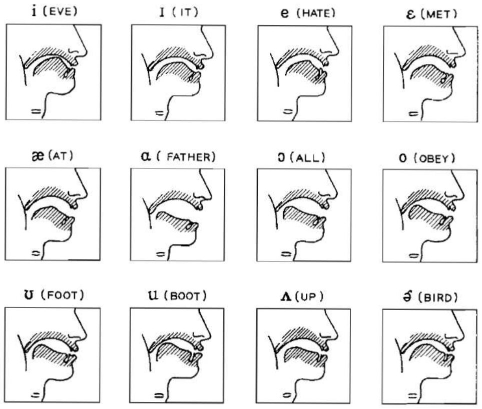


Figure 2.5: Schematic vocal tract profiles for the production of English vowels. (Adapted from Potter, Kopp and Green)

be whispered. In normal articulation, the tract is maintained in a relatively stable configuration during most of the sound. The vowels are further characterized by negligible (if any) nasal coupling, and by radiation only from the mouth (excepting that which passes through the cavity walls). If the nasal tract is coupled to the vocal tract during the production of a vowel, the vowel becomes nasalized. The distinction between nasalized and non-nasalized versions of any particular vowel is phonemic in some languages (e.g., French), but not in English; thus, for example, some comedians produce an entertaining effect by nasalizing all of their vowels.

When the 12 vowels of GA speech are classified according to the tongue-hump-position degree-of-constriction scheme, they may be arranged as shown in Table 2.1. Along with each vowel is shown a key word containing the vowel.

The approximate articulatory configurations for the production of these sounds

Table 2.2: All consonants may be divided into four broad manner classes, using the two binary features [sonorant] and [continuant]. The opposite of "sonorant" is "obstruent," sometimes denoted [−sonorant] the opposite of "continuant" is "discontinuant," sometimes denoted [−continuant]

	Continuant [+continuant]	Discontinuant [−continuant]
Sonorant [+sonorant]	Glides (/w,j/) and Semivowels (/l,r/)	Nasals (/m,n,ŋ/)
Obstruent [−sonorant]	Fricatives (/f,v,θ,ð,s,z,ʃ,ʒ,h/)	Stops (/p,b,t,d,k,g/) and Affricates (/tʃ,dʒ/)

(exclusive of the two unstressed vowels) are shown qualitatively by the vocal tract profiles in Fig. Fig. 2.5 (Potter *et al.* [1947]). The physiological basis for the front-back/high-low classification is particularly well illustrated if the profiles for the vowels /i,æ,a,u/ are compared[2].

2.2.2 Consonants

Consonants are sounds produced with a constriction at some point in the vocal tract. Consonants may be further divided into four classes, based on two binary manner features: the feature **sonorant**, and the feature **continuant**.

The literal meaning of the word "sonorant" is "song-like." A sonorant consonant is a consonant with no increase of air pressure inside the vocal tract, either because the vocal tract constriction is not very tight (/w,j,r,l/), or because the soft palate is opened, allowing air to escape through the nose (/m,n,ŋ/). Because there is no increase in air pressure, the voicing of a sonorant consonant is free and easy, and, for example, it is possible to sing a sonorant consonant.

A discontinuant consonant is produced with a complete closure at some point in the vocal tract. Because of this complete closure, the transition between a discontinuant consonant and its neighboring vowel is always marked by a sudden acoustic discontinuity, when the sound quality changes dramatically in a space of one or two milliseconds. A continuant consonant has no complete vocal tract closure.

Based on these two binary features, it is possible to divide all consonants into four broad manner classes, as shown in Table 2.2.

---

[2] These profiles, and the ones shown subsequently in this chapter, mainly illustrate the oral cavity. The important pharynx cavity and the lower vocal tract are not drawn. Their shapes may be deduced from x-rays (see Figs. Fig. 5.34 through Fig. 5.36, for example).

Fricative Consonants

Fricatives are produced from an incoherent noise excitation of the vocal tract. The noise is generated by turbulent air flow at some point of constriction. In order for the air flow through a constriction to produce turbulence, the Reynold's number $R_{ee} = ud\rho/\mu$ must be larger than 1800, where u is the air particle velocity, ρ and μ are the density and viscosity of air, and d is the smallest cross-sectional width of the constriction (all expressed in consistent units, so that the Reynolds number itself is dimensionless). Since velocity is inversely proportional to the area of the constriction, small constrictions lead to high Reynolds numbers; the threshold Reynolds number for turbulence is usually reached by by constrictions of less than about 3mm width. In order to produce a fricative, a talker must position the tongue or lips to create a constriction with a width of 2-3mm, and allow air pressure to build up behind the constriction, so that the air flow through the constriction is turbulent. If the constriction is too wide, it will not produce turbulence; if it is too narrow, it will stop the air flow entirely. Because of the precise articulation required, fricatives are rarely the first phonemes acquired by infants learning to speak.

Common constrictions for producing fricative consonants are those formed by the tongue behind the teeth (dental: /θ,ð/), the upper teeth on the lower lip (labiodental: /f,v/), the tongue to the gum ridge (alveolar: /s,z/), the tongue against the hard palate (palatal: /ʃ,ʒ/), and the vocal cords constricted and fixed (glottal: /h/). Radiation of fricatives normally occurs from the mouth. If the vocal cord source operates in conjunction with the noise source, the fricative is a voiced fricative. If only the noise source is used, the fricative is unvoiced.

Both voiced and unvoiced fricatives are continuant sounds. Because a given fricative articulatory configuration can be excited either with or without voicing, the voiced and voiceless fricatives form complementary pairs called cognates. The fricative consonants of the GA dialect are listed in Table 2.3, along with typical "places" of articulation and key words for pronunciation.

Vocal tract profiles for these sounds are shown in Fig. 2.6. Those diagrams in which the vocal cords are indicated by two small lines are the voiced fricatives. The vocal cords are shown dashed for the glottal fricative (/h/).

The phoneme /h/ is a special case because, like the sonorant consonants, it requires no increase of air pressure within the vocal tract. For this reason, some phoneticians class /h/ as a glide rather than a fricative (e.g., (Stevens [1999])). In inter-vocalic context (e.g., in the word "ahead"), the acoustic quality of /h/ may be very sonorant-like, e.g., the amplitude of voicing may not decrease at all. In other contexts, /h/ may have weakened voicing (like a typical voiced fricative), or it may be completely unvoiced (like a typical unvoiced fricative). All of these different allophones are produced and perceived interchangeably, by native speakers of English, as examples of the same underlying phoneme /h/.

Table 2.3: Fricative consonants

Place of articulation	Voiced		Voiceless	
Labio-dental	/v/	vote	/f/	for
Dental	/ð/	then	/θ/	thin
Alveolar	/z/	zoo	/s/	see
Palatal	/ʒ/	azure	/ʃ/	she
Glottal			/h/	he

Table 2.4: Stop consonants

Place of articulation	Voiced		Voiceless	
Labial	/b/	be	/p/	pay
Alveolar	/d/	day	/t/	to
Palatal/velar	/g/	go	/k/	key

Stop Consonants

Among those consonants which depend upon vocal tract dynamics for their creation are the stop consonants. To produce these sounds a complete closure is formed at some point in the vocal tract. The lungs build up pressure behind this occlusion, and the pressure is suddenly released by an abrupt motion of the articulators. Stops are distinguished from other phonemes by complete closure, followed by a characteristic acoustic "explosion" called a "transient." The transient typically lasts only one or two milliseconds, but it may be followed by a fricative burst of 5-10ms in duration, if the lips or tongue pass too slowly through the 2-3mm frication region.

Stops in English come in voiced/unvoiced cognate pairs, as do fricatives. Cognate pairs are distinguished in two ways. First, the vocal folds may continue to vibrate during the closure interval of a voiced stop. Closure voicing is often heard in carefully produced speech, but rarely in casual speech. Instead, most speakers of GA English signal that a stop is voiced by allowing the vocal folds to begin vibrating immediately after stop release. An unvoiced stop, by contrast, has a period of aspiration following release, during which the vocal folds are held open and turbulence is produced at the glottis. The acoustic effect is exactly what one would achieve by producing an unvoiced stop followed immediately by an /h/.

The cognate pairs of stops, with typical places of articulation, are shown in Table 2.4. Articulatory profiles for these sounds are shown in Fig. Fig. 2.7. Each position is that just prior to the pressure release.

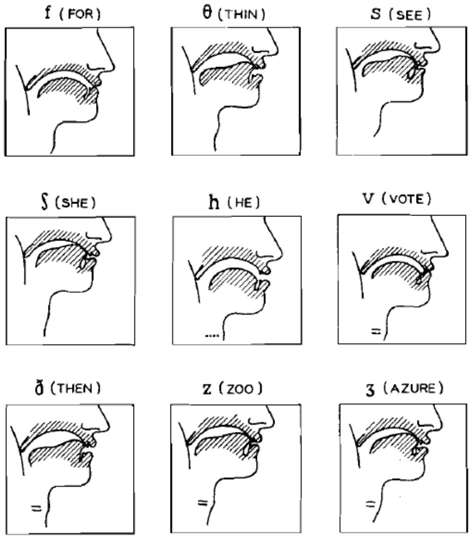


Figure 2.6: Vocal tract profiles for the fricative consonants of English. The short pairs of lines drawn on the throat represent vocal cord operation. (Adapted from Potter, Kopp and Green)

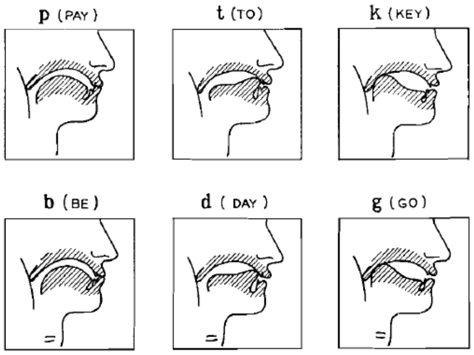


Figure 2.7: Articulatory profiles for the English stop consonants. (After Potter, Kopp and Green)

Table 2.5: Nasals

Place		
Labial	/m/	me
Alveolar	/n/	no
Palatal/velar	/ŋ/	sing (no initial form)

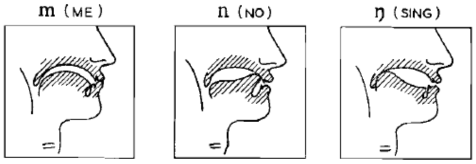


Figure 2.8: Vocal profiles for the nasal consonants. (After Potter, Kopp and Green)

Nasal Consonants

The nasal consonants, or nasals, are sonorant consonants; they are normally voiced in GA English, although unvoiced allophones might be heard in some contexts (e.g., some speakers will devoice the /n/ in “fishnet”). A complete closure is made toward the front of the vocal tract, either by the lips, by the tongue at the gum ridge, or by the tongue at the hard or soft palate. The velum is opened wide and the nasal tract provides the main sound transmission channel. Most of the sound radiation takes place at the nostrils. The closed oral cavity functions as a side branch resonator coupled to the main path, and it can substantially influence the sound radiated. Because there is no increase of the air pressure in the mouth, nasals are classed as sonorant consonants; because there is a complete closure within the vocal tract, they are discontinuant. The GA nasal consonants are listed in Table 2.5, and their vocal profiles are illustrated in Fig. Fig. 2.8.

Glides and Semivowels

Two small groups of consonants contain sounds that greatly resemble vowels. These are the glides /w,j/ and the semivowels /r,l/ (Fairbanks [1940]). Both are characterized by sonorant voicing, no effective nasal coupling, and sound radiation from the mouth. All four phonemes may be optionally devoiced (as in “which” or “rheum”); speakers of GA English usually consider voiced and devoiced allophones to be examples of the same underlying phoneme.

The glides /w/ and /j/, respectively, may be interpreted as extreme examples of the vowels /u/ and /i/—the former involves an extreme lip constriction, the latter an extreme palatal tongue constriction. In both cases, the constriction is a dynamic one, released gradually into the following vowel.

The semivowels, by contrast, may be produced either dynamically or in a rela-

Table 2.6: Glides and semi-vowels

Place		
Palatal	/j/	you
Labial	/w/	we (no final form)
Palatal	/r/	read
Alveolar	/l/	let

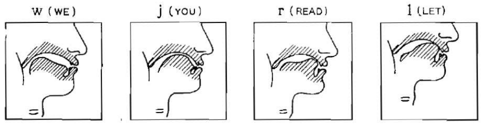


Figure 2.9: Vocal tract configurations for the beginning positions of the glides and semivowels. (After Potter, Kopp and Green)

tively static configuration; in fact, either of these two consonants may be produced as the nucleus of a syllable in English (e.g., in the words “bird” and “bull”; when produced as a syllable nuclei, these phonemes may be transcribed as /ɜʳ/ and /l̩/, respectively). Both sounds are most reliably identified by a unique acoustic pattern: /r/ is the only sound in English with a third formant below 2000Hz, and /l/ is one of the few sounds in English with a third formant above 3000Hz. Both sounds are typically produced in syllable-initial position with both a tongue body constriction and a tongue tip constriction; in syllable-final position, both sounds are optionally produced with only a tongue body constriction. The tongue tip constriction for /r/ is curled back (“retroflex”), and the tongue body constriction is tightly bunched in the middle of the hard palate. The tongue tip constriction for /l/ is made with the tip touching the gum ridge like a /d/, but open on the left and/or right (“lateral”); the tongue body constriction is near the uvula. These are the only retroflex and lateral phonemes in English, but other languages have other phonemes (in some cases, stops and fricatives) with similar tongue tip positions.

The glides and semivowels for the GA dialect are listed, according to place of articulation, in Table 2.6. Their profiles, for the beginning positions, are given in Fig. Fig. 2.9.

Combination Sounds: Diphthongs and Affricates

Some of the preceding vowel or consonant elements can be combined to form basic sounds whose phonetic values depend upon vocal tract motion. An appropriate pair of vowels, so combined, form a diphthong. The diphthong is vowel-like in nature, but is characterized by change from one vowel position to another. For example, if the vocal tract is changed from the /e/ position to the /ɪ/ position, the diphthong /eɪ/ as in say is formed. Other GA diphthongs are /ɪu/ as in new, /ɔɪ/ as in boy; /ɑʊ/ as in out, /ɑɪ/ as in I, and /oʊ/ as in go.

As vowel combinations form the diphthongs, stop-fricative combinations likewise create the two GA affricates. These are the /tʃ/ as in chew and the /dʒ/ as in jar.

Each of these combination sounds is perceived to be a phoneme by typical speakers of GA English. For example, in games where subjects are asked to reverse the order of phonemes in a word (turning "scram" into "marks," for example), an affricate or diphthong will be treated as a single phoneme (e.g., turning "chide" into "daytch" rather than "dyasht"). The acoustic signal also gives us one reason to treat a combination sound as if it were a single phoneme: the average duration of a combination phoneme is shorter than the average total duration of its component phonemes (e.g., the average duration of /tʃ/ is shorter than the sum of the average durations of /t/ and /ʃ/). In most other respects, however, a combination sound has exactly the same articulatory and acoustic characteristics as a sequence of two separate phonemes, e.g., a /tʃ/ is produced with an unvoiced alveolar closure that looks (e.g., if viewed using MRI) and sounds exactly like a /t/ closure, followed by an unvoiced palatal fricative that looks and sounds exactly like an /ʃ/. The standard IPA notation for these sounds writes them as the sequence of two phonemes (e.g., /t/ followed by /ʃ/) in order to emphasize their articulatory and acoustic decomposibility.

2.3 Quantitative Description of Speech

The preceding discussion has described the production of speech in a completely qualitative way. It has outlined the mechanism of the voice and the means for producing an audible code which, within a given language, consists of distinctive sounds. However, for any transmission system to benefit from prior knowledge of the information source, this knowledge must be cast into a tractable analytical form that can be employed in the design of signal processing operations. Detailed inquiry into the physical principles underlying the speech-producing mechanism is therefore indicated.

The following chapter will consider the characteristics of the vocal system in a quantitative fashion. It will treat the physics of the vocal and nasal tracts in some depth and will set forth certain acoustical properties of the vocal excitations. The primary objective–as stated earlier–is to describe the acoustic speech signal in terms of the physical parameters of the system that produced it. Because of physiological and linguistic constraints, such a description carries important implications for analysis-synthesis telephony.

2.4 Homework

Problem 2.1

Mary and John are talking about something they found on the internet. What are they saying?

Mary: heidʒɑnlʊkætðɪs
John: wʌt
Mary: sʌmgɑisɛzhikənridɑipipieinoʊteiʃʌnwɪðaʊtənɑipipieitʃɑrt
John: wʌtsiriʌsli
Mary: bɔiɑiwɪʃɑikʊdduðæt

Problem 2.2

Create IPA transcriptions of the following sentences.

a. A bird in the hand is worth two in the bush.

b. A stitch in time saves nine.

c. Measure twice, cut once.

d. How much wood would a woodchuck chuck if a woodchuck could chuck wood?

Problem 2.3

Create a table showing the manner, place, and voicing features of all phonemes in the phrase "better speech."

Chapter 3

Acoustics of speech and hearing

The production of speech by the vocal tract and the reception of sound by the cochlea, have a common mathematical framework in that they have both been represented as inhomogeneous transmission lines. While these representations may be improved upon, they capture the essence of the underlying physical mechanisims of production and reception processes. In this chapter we shall outline the underlying theory required to represent both these system. Then in the following chapter (4) we shall concentratate on production, and then in Chapter 7, develop models of the inner ear.

3.1 Equivalent Circuit for the Lossy Cylindrical Pipe

Consider the length dx of lossy cylindrical pipe of area A shown in Fig. 3.1a. Assume plane wave transmission so that the sound pressure and volume velocity are spatially dependent only upon x. Because of its mass, the air in the pipe exhibits an inertance which opposes acceleration. Because of its compressibility the volume of air exhibits a compliance. Assuming that the tube is smooth and hard-walled, energy losses can occur at the wall through viscous friction and heat conduction. Viscous losses are proportional to the square of the particle velocity, and heat conduction losses are proportional to the square of the sound pressure.

The characteristics of sound propagation in such a tube are easily described by drawing upon elementary electrical theory and some well known results for one-dimensional waves on transmission lines. Consider sound pressure analogous to the voltage and volume velocity analogous to the current in an electrical line. Sound pressure and volume velocity for plane wave propagation in the uniform tube satisfy the same wave equation as do voltage and current on a uniform transmission line. A dx length of lossy electrical line is illustrated in Fig. 3.1b. To develop the analogy let us write the relations for the electrical line. The per-unitlength inductance, capacitance, series resistance and shunt conductance are $L(x)$, $C(x)$, $R(x)$, and $G(x)$ respectively.

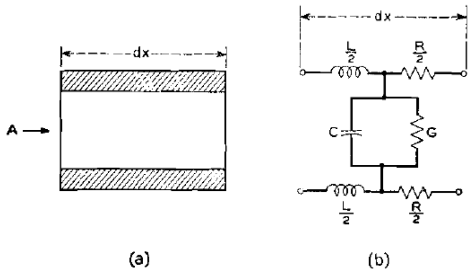


Figure 3.1: Incremental length of lossy cylindrical pipe. (a) acoustic representation; (b) electrical equivalent for a one-dimensional wave

Assuming sinusoidal time dependence for voltage and current, ($i(x,t) = I(x,\omega)e^{j\omega t}$ and $e(x,t) = E(x,\omega)e^{j\omega t}$), the differential current loss and voltage drop across the dx length of line are

$$dE = -\mathcal{Z}(x,s) I dx \quad \text{and} \quad dI = -\mathcal{Y}(x,s) E dx, \tag{3.1}$$

where $\mathcal{Y}(x,s) = G(x) + sC(x)$ and $\mathcal{Z}(x,s) = R(x) + sL(x)$, with complex *Laplace frequency* given by $s = \sigma + j\omega$.

Radian frequency ω is appropriate when taking the *Fourier transform* of say a voltage or a current, pressure, displacement, etc., however the Laplace frequency is the appropriate variable when taking the *Laplace transform* of causal functions,[1] such as an impedance $\mathcal{Z}(s)$ or an admittance $\mathcal{Y}(s)$, which due to their inherent causality must be *analytic* functions of complex frequency s, containing poles and zeros, or other singularities in the complex plane (e.g., branch cuts). More on this topic will be found in Sec. ??. When the Laplace frequency s is used typically the right half s plane ($\sigma > 0$). It will be used with the Laplace transform (indicated as $\leftrightarrow$) when the frequency dependence indicates any causal function, such as the impulse response of a capacitor $\frac{1}{sC} \leftrightarrow \frac{1}{C}\Delta(t) = \int_{\omega=-\infty}^{\infty} \frac{1}{Cs} e^{-st} ds$, where $\Delta(t)$ is the Heaviside *unit step function*, zero for $t < 0$, 1 for $t > 0$ and undefined at $t = 0$.

A short review secti
Laplace in Sec. ??.

In matrix form these equations may be written

$$\frac{d}{dx}\begin{bmatrix} E(x,\omega) \\ I(x,\omega) \end{bmatrix} = -\begin{bmatrix} 0 & \mathcal{Z}(x,s) \\ \mathcal{Y}(x,s) & 0 \end{bmatrix}\begin{bmatrix} E(x,\omega) \\ I(x,\omega) \end{bmatrix}. \tag{3.2}$$

This is the 2×2 matrix form of the *Webester Horn equation* Webster [1919], which allows for the variation of the series *per-unit-length impedance* $\mathcal{Z}(x,s)$ and shunt *per-unit-length admittance* $\mathcal{Y}(x,s)$, as functions both the Horn area $A(x)$, and when losses are considered, the Laplace frequency s.

---

[1] A *causal function* is zero for $t < 0^-$.

3.1.1 The homogeneous transmission line

Shown in Fig. 3.1b is the case of an elementary *homogeneous* piece of transmission line, where $\mathcal{Z}(s)$ and $\mathcal{Y}(s)$ are independent of position x, the equations for the voltage E and current I reduce to the wave equation

$$\frac{d^2E}{dx^2} - \mathcal{Z}\mathcal{Y}\, E = 0 \quad \text{and} \quad \frac{d^2I}{dx^2} - \mathcal{Z}\mathcal{Y}\, I = 0, \tag{3.3}$$

having solutions

$$\begin{bmatrix} E(x,\omega) \\ I(x,\omega) \end{bmatrix} = \begin{bmatrix} E_\omega^+ & E_\omega^- \\ I_\omega^+ & I_\omega^- \end{bmatrix} \begin{bmatrix} e^{-\gamma x} \\ e^{+\gamma x} \end{bmatrix} \tag{3.4}$$

where $\gamma(s) \equiv +\sqrt{\mathcal{Z}\mathcal{Y}} = \alpha(s) + j\beta(s)$ is the so-called *propagation constant* (the sign determines the wave direction), and the frequency-dependent *integration constants* E_ω^+, E_ω^-, I_ω^+ and I_ω^- are frequency dependent integration factors, determined by the 2-port terminal *boundary conditions*. Again, solutions Eq. 3.4 do *not* apply to the Webster Horn equation (Eq. 3.2), where a very different solution approach must be used due to the inhomogeneous *constituent relations* $\mathcal{Z}(x,s)$ and $\mathcal{Y}(x,s)$.

Transmission ($\top$-matrix) ABCD representations Specifically, as shown in Fig. 3.1 for a *finite* piece of homogeneous line, having length l, with sending-end (left) voltage and current E_1 and I_1, the receiving-end (right) voltage and current E_2 and I_2 are given by[2]

$$\begin{bmatrix} E_1(x,\omega) \\ I_1(x,\omega) \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} E_2(x+l,\omega) \\ I_2(x+l,\omega) \end{bmatrix}, \tag{3.5}$$

where

$$A(s) \equiv \left.\frac{E_1}{E_1}\right|_{I_2=0}, \quad B(s) \equiv \left.\frac{E_1}{I_2}\right|_{E_1=0}, \quad C(s) \equiv \left.\frac{I_1}{E_2}\right|_{I_2=0} \quad D(s) \equiv \left.\frac{I_1}{I_2}\right|_{E_2=0}.$$

The inverse Laplace transforms of the $\top$-matrix parameters $a(t) \leftrightarrow A(s)$, $b(t) \leftrightarrow B(s)$, $c(t) \leftrightarrow C(s)$ and $d(t) \leftrightarrow D(s)$ must be analytic functions of Laplace frequency s in the right-half plane because they must represent causal functions (e.g., $a(t < 0^-) = 0$). This follows from the fact the the voltages and currents of the circuit must all obey causal relationships. Furthermore, the determinite of the $\top$-Matrix must be unity, that is $\Delta_\top \equiv AB - CD = 1$. This follows from the requirement that the passive physical system being modeled is reciprocal.

Given our example in Fig. 3.1a of a stub of transmission line of length l,

$$\begin{bmatrix} E_1(x,\omega) \\ I_2(x,\omega) \end{bmatrix} = \begin{bmatrix} \cosh(\gamma l) & Z_0 \sinh(\gamma l) \\ Y_0 \cosh(\gamma l) & \sinh(\gamma l) \end{bmatrix} \begin{bmatrix} E_2 \\ I_2 \end{bmatrix}, \tag{3.6}$$

---

[2] The transmission element C should not be confused with the acoustic compliance "C".

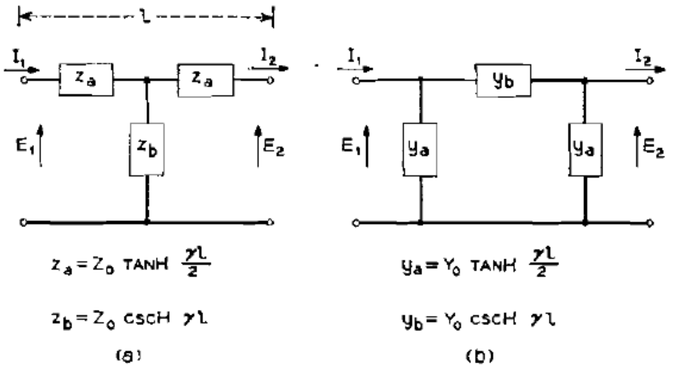


Figure 3.2: Equivalent 2-port networks for a length l of uniform transmission line. (a) T-section; (b) Π-section. These represent two alternative ways of specifying the impedances. For example the pipe representation of Fig. 3.1b is a T-section with $z_a = sL/2$, $z_c = R/2$ and $z_b = 1/(G + sC)$. in (a) z_c and (b) y_c need to be added.

where

$$Z_0 = 1/Y_0 \equiv \sqrt{\mathcal{Z}/\mathcal{Y}} \tag{3.7}$$

is the *characteristic impedance* (Y_0 is the *characteristic admittance*) of the line and

$$\gamma \equiv +\sqrt{\mathcal{Z}\mathcal{Y}} \tag{3.8}$$

is the wave's *propagation function*. This relationship follows from Eq. 3.4, the solution of the wave equation.

One recalls from conventional circuit theory the lossless case corresponds to $\gamma = \sqrt{\mathcal{Z}\mathcal{Y}} = j\beta = s\sqrt{LC}$, and $Z_0 = \sqrt{L/C}$. The hyperbolic functions then reduce to circular functions which are purely reactive. Notice, too, for small loss conditions, (that is, $R \ll \omega L$ and $G \ll \omega C$) the attenuation and phase constants are approximately[3]

$$\alpha \approx \frac{Y_0 R}{2} + \frac{Z_0 G}{2}$$

$$\beta \approx \frac{\omega}{c}, \tag{3.9}$$

with wave velocity (i.e., speed of sound) $c \equiv 1/\sqrt{LC}$.

The $\top$-matrix is useful for calculations since it is designed to be cascaded, to build a network. However it is frequenly more convient to define a system in terms of the *impedance matrix* $\mathbb{Z}(s)$. We need to know how to transfrom between the $\top$ and $\mathbb{Z}$ matrix forms.

---

[3] In the case of losses, the signs of α and β for the forward and reverse traveling waves must be determined such that the solution is both causal and stable. This involves the careful choice of the branch cut resulting from the square root in Eq. 3.8.

	$[\mathbb{Z}]$	$[\top]$	T-section (Fig. 3.2a)
$[\mathbb{Z}]$	$\begin{matrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{matrix}$	$\begin{matrix} A/C & \Delta_\top/C \\ 1/C & D/C \end{matrix}$	$\begin{matrix} z_a + z_c & z_b \\ z_b & z_a + z_c \end{matrix}$
$[\top]$	$\begin{matrix} z_{11}/z_{21} & \Delta_{\mathbb{Z}}/z_{21} \\ 1/z_{21} & z_{22}/z_{21} \end{matrix}$	$\begin{matrix} A & B \\ C & D \end{matrix}$	$\begin{matrix} (z_a + z_c)/z_b & z_a^2/z_b + 2z_a \\ 1/z_b & (z_a + z_c)/z_b \end{matrix}$

Table 3.1: Table of transformations between $\mathbb{Z}$-matrix (Eq. 3.10) and $\top$-matrix (Eq. 3.5) representations. Note that the sign of I_2 in the $\top$-matrix is switched for the $\mathbb{Z}$-matrix (i.e., $I_2' = -I_2$).
On the right-most column the forms are defined in terms of the T-secion parameters of Fig. 3.2a. $\Delta_\top$ and $\Delta_{\mathbb{Z}}$ represent the determinates of $[T]$ and $[\mathbb{Z}]$.

Impedance matrix representation The general 2-port *impedance matrix* $\mathbb{Z}(s)$ representation is defined as

$$\begin{bmatrix} E_1(x,\omega) \\ E_2(x,\omega) \end{bmatrix} = \mathbb{Z}(s) \begin{bmatrix} I_1 \\ I_2' \end{bmatrix} \equiv \begin{bmatrix} z_{11}(s) & z_{12}(s) \\ z_{21}(s) & z_{22}(s) \end{bmatrix} \begin{bmatrix} I_1 \\ I'_2 \end{bmatrix} \tag{3.10}$$

with $I'_2 \equiv -I_2$.

The T-section format for the l length of line is shown in Fig. 3.2a, along with the Π-section format in Fig. 3.2b. These forms (T and Π sections) are the most schematic way of specifying a transmission line in terms of its physical parameters. From the T-section parameters (e.g., z_a, z_b) the $\top$-matrix and $\mathbb{Z}$-matrix parameters may be easily derived.

This relationship between the T-section parameters (Fig. 3.2a), the $\top$-matrix parameters $A(s)$, $B(s)$, $C(s)$ and $D(s)$ and the $\mathbb{Z}$-matrix elements z_{ij}, etc., are summarized in Table 3.1. By inspection of Eq. 3.10 and Fig. 3.2a,

$$z_{11} \equiv \frac{E_1}{I_1}\bigg|_{I'_2=0} = z_a + z_b \qquad z_{12} \equiv \frac{E_1}{I'_2}\bigg|_{I_1=0} = z_b$$

$$z_{21} \equiv \frac{E_2}{I_1}\bigg|_{I'_2=0} = z_b \qquad z_{22} \equiv \frac{E_2}{I'_2}\bigg|_{I_1=0} = z_a + z_c.$$

A system is *reciprocal* when $z_{12} = z_{21}$, or in terms of the $\top$-matrix representation, when $\Delta_\top \equiv AC - BD$. If a network is identical when run in reverse (if the system is invariant to having the ports flipped), the networks is called *symmetrical*, leading to $z_{11} = z_{22}$, or equivalently $A = D$.

It is important to note that there are cases where a physical system does not have an impedance representation, for example, when $C = 0$. The $\top$-matrix form, on the otherhand, always exists. Thus the prefered form must be for the $\top$-matrix form.

3.2 Acoustic transmission lines

Having reviewed the relations for the uniform, lossy electrical line, we may next interpret plane wave propagation in a uniform, lossy pipe (Fig. 3.1a) in analogous terms. Since the sound pressure $p(t) \leftrightarrow P(\omega)$ is considered analogous to voltage and the acoustic volume velocity, $u(t) \leftrightarrow U(\omega)$ analogous to current, the lossy, one dimensional, sinusoidal sound propagation is described by Eq. 3.3. The propagation constant is complex (that is, the velocity of propagation is in effect complex) and therefore the wave attenuates as it travels. In a smooth hard-walled tube the viscous and heat conduction losses can be represented, in effect, by an I^2R loss and an E^2G loss, respectively. The inertance of the air mass is analogous to the electrical inductance, and the compliance of the air volume is analogous to the electrical capacity. We can draw these parallels quantitatively.[4]

3.2.1 The Acoustic Mass "L"

The mass of air contained in the dx length of pipe in Fig. 3.1a is $\rho_0 A dx$, where ρ_0 is the air density. The *excess pressure* $p(x,t)$ accelerating this mass is (i.e., Newton's law):

$$dp = \rho_0 dx \frac{dv}{dt} = \rho_0 \frac{dx}{A} \cdot \frac{du(x,t)}{dt},$$

where $v(x,t)$ is particle velocity and $u(x,t)$ is the *volume velocity*, define as the area times the normal component of the particle velocity, namely

$$u(x,t) \equiv Av. \tag{3.11}$$

If we take the *Fourier Transform*[5] of these relations [e.g., $p(x,t) \leftrightarrow P(x,\omega)$ and $u(x,t) \leftrightarrow U(x,\omega)$] we obtain frequency domain relationships

$$\frac{d}{dx} P(x,\omega) = sL_a U(x,\omega), \tag{3.12}$$

where

$$L_a = \rho_0 / A \tag{3.13}$$

is the *per-unit-length acoustic inertance* .

It is our convention to use radian frequency ω for non-causal functions, having a Fourier transfrom, and complex radian frequency $s = \sigma + j\omega$ for *causal functions* of frequency having a *Laplace transform* (e.g., impedance sL_a defined via Eq. 3.12). We shall keep track of causality with this notational difference in frequency dependence (i.e., ω vs. s).

---

[4] The first-time reader might wish to omit the following detailed sections and jump to the summary Eq. 3.46 in Section 3.2.5.

[5] The *Fourier Transform relations* are defined as $P(\omega) \equiv \int_{-\infty}^{\infty} p(t) e^{-j\omega t} dt \quad \leftrightarrow \quad p(t) \equiv \int_{-\infty}^{\infty} P(\omega) e^{j\omega t} d\omega / 2\pi$.

3.2.2 The Acoustic "C"

The analogous acoustic capacitance, or compliance, arises from the compressibility of the volume $\mathcal{V}$ of air contained in the dx length of tube shown in Fig. 3.1a. The elemental air volume $\mathcal{V} = Adx$ experiences compressions and expansions following the adiabatic gas law

$$p\mathcal{V}^{\eta} = \text{constant},$$

where p is the total pressure (the static + excess), $\mathcal{V}$ is the volume of the gas, and η is the adiabatic constant defined as the ratio of specific heat at constant pressure c_p to that at constant volume c_v.[6]

After taking the log and differentiating with respect to time

$$\frac{1}{p}\frac{dp}{dt} = -\frac{\eta}{\mathcal{V}}\frac{d\mathcal{V}}{dt}.$$

---

[6] For diatomic air at normal conditions, $\eta = c_p/c_v = 1.4$.

The diminution of the original air volume, owing to compression caused by an increase in pressure, must equal the volume current into the compliance; that is,

$$u = -\frac{d\mathcal{V}}{dt},$$

and

$$\frac{1}{p}\frac{dp}{dt} = \frac{\eta}{\mathcal{V}}u.$$

For sinusoidal time dependence $P = P_0 + P(x,\omega)e^{j\omega t}$, where P_0 is the quiescent (ambient) static pressure, large[7] compared with the *excess pressure* $P(x,\omega)$.[8]

---

[7] Under standard conditions $P_0 = 10^5$ [Pa].

[8] The *excess pressure* is the small (e.g. less than 10 Pascals) acoustic time-varying component riding on the ambient static pressure.

The volume flow into the compliance of volume Adx is therefore

$$\frac{d}{dx}U(x,\omega) = s\frac{A}{P_0\eta}P(x,\omega). \tag{3.14}$$

The *per-unit-length acoustic compliance* is defined via Ohm's Law as the ratio of the volume velocity over the excess pressure

$$sC_a \equiv \frac{U(x,\omega)}{P(x,\omega)},$$

with

$$C_a = \frac{A}{P_0\eta} = \frac{A}{\rho_0 c^2}. \tag{3.15}$$

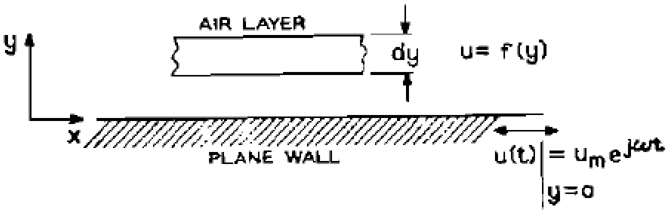


Figure 3.3: Relations illustrating viscous loss at the wall of a smooth tube

This last relation follows from the derivation of the acoustic wave equation Eq. 3.3 and the propagation constant Eq. 3.8, the speed of sound, which is given by [Morse, 1948]

$$c = \sqrt{\frac{P_0 \eta}{\rho_0}}. \tag{3.16}$$

Thus $P_0\eta = \rho_0 c^2$.

3.2.3 The Acoustic "R"

Add exact Helmholt
for a pipe, following

The acoustic R represents the power dissipated in viscous friction at the tube wall, a loss proportional to $|U|^2$ [Ingard, 1953]. The history of this analysis goes back to Stokes, Helmholtz and Kirchhoff, and Rayleigh [Keefe, 1984]. The easiest way to put in evidence this equivalent surface resistance is to consider the situation shown in Fig. 3.3. Imagine that the tube wall is a plane surface, large in extent, and moving sinusoidally in the x-direction with harmonic velocity $U(x, y, \omega)$. The air particles proximate to the wall experience a force owing to the viscosity μ of the medium. The power expended per-unit-area in dragging the air with the plate is the loss to be determined.

Consider a layer of air dy thick and of unit area normal to the y axis, The net force on the layer is

$$\mu \left[\left(\frac{\partial u}{\partial y} \right)_{y+dy} - \left(\frac{\partial u}{\partial y} \right)_{y} \right] = \rho_0 dy \frac{\partial u}{\partial t},$$

where $u(x, y, \omega t)$ is the harmonic particle velocity in the x-direction. The diffusion equation specifying the air particle velocity as a function of the distance above the wall is then [Hildebrand, 1948]

$$\frac{\partial^2 u}{\partial y^2} = \frac{\rho_0}{\mu} \frac{\partial u}{\partial t}. \tag{3.17}$$

Taking a Fourier Transform we obtain the harmonic time dependence

$$\frac{d^2 U}{dy^2} = s \frac{\rho_0}{\mu} U = k_v^2 U, \tag{3.18}$$

where $k_v = \sqrt{s\rho_0/\mu} = (1+j)\sqrt{\omega\rho_0/2\mu}$,[9] The harmonic velocity distribution is

$$U(x,y,\omega) = U(x,y=0,\omega)e^{-k_v y} = U(x,y=0,\omega)\cdot e^{-\sqrt{\omega\rho_0/2\mu}\,y}\cdot e^{-j\sqrt{\omega\rho_0/2\mu}\,y} \tag{3.20}$$

The distance required for the particle velocity to diminish to $1/e$ of its value at the driven wall is often called the *boundary-layer thickness* and is

$$\delta_v = \sqrt{2\mu/\omega\rho_0}. \tag{3.21}$$

I prefer the more natural anaytic bounary-layer.

In air at a frequency of 100Hz, for example, $\delta_v \approx 0.2$ [mm].[10]

The viscous drag, per-unit area, on the plane wall is

$$F = -\mu\left.\frac{\partial U}{\partial y}\right|_{y=0} = \mu k_v U(y=0),$$

or

$$F = U\sqrt{s\mu\rho_0}. \tag{3.22}$$

Notice that this force has a real part and a positive reactive part. The latter acts to increase the apparent acoustic L. The average power dissipated per-unit surface area in this drag is

$$\bar{P} = \frac{1}{2}|F||U_m|\cos\theta = \frac{1}{2}|U_m|^2 R_s, \tag{3.23}$$

where $R_s = \sqrt{\omega\rho_0\mu/2}$ is the *per-unit-area surface resistance* and $\theta = \pi/2$ is the phase angle between F and U_m (e.g., 45°). For a length l of the acoustic tube, the inner surface area is Sl, where S is the circumference. Therefore, the average power dissipated per-unit length of the tube is $\bar{P}S = \frac{1}{2}u_m^2 SR$ or in terms of the acoustic volume velocity

$$\bar{P}S = \frac{1}{2}U_m^2 R_a,$$

where

$$R_a = \frac{S}{A^2}\sqrt{\omega\rho_0\mu/2}, \tag{3.24}$$

and A is the cross-sectional area of the tube. R_a is then the *per-unitlength acoustic resistance* for the analogy shown in Fig. 3.1.

As previously mentioned, the reactive part of the viscous drag contributes to the acoustic inductance per-unit-length. In fact, for the same area and surface relations

---

[9] Note that this involves the *"usual 1/2 derivative"* formula, resulting from $\sqrt{s}$ [Lighthill, 1978, page 21, Eq. 77], which may be implimented by a time-convolution with the operator (and its Laplace transform)

$$\frac{\Delta(t)}{\sqrt{\pi t}} \star \frac{d}{dt} \leftrightarrow \frac{s}{\sqrt{s}} = \sqrt{s}. \tag{3.19}$$

[10] Some prefer the alternate definition of a more general *analytic boundary-layer thickness* $\delta_v = \sqrt{\mu/s\rho_0}$, without the factor of 2.

applied above, the acoustic inductance obtained in the foregoing section should be increased by the frequency dependent factor $\frac{A^2}{S}\sqrt{\mu\rho_0/2\omega}$, or

$$L_a \approx \frac{\rho_0}{A}\left(1+\frac{S}{A}\sqrt{\frac{\mu}{2\rho_0\omega}}\right). \tag{3.25}$$

Thus, the viscous boundary layer increases the apparent acoustic mass and slightly diminishes the cross-sectional area. For vocal tract analysis, the viscous boundary layer is typically so thin that the second term in Eq. 3.25 is negligible. For example, for a circular cross-section of 9 cm^2, the second term at a frequency of 500Hz is about (0.006) ρ_0/A (i.e., factor of 0.6%), and even smaller at higher frequencies.

The 2-port mass

Each acoustic element may also be written as a 2-port transmission matrix. For the acoustic mass element "L" this is equivilent to chosing the elemental dx to be a finite length, say l meters long. In this case we may describe the mass as the following 2-port transmission matrix notation

$$\begin{bmatrix} P_1(x,\omega) \\ U_1(x,\omega) \end{bmatrix} = \begin{bmatrix} 1 & R_a + sM_a \\ 0 & 1 \end{bmatrix} \begin{bmatrix} P_2(x,\omega) \\ U_2(x,\omega) \end{bmatrix}, \tag{3.26}$$

The port variables on the left $[P_1, U_1]$ represent the input with volume velocity U_1 defined into the port, while the variables on the right $[P_2, U_2]$ represent the output, with volume velocity U_2 out of the port. The acoustic mass is given by

$$M_a l \equiv L_a = \frac{l\rho_0}{A}\left(1+\frac{S}{A}\sqrt{\frac{\mu}{2\rho_0\omega}}\right). \tag{3.27}$$

where l is the length of the tube, ρ_0 is the density of air and A is the tube area. As will be discussed in greather depth below, the length must be modified to account for spreading of the waves as it leaves the tube, or if one tube is connected to a second having a different area. This end correction is a function of the ratio of the two tube areas.

3.2.4 The Acoustic "G"

Add exact Helmholt
for a pipe, following

The analogous shunt conductance provides a power loss proportional to the square of the local sound pressure. Such a loss arises from heat conduction at the walls of the tube. The *per-unit-length conductance* can be deduced in a manner similar to that for the viscous loss. As before, it is easier to treat a simpler situation and extend the result to the vocal tube.

Consider a highly conductive plane wall of large extent, such as shown in Fig. 3.4. The air above the boundary is essentially at constant pressure and has a coefficient of

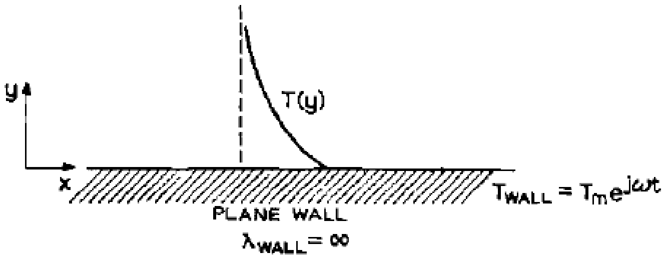


Figure 3.4: Relations illustrating heat conduction at the wall of a tube

heat conduction λ and a specific heat c_p. Suppose the wall is given an oscillating temperature $T|_{x,y=0} = T_m(y=0)e^{j\omega t}$. The vertical temperature distribution produced in the air is described by the diffusion equation[Hildebrand, 1948].

$$\frac{\partial^2 T}{\partial y^2} = \frac{c_p \rho_0}{\lambda}\frac{\partial T}{\partial t},$$

Mixing freq and time domain!?

which in the frequency domain is

$$\frac{\partial^2 T_m}{\partial y^2} = j\omega \frac{c_p \rho_0}{\lambda} T_m. \tag{3.28}$$

The solution is therefore $T = T_m(y=0,\omega)e^{-k_h y + j\omega t}$, where

$$k_h = \sqrt{\frac{s c_p \rho_0}{\lambda}} \tag{3.29}$$

which is the same form as the velocity distribution due to viscosity. In a similar fashion, the boundary layer depth for temperature is $\delta_h = \sqrt{2\lambda/\omega c_p \rho_0}$ (or complex $\delta_h = \sqrt{\lambda/s c_p \rho_0}$), thus $k_h = (1+j)/\delta_h$.

fix def. of δ_h

Now consider more nearly the situation for the sound wave. Imagine an acoustic pressure wave moving parallel to the conducting boundary, that is, in the x-direction. We wish to determine the temperature distribution above the wall produced by the sound wave. The conducting wall is assumed to be maintained at some quiescent temperature and permitted no variation, that is, $\lambda_{wall} = \infty$. If the sound wavelength is long compared to the boundary extent under consideration, the harmonic pressure variation above the wall may be considered as $P = P_0 + p$, where P_0 is the quiescent atmospheric pressure and $p = p_m e^{j\omega t}$ is the pressure variation. (That is, the spatial variation of p with x is assumed small.) The gas laws prescribe

$$PV^\eta = \text{constant} \quad \text{and} \quad PV = RT \quad \text{(for unit mass)}.$$

Taking differentials gives

$$\frac{d\mathcal{V}}{\mathcal{V}} = -\frac{1}{\eta}\frac{dP}{P} \quad \text{and} \quad \frac{dP}{P} + \frac{dV}{V} = \frac{dT}{T} \tag{3.30}$$

Combining the equations yields

$$\frac{dP}{P}\left(1-\frac{1}{\eta}\right)=\frac{dT}{T}, \tag{3.31}$$

where

$$dP = p = p_m e^{j\omega t}$$
$$dT = \tau = \tau_m e^{j\omega t},$$

so from (Eq. 3.31)

$$\tau_m = \frac{T_0}{P_0}\left(\frac{\eta-1}{\eta}\right)p_m \tag{3.32}$$

At the wall, $y = 0$ and $\tau(0) = 0$ (because $\lambda_{wall} = \infty$). Far from the wall (i.e., for y large), $|\tau(y)| = \tau_m$ as given in (Eq. 3.32). Using the result of (Eq. 3.29), the temperature distribution can be constructed as

$$\tau(y,t) = \left[1 - e^{-k_{hy}}\right]\tau_m e^{j\omega t},$$

or

$$\tau(y,t) = \frac{P_0}{T_0}\left(\frac{\eta-1}{\eta}\right)\left[1 - e^{-k_h y}\right]p_m e^{j\omega t}. \tag{3.33}$$

This equation is the phasor at frequency ω, thus the ratio $\tau/p_m e^{j\omega t}$ is an admittance, and therefore an analytic function of s in the right-half plane.

Now consider the power dissipation at the wall corresponding to this situation. A long wavelength sound has been assumed so that the acoustic pressure variations above the boundary can be considered $p = p_m e^{j\omega t}$, and the spatial dependence of pressure neglected. Because of the temperature distribution above the boundary, however, the particle velocity will be nonuniform, and will have a component in the y-direction. The average power flow per-unit surface area into the boundary is $p\bar{u}_{y0}{}^{t}$, where u_{y0} is the velocity component in the y direction on the boundary. To examine this quantity, u_y is needed.

Conservation of mass in the y-direction requires

$$\rho_0\frac{\partial u_y}{\partial y} = -\frac{\partial \rho_0}{\partial t}. \tag{3.34}$$

Also, for a constant mass of gas $d\rho_0/\rho_0 = -dV/V$ which with the second equation in (Eq. 3.30) requires

$$\frac{dP}{P} - \frac{d\rho_0}{\rho_0} = \frac{dT}{T}. \tag{3.35}$$

Therefore,

$$\frac{\partial u_y}{\partial y} = \left(\frac{1}{T_0}\frac{\partial \tau}{\partial t} - \frac{1}{P_0}\frac{\partial p}{\partial t}\right), \tag{3.36}$$

and

$$u_y = \int \frac{\partial u_y}{\partial y} \cdot dy$$

$$y_y = \frac{j\omega p}{P_0} \left\{ \frac{\eta - 1}{\eta} \left(y + \frac{e^{-k_y y}}{ky} \right) - y \right\}. \tag{3.37}$$

And,

$$u_{y_0} = p\frac{\omega}{c}\frac{\eta - 1}{\rho_0 c}\frac{j}{1+j}\delta_h. \tag{3.38}$$

The equivalent energy flow into the wall is therefore

$$W_h = p\bar{u}_{y_0}{}^t = \frac{\omega}{c}\frac{\eta - 1}{\rho_0 c}\delta_h \frac{1}{\sqrt{2}}\frac{1}{T}\int_0^T P_m^2 \cos\left(\omega t + \frac{\pi}{4}\right)\cos\omega t \cdot dt$$

$$W_h = \frac{1}{4}\frac{\omega}{c}\frac{\eta - 1}{\rho_0 c}\delta_h p_m^2 = \frac{1}{2}G_\alpha p_m^2, \tag{3.39}$$

where G_α is an equivalent conductance per-unit wall area and is equal

$$G_\alpha = \frac{1}{2}\frac{\omega}{c}\frac{\eta - 1}{\rho_0 c}\sqrt{\frac{2\lambda}{\omega c_p \rho_0}}. \tag{3.40}$$

The equivalent conductance per-unit-length of tube owing to heat conduction is therefore

$$G_\alpha = S\frac{\eta - 1}{\rho_0 c^2}\sqrt{\frac{\lambda\omega}{2c_p\rho_0}}, \tag{3.41}$$

where S is the tube circumference. To reiterate, both the heat conduction loss G; and the viscous loss R_α are applicable to a smooth, rigid tube. The vocal tract is neither, so that in practice these losses might be expected to be somewhat higher. In addition, the mechanical impedance of the yielding wall includes a mass reactance and a conductance which contribute to the shunt element of the equivalent circuit. The effect of the wall reactance upon the tuning of the vocal resonances is generally small, particularly for open articulations. The contribution of wall conductance to tract damping is more important. Both of these effects are estimated in a later section.

The 2-port compliance

The acoustic compliance element "$G_a + sC_a$" is equivilent to chosing the elemental dx to be a finite l meters long. In this case we may describe the mass as the following 2-port transmission matrix notation

$$\begin{bmatrix} P_1(x,\omega) \\ U_1(x,\omega) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ G_a + sC_a & 1 \end{bmatrix} \begin{bmatrix} P_2(x,\omega) \\ U_2(x,\omega) \end{bmatrix}, \tag{3.42}$$

The port variables on the left $[P_1, U_1]$ represent the input with volume velocity U_1 defined into the port, while the variables on the right $[P_2, U_2]$ represent the output, with volume velocity U_2 out of the port. The acoustic compliance is given by

$$C \equiv lC_a = \frac{Al}{\eta P_0} \tag{3.43}$$

where A is the tube area.

3.2.5 Summary of the Analogous Acoustic Elements

In matrix form the acoustic equations may be written in a manner similar to the electrical case, but in terms of the pressure P playing the role of a force, and the volume velocity U playing the role of the flow

$$\frac{d}{dx}\left[\begin{array}{c} P(x,\omega) \\ U(x,\omega) \end{array}\right] = -\left[\begin{array}{cc} 0 & \mathcal{Z}(x,s) \\ \mathcal{Y}(x,s) & 0 \end{array}\right]\left[\begin{array}{c} P(x,\omega) \\ U(x,\omega) \end{array}\right]. \tag{3.44}$$

with

$$\mathcal{Z} \equiv R_a + sL_z \qquad \text{and} \qquad \mathcal{Y} \equiv G_a + sC_a. \tag{3.45}$$

The per-unit-length analogous constants for the uniform pipe are

$$\begin{array}{ll} L_a = \frac{\rho_0}{A}, & C_a = \frac{A}{\rho_0 c^2}, \\ R_a = \frac{S}{A^2}\sqrt{\frac{\omega\rho_0\mu}{2}}, & G_a = S\frac{\eta-1}{\rho_0 c^2}\sqrt{\frac{\lambda\omega}{2c_p\rho_0}}, \end{array} \tag{3.46}$$

where A is tube area, S is tube circumference, ρ_0 is air density, c is sound velocity, μ is viscosity coefficient, A is coefficient of heat conduction, η is the adiabatic constant, and c_p is the specific heat of air at constant pressure.[11]

Having set down these quantities, it is possible to approximate the nonuniform vocal tract with as many right circular tube sections as desired by cascading the transmission matrix of each tube. The transmission characteristics can be determined

---

[11] Summarizing the constants:

$P_0 = 10^5$ Pa
$\rho_0 = 114$ kgm/m^3 (moist air at body temperature 37deg C)
$c = \sqrt{\gamma P_0/\rho_0} = 343$ m/sec (moist air at body temperature, 37deg C)
$\eta = c_p/c_v = 1.4$
$c_p = 0.24$ cal/gm-degree(Odeg C, 1 atmos)transform to MKS units
$\mu = 1.86 \times 10^{-4}$ dyne-sec/cm^2(20C, 0.76 m.Hg)transform to MKS units
$\lambda = 0.055 \times 10^{-3}$ cal/gm-sec-deg(0deg C)transform to MKS units

either from calculations on equivalent network sections such as shown in Fig. **??**, or from electrical circuit simulations of the clements. When the approximation involves more than three or four network loops, manual computation becomes prohibitive. Computer techniques can then be used to good advantage.

A further level of approximation can be made for the equivalent networks in Fig. **??**. For a given length of tube, the hyperbolic elements may be approximated by the first terms of their series expansions, namely,

$$\tanh x = x - \frac{x^3}{3} + \frac{2x^5}{15} \cdots$$

and

$$\sinh x = x + \frac{x^3}{3!} + \frac{x^5}{5!} \cdots$$

so that

$$z_a = Z_0 \tanh \frac{\gamma l}{2} \approx \frac{1}{2}(R_a + j\omega L_a) l$$

and

$$\frac{1}{z_b} = \frac{1}{Z_0} \sinh \gamma l \approx (G_a + j\omega C_a) l. \tag{3.47}$$

The error incurred in making this approximation is a function of the elemental length l and the frequency, and is

$$\left(1 - \frac{x}{\tanh x}\right) \text{ and } \left(1 - \frac{x}{\sinh x}\right),$$

respectively. In constructing electrical analogs of the vocal tract it has been customary to use this approximation while keeping l sufficiently small. We shall return to this point later in the chapter.

We will presently apply the results of this section to some simplified analyses of the vocal tract. Before doing so, however, it is desirable to establish several fundamental relations for sound radiation from the mouth and for certain characteristics of the sources of vocal excitation.

3.3 Acoustic Horns

In this section we move from uniform tubes of constant crossection to the topic of horns, which have an area that is changing in the direction of wave propagatation. This is a very important topic to communication acoustics. First when the area changes, the impedance also must change. Thus horns are used to transform the acoustic impedance from one end to the other end of the horn. Second horns are a special case on inhomogeneous media, a very important topic to speech production and hearing. To deal with these issues we must start from the basic equations of acoustics in 3 dimensions. The equation of a horn is then typically an approximation which reduces the 2 or 3 dimentional wave propagation to a function of the axial variable. The methods used to do this are reviewed next.

In three dimensions the basic acoustic equations are based on two laws, Newton's second law of conservation of momentum

$$\nabla P = -s\rho_0\mathbf{U} \quad [\texttt{eq : gradP}] \tag{3.48}$$

and Hooke's Law for the adiabatic compressibility of air

$$\nabla\cdot\mathbf{U} = -\frac{s}{\eta P_0}P, \quad [\texttt{eq : divU}] \tag{3.49}$$

where $P(\omega)$ is the pressure, $\mathbf{U}(\omega)$ is the vector particle velocity, with s the Laplace frequency $s = \sigma + j\omega$, ρ_0 the density of air, $\eta \equiv c_p/c_v$, P_0 is the static pressure of air. We refer to the ratio of pressure to particle velocity as the *specific acoustic impedance* in [Rayls], and the pressure over a volume velocity is the *acoustic impedance* in [*acoustic ohms*].

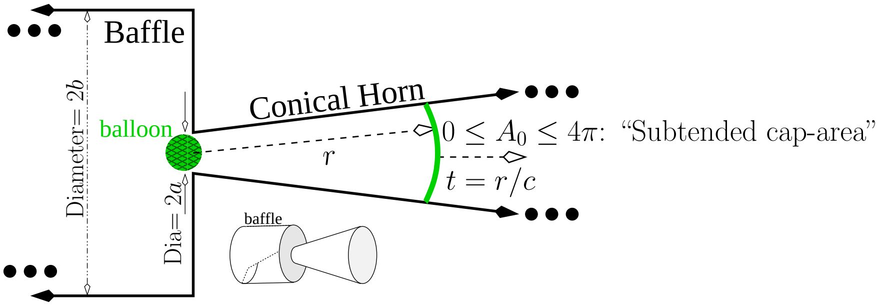


Figure 3.5: *Experimental setup showing a large pipe on the left terminating the wall containing a small hole with a balloon, shown in green. At time $t = 0$ the ballon is pricked and a pressure pulse is created. The baffel on the left is meant to represent an ∞ long tube having a very large radius compared to the horn input diameter $2a$, such that the acoustic admittance looking to the left ($A/\rho_0 c$ with $A \to \infty$), is very large compared to that looking into the horn, $Y_+(a, s)$ (Eq. 3.61). At time $t = b/c$ the outbound pressure pulse $\delta(t - b/c)/r$ has reached radius b.* `[fig:Exp-horn]`

3.3.1 Spherical acoustics

The conical horn is a special case with an exact solution due to its cylindrical symmetry. In spherical coordinates the pressure is given by Eq. 3.48

$$\nabla_r P \equiv \frac{dP}{dr} = -s\rho_0 U_r \tag{3.50}$$

and the velocity is determined by Eq. 3.49

$$\nabla_r \cdot \mathbf{U} \equiv \frac{1}{r^2}\frac{d}{dr}r^2 U_r = -\frac{s}{\eta P_0}P, \tag{3.51}$$

where the subscript r indicates the radial component of the pressure gradient and velocity divergence.

If we define the *per-unit length series acoustic impedance*

$$\mathcal{Z}(r,s) = s\frac{\rho_0}{A(r)} \tag{3.52}$$

and the *per-unit length shunt acoustic admittance*

$$\mathcal{Y}(r,s) = s\frac{A(r)}{\eta P_0}, \tag{3.53}$$

with the horn's *subtended cap area* as $A(r) = A_0 r^2$ (A_0 is the fraction of subtended cap area relative to 4π, as shown in Fig. 3.5), then the transformed spherical equations may be written in *Webster form* (i.e. see Eq. 3.2) in terms of the pressure P volume velocity $V = A(r)U_r$ as

$$\frac{d}{dr}\begin{bmatrix} P(r,\omega) \\ V(r,\omega) \end{bmatrix} = -\begin{bmatrix} 0 & \mathcal{Z}(r,s) \\ \mathcal{Y}(r,s) & 0 \end{bmatrix}\begin{bmatrix} P(r,\omega) \\ V(r,\omega) \end{bmatrix}. \text{[eq : Webster2]} \tag{3.54}$$

This is a scaled form of the conical horn (it is a horn because the *per-unit-length impedance* $\mathcal{Z}$ and admittance $\mathcal{Y}$ are a function of r), with an angle that subtends 4π [sr] (steradians). Expressing the solution in terms of the volume velocity after scaling $A(r)$, we obtain the traditional conical horn equation Eq. 4.67 [Salmon, 1946a,b, Morse, 1948, p. 271].

Thus the spherical wave solution may be expressed as a Webster (conical) horn equation. If one reduces this to a second-order equation in pressure, the classic Webster horn equation results [Webster, 1919, Salmon, 1946a,b, Morse, 1948, Kinsler and Frey, 1962, Leach, 1996, Pierce, 1981].

The functions $\mathcal{Z}$ and $\mathcal{Y}$ define the *acoustic characteristic impedance (resistance)* that depends on the radius

$$Z_0(r) \equiv \sqrt{\frac{\mathcal{Z}}{\mathcal{Y}}} = \frac{\sqrt{\rho_0 \eta P_0}}{A(r)} = \frac{\rho_0 c}{A(r)} = \frac{1}{Y_0(r)} \text{ [eq : Z0]} \tag{3.55}$$

and a wave *propagation function*, closely related to the sound speed by

$$\gamma \equiv \sqrt{\mathcal{Z}\mathcal{Y}} = s/c \tag{3.56}$$

that for all horns, is independent of the axial coordinate r [Morse, 1948].

Equation 3.54 may be reexpressed as a single equation in pressure as

$$\frac{d^2}{dr^2} rP = \gamma^2 rP. \tag{3.57}$$

Following d'Alembert (1747), the solution to this equation in spherical coordinates, corresponding to a spherical symmetry, is given by the sum of an outbound $\mathcal{P}^+(r,\omega)$

and an inbound $\mathcal{P}^-(r,\omega)$ wave [Salmon, 1946a,b, Morse, 1948, Pierce, 1981, Leach, 1996]

$$P(r,\omega) = \mathcal{P}^+_\omega \frac{e^{-sr/c}}{r} + \mathcal{P}^-_\omega \frac{e^{+sr/c}}{r} \quad \texttt{[eq:dAlembert]} \tag{3.58}$$

where $\mathcal{P}^+_\omega(a) \leftrightarrow p^+(t)$ and $\mathcal{P}^-_\omega(a) \leftrightarrow p^-(t)$ are the Fourier "source strengths" and a is the radius corresponding to the source location (the radius the waves are launched from). When $\mathcal{P}^+_\omega(a) = 1$ and $\mathcal{P}^-_\omega(a) = 0$, the resulting outbound pressure wave is a Dirac delta function

$$p(r,t) = \frac{\delta(t - r/c)}{r} \leftrightarrow P(r,\omega) = \frac{e^{-sr/c}}{r}. \quad \texttt{[eq:p+]} \tag{3.59}$$

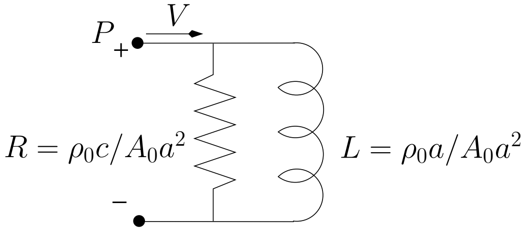


Figure 3.6: *Equivalent circuit of the* radiation acoustic impedance *of a conical horn for Eq. 3.61. Parameter* A_0 *is a number between* 0 *and* 4π *determined by the horn's* subtended cap area, *as defined by Fig. 3.5, while* a *is the horn's throat radius. The resistance* R *and the reactance* sL *are equal at frequency given by* $ka = 1$, *or* $\omega_c = c/a$. *This makes the connection between resonant scattering and the reactive component of the mass. This mass is also called the* spreading inertance. `[fig:Zrad]`

3.3.2 Particle velocity and the radiation impedance

Substituting the expression for the pressure (Eq. 3.58) into Eq. 3.48 results in an expression for the radial component of the particle velocity

$$U_r = -\frac{1}{\mathcal{Z}}\frac{\partial}{\partial r}\left[\mathcal{P}^+_\omega \frac{e^{-sr/c}}{r} + \mathcal{P}^-_\omega \frac{e^{+sr/c}}{r}\right] = Y^+\mathcal{P}^+ - Y^-\mathcal{P}^-, \quad \texttt{[eq:Ur]} \tag{3.60}$$

corresponding to out and inbound velocity waves $\mathcal{U}^\pm_r = Y^\pm \mathcal{P}^\pm$. From the above definition the *acoustic radiation admittances* at $r = a$ for outbound (Y^+) and inbound (Y^-) waves are

$$Y^\pm(a,s) \equiv \frac{A_0 a^2}{\sqrt{\rho_0 \eta P_0}} \pm \frac{A_0 a^2}{s\rho_0 a}. \quad \texttt{[eq:Ypm]} \tag{3.61}$$

Here the first, real term which is the characteristic admittance $Y_0(a)$, corresponds to radiated (absorbed) energy, while the second, reactive complex term, to the stored kinetic energy (wave momentum). As depicted in Fig. 3.6, the sum of two admittances represent the parallel combination of impedances, in this case the characteristic resistance $Z_0 \equiv \rho_0 c/A_0 a^2$ and the acoustic mass $\pm\rho_0 a/A_0 a^2$ [Bauer, 1944, Salmon, 1946a,b, Morse, 1948]. The minus sign on the inbound wave U_r^- has been chosen to correspond to the direction of mass flow so that the real part of the radiation admittance remains positive for waves of both directions. It is obvious that a converging wave and a diverging wave cannot have the same impedance. This shows up as the converging wave having a negative mass, as indicated by the sign of Y^- in /EqYpm, corresponding to the inbound wave.

Setting the expression for the admittance to zero gives the frequencies of the poles $s_\pm(r) = \pm c/a$, which are related to the size of the acoustic wavelength relative to the size of the sphere. This is the same as $ka = 1$ where $k = 2\pi/\lambda$ is the wave number and λ is the wavelength. One may conclude that the reactance is related to sphereical *resonant scattering*.

This reactive term is well known in loud speaker design, as it explains why loudspeakers cannot radiate energy at low frequencies. This shunt reactance in the radiation impedance limits the power radiated for frequencies greater than the cutoff frequency $s_\pm$. This is the same for an electrical antenna smaller than the wavelength (e.g., $ka < 1$).

In the time domain, an inverse Laplace transform of the two admittances is

$$y_\pm(a,t) = Y_0\delta(t) \pm \frac{A_0 a}{\rho_0}\Delta(t) \leftrightarrow Y^\pm(a,s), \quad \texttt{[eq:ypm]} \tag{3.62}$$

where $\delta(t)$ is the Dirac impulse function and $\Delta(t)$ is its first integral, the unit step function. Since the velocity is the product of the admittance and the pressure, the time domain admittance $y^+(a,s)$ represents the radial particle velocity corresponding to an outbound wave pressure wave impulse, launched from a spherical radiator having radius a (e.g., Eq. 3.59).

In the case of a uniform tube, there is no mass reactance and the forward traveling pressure wave is simply $p^+\delta(t - x/c)$, with a corresponding velocity wave $v^+ = Z_0 p^+\delta(t - x/c)$.

The velocity wave, found by convolution of the pressure impulse Eq. 3.59 with the admittance $y_+(t)$, consists of a delta function plus a outbound constant velocity, that decreases with the radius r. The spherical wave on the other hand has a mass reactance (i.e. $s\rho_0 r$) in the wave radiation impedance. The outbound step function, being the inverse transform of the imaginary part of the admittance, represents the reactive storage of energy. This reactive part is the part not radiated, but corresponds to the velocity component that is delayed, relative to the pressure component, representing stored energy in the acoustic field.

The relations between pressure and velocity $[P, U_r]$ and the out and inbound pressure $[\mathcal{P}^+, \mathcal{P}^-]$, given at a, by Eq. 3.58 and Eq. 3.60, may be summarized in matrix form as, evaluated at the input $(r = a)$

$$\begin{bmatrix} P(a,\omega) \\ U_r(a,\omega) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ Y^+(a,s) & -Y^-(a,s) \end{bmatrix} \begin{bmatrix} \mathcal{P}^+(a,\omega) \\ \mathcal{P}^-(a,\omega) \end{bmatrix} \quad \texttt{[eq : ImpedanceWaves]} \tag{3.63}$$

with $\mathcal{P}^+$ being the outbound and $\mathcal{P}^-$ being the inbound d'Alembert (1746) waves for a spherical radiator of radius a, where $\mathcal{P}^\pm(a,\omega) = \mathcal{P}^\pm_\omega(a)e^{\mp sa/c}/a$ and the radiation admittances $Y^\pm(a,s)$ are given by Eq. 3.61.

3.3.3 Wave variables

A second order linear differential equation may be transformed into another set of basis functions. The natural basis set is the d'Alembert waves of Eq. 3.58, but other transformations are possible. In the following we shall transform Eq. 4.67 using *wave variables* [Fettweis, 1986].

Wave variables are linearly related to the pressure and velocity (Eq. 3.63) via the relations $P = P^+ + P^-$ and $U_r = U^+ - U^-$. In order that this linear transformation be unique, we further require that the wave variable ratios to be constrained such that they are equal to the characteristic impedance, which in general may depend on the positional coordinate, and if losses are considered, on the Laplace frequency s

$$\frac{P^\pm}{U_r^\pm} = Z_0(r,s) \equiv \sqrt{\frac{\mathcal{Z}(r,s)}{\mathcal{Y}(r,s)}}. \quad \texttt{[eq : z0]} \tag{3.64}$$

Defining the ratio of wave variables in this way leads to a unique linear transformation between *impedance variables* (i.e., the usual pressure and velocity variables $[P, U_r]$, the ratio of which define the impedance) and wave variables $[P^+, P^-]$. Eq. 3.64 makes them similar to localized plane waves. Wave variables (i.e., the localized plane waves) are special in that they characterize causal wave-fronts.

The uniqueness of the relations is proved by writing them in matrix form

$$\begin{bmatrix} P(r,\omega) \\ U_r(r,\omega) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ Y_0 & -Y_0 \end{bmatrix} \begin{bmatrix} P^+(r,\omega) \\ P^-(r,\omega) \end{bmatrix}, \quad \texttt{[eq : Wavepm1]} \tag{3.65}$$

and noting that the determinant (i.e., -$2y_0(r) = -2\sqrt{\mathcal{Y}/\mathcal{Z}}$) is non-zero. In homogeneous media, wave variables identically reduce to plane waves.

From the inverse of Eq. 3.66 we may determine the wave variables given $\mathcal{P}^\pm$ as

$$\begin{bmatrix} P^+(r,\omega) \\ P^-(r,\omega) \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & +Z_0 \\ 1 & -Z_0 \end{bmatrix} \begin{bmatrix} P(r,\omega) \\ U_r(r,\omega) \end{bmatrix}, \quad \texttt{[eq : Wavepm1]} \tag{3.66}$$

We view this transformation as a differential form of Weyl's famous integral expansion of the spherical wave in terms of plane waves. If we cascade Eq. 3.63 with

Need a ref to Wey
wave expansion of
Talk to Scott about

Eq. 3.66 we may directly write the wave variables in terms of the d'Alembert solutions $\mathcal{P}^{\pm}$

$$\begin{bmatrix} P^+(r,\omega) \\ P^-(r,\omega) \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & +Z_0 \\ 1 & -Z_0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ Y^+(a,s) & -Y^-(a,s) \end{bmatrix} \begin{bmatrix} \mathcal{P}^+(a,\omega) \\ \mathcal{P}^-(a,\omega) \end{bmatrix}, \quad \text{[eq : Wavepm2]} \tag{3.67}$$

This transformation represents a linear mapping from d'Alembert waves to that obey Eq. 3.64, similar to plane waves. The price for this much simpler impedance relationship (e.g., it is real when the medium is lossless), is that both types of wave variables are required to expand each of the d'Alembert waves.

As a specific example, when there is only an outbound wave (e.g., an infinite line driven by an impulse source), $\mathcal{P}^+_\omega = 1$, $\mathcal{P}^-_\omega = 0$ and $\mathcal{P}^+ = e^{-sr/c}/r \leftrightarrow \delta(t - r/c)/r$

$$P^{\pm}(r,\omega) = \frac{1 \pm Z_0 Y^+}{2} e^{-sr/c} \quad \text{[eq : Wavepm3]} \tag{3.68}$$

thus

$$P^{+}(r,\omega) = \left(1 + \frac{c}{2sa}\right) \frac{e^{-sr/c}}{r} \quad \text{[eq : Wavepm4]} \tag{3.69}$$

and

$$P^{-}(r,\omega) = -\frac{c}{2sa} \frac{e^{-sr/c}}{r} \quad \text{[eq : Wavepm5]} \tag{3.70}$$

resulting in a reflectance of

$$R(s) = -\frac{c}{s2a + c} \tag{3.71}$$

which is -1 for frequencies below resonance, and goes to zero as 1 over frequency, above.

3.4 Homework

Problem 3.1

Show that

$$\nabla_r^2\Phi \equiv \frac{1}{r^2}\frac{d}{dr}r^2\frac{d\Phi}{dr} = \frac{d^2\Phi}{dr^2} + \frac{2}{r}\frac{d\Phi}{dr} = \frac{1}{r}\frac{d^2r\Phi}{dr^2}$$

and discuss why this useful.

Problem 3.2

Identify the following people, the century they made their most significant contributions, and what that contribution was

- Newton Solution: Principia 1686: First calulation of the speed of sound. His answer was off however by $\sqrt{\eta}$.
- Laplace Solution: Pointed in 1822 out that the expansion and condensations take place adiabatically [Brillouin, 1953, Chapter I].
- Bernoulli (father and son) Solution: c1727-1753 made many contributions, including superposition required for a Fourier Series.
- Euler Solution: 1748 developed the first solutions of wave propagation on a string using partial differential equations.
- d'Alembert Solution: Showed the basic solutions of the wave equation 1747 of the from $f(x-t/c)+g(x+t/c)$. Brilloiuin attributes this to Euler for some reason, which doenst seem right.
- Lagrange Solution: 1759
- Cauchy Solution:
- Baden-Powell Solution: 1841 computed the velocity of a wave along a cubic lattice, and provided the first physical model, based on a 2-mass model, of a band-stop filter [Brillouin, 1953].
- Kelvin Solution: Was the first to point out that the lattice was a low-pass filter and had a cutoff frequency [Brillouin, 1953, Page 10]. 1881 "Popular Lectures"
- Fourier Solution: 1807 proved his Theorem that any function may be expanded in a series of sin and cos functions.
- Heaviside Solution: 1887 Important contributions to the theory of transmission lines. His work lead Pupin (1900) and Campbell (1906) to the loaded line, which made the telephone network practical.

- Pupin Solution: Inventor of the loaded line. This was somewhat controvirtal as Campbell many actually have been ahead of Pupin in the details, but Pupin was the first to the patent office.

- Campbell Solution: The inventor of the wave filter in 1911 which he first built around 1906.

- Born Solution: 1912 Max Born realistically modeled NaCl using a 3-dimensional lumped parameter mechanical model Brillouin [1953]. He also wrote a key book on optics still in use today Born and Wolf [1959].

- Brillouin Solution: Physicist from IBM research Laboratory, who generalized wave dispersion relations to 3 dimensions, which are now named in his honor. Following is a quote from Wikipedia 9/14/08:

  > Until 1942, Brillouin was a visiting professor at the University of Wisconsin at Madison, and then he was a professor at Brown University, in Providence, Rhode Island, until 1943. For the next two years, he was a research scientist with the National Defense Research Committee at Columbia University, working in the field of RADAR. From 1947 to 1949, he was professor of applied mathematics at Harvard University. During the period 1952 to 1954, he was with IBM Corporation in Poughkeepsie, New York, as well as a staff member of the IBM Watson Laboratory at Columbia University. In 1954, he became an adjunct professor at Columbia University. He lived in New York City until he died in 1969.

Problem 3.3

The pressure at the input of 10 cm of acoustic transmission line is a 1 [Pa] pure tone of 1 kHz having a source impedance of $2\rho_0 c/A_s$ where $A_s = \pi r_s^2$. At the end of the line is a 1 cc cavity. Define $x = L$ as the source end and $x = 0$ as the location of the cavity load.

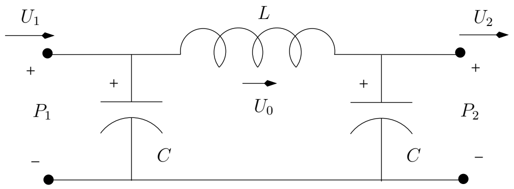
a. Find the pressure half way down the line (at 5 cm). Solution: Convert the pressure into a forward and backward waves, and then propagate the pressure using the formula $P^+(x) = e^{-\gamma x/c} P^+(0)$, which just says that the forward going wave is delayed as it propagates, as described by d'elamber (e.g., $p^+(t, t-x/c) = p^+(t, 0)$).

b.

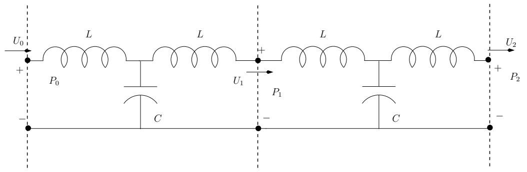
Problem 3.4

A transmission line is composed of series mass' and shunt springs. Find the frequency response (e.g., $H = U_2/U_1$). Assume that $C = 1$ and $L = 1$.

a. Single PI stage



b. Double PI stage



Problem 3.5

The *intensity* of an acoustic wave is the average product of its acoustic pressure, $p(t)$, times its acoustic air particle velocity, $u(t)$:

$$I = E\left[p(t)u(t)\right]$$

where $p(t)$ is measured in Pascals, and $u(t)$ is measured in m/s. What are the units of intensity? How does intensity relate to power?

Problem 3.6

The back of a loudspeaker is typically encased in a vented cabinet. A "vented" cabinet is a wooden box with the loudspeaker on one side, and a hole on the other side. The simplest model of a vented cabinet is a one-dimensional resonator, with the loudspeaker mounted at $x = -L$, and the vent open at $x = 0$. The boundary conditions are

$$\begin{aligned} U(x=-L,\Omega) &= -U_s(\Omega) \\ P(x=0,\Omega) &= 0 \end{aligned}$$

where $U_s(\Omega)$ is the loudspeaker velocity. Assume that the area of the loudspeaker and the area of the vent are both A. Then the total volume velocity radiating out of the cabinet is

$$Q_s(\Omega) = A(U_s(\Omega) + U(0, \Omega))$$

Find the "cabinet transfer function" $Q_s(\Omega)/U_s(\Omega)$. Sketch $Q_s(\Omega)/U_s(\Omega)$ for $0 \leq \Omega \leq 2\pi c/L$. Over what range of frequencies is $\frac{Q_s(\Omega)}{U_s(\Omega)} > 0$?

Problem 3.7

Glottal vibration is the result of two kinds of forces: aerodynamic forces, and stiffness and damping of the vocal fold. For now, let us only consider the stiffness of the vocal fold:

$$\frac{d^2x}{dt^2} = -\left(\frac{k}{m}\right)x \tag{3.72}$$

In acoustic terms, Eq. 4.63 governs the undriven, undamped, collision-free vocal fold behavior—undriven because we are ignoring aerodynamic forces, undamped because we are ignoring the viscosity of the tissue, and collision-free because we assume that the two vocal folds are far enough apart to avoid collision. Eq. 4.63 is a good place to start our understanding of vocal fold mechanics, because it isolates the "control knob" that most talkers use, most of the time, to control pitch: the stiffness, k, of the vocal fold. Stiffness of the vocal fold can be increased by stretching it; stiffness can be decreased by shortening the vocal fold.

a. Demonstrate that, with no driving forces and no damping and no collisions, the vocal folds can vibrate forever. Hint: show that $x(t) = A\cos(\omega t - \phi)$ is a solution to Eq. 4.63.

b. The moving part of the vocal fold is a ribbon of tissue about 1cm long, about 3mm deep, and about 1mm wide. This ribbon of tissue has the density of water (1 gram/cm^3). What is its total mass?

c. Suppose that a particular talker speaks at 200Hz. According to the model given in Eq. 4.63, what is the stiffness of her vocal folds? Be sure to tell me what the units are.

d. Assume that the vocal fold displacement, $x(t)$, has an amplitude of 1mm. Plot (by hand or using any program of your choice) one full period of the vocal fold displacement $x(t)$, of the vocal fold velocity dx/dt, and of the stiffness force $f(t) = -kx(t)$.

e. Now suppose that the same talker increases her pitch to 300Hz (an increase in pitch of one musical fifth), without changing her vocal fold mass. What is the new value of her vocal fold stiffness?

Problem 3.8

During production of the vowel /ɑ/, the pharynx is quite narrow (about 1cm^2), while the oral cavity is quite wide (about 8cm^2). Let the boundary between these two parts of the vocal tract be called $x = 0$.

a. Draw a schematic picture of this situation.

b. Pressure $p(x, t)$ (in Pascals) and volume velocity $u(x, t)$ (in liters/second) must be continuous across the boundary, i.e.

$$p(0_-, t) = p(0_+, t) \tag{3.73}$$

$$u(0_-, t) = u(0_+, t) \tag{3.74}$$

Re-write Eqs. 4.64 and 4.65 in terms of the forward-going and backward-going waves, whose phasors are p_{1+}, p_{1-}, p_{2+}, and p_{2-}.

c. Show that the outgoing waves from, p_{2+} and p_{1-}, may be written in terms of the incoming waves, p_{2-} and p_{1+}, and in terms of a reflection coefficient γ. Write γ in terms of the front cavity and back cavity areas.

d. Suppose that the glottis is a perfect source, i.e., regardless of what the backward-going wave p_{1-} may be, the forward-going wave is always a perfect cosine $p_{1+} = 1$. Find the forward-going and backward-going waves in the front cavity, p_{2+}, and p_{2-}, as a function of the front cavity length L_f, and the reflection coefficient γ. Assume a zero-pressure termination at the lips.

e. Find the air velocity at the lips, $v(L_f, \omega)$, as a function of L_f, ω, and γ. Assume that $p_{1+} = 1$ at all frequencies.

f. Plot $v(L_f, \omega)$ as a function of ω.

Problem 3.9

Assume a perfectly decoupled back and front cavity, where $A_b \gg A_f$. Assume that $L_b + L_f = 17$cm. Calculate the first three formant frequencies for the following back cavity lengths: $L_b \in \{1, 3, 5, 7, 9, 11, 13, 15\}$cm. Remember to consider the Helmholtz resonance. Plot F_1, F_2, and F_3 (in Hertz) on the same axes, as a function of L_b. This plot is called a “nomogram;” it is considered by many to be a convenient summary of the relationship between vocal tract shape and vowel quality.

Problem 3.10

The vocal tract configuration for an /r/ consonant, in American English, is roughly as follows: the back cavity, behind the tongue tip constriction, has a length of about 15cm, with a relatively large average cross-sectional area (about 10cm^2). The front cavity, between the tongue tip constriction and the lips, has a length of about 4cm, and a cross-sectional area of about 6cm^2. The side branch, under the tongue, has a length of about 4cm, and a cross-sectional area of about 16.5cm^2.

a. Sketch the three-tube model for /r/. Label all areas and lengths of all tube sections, and be sure to show whether each tube is closed or open.

b. Find F_1 of the /r/ configuration, using the low-frequency approximations given in the lecture notes.

c. Find F_1 of the neighboring vowel: set the length of the side branch to 0cm, and then solve the same equations that you solved in part (b). How does your answer compare to part (b)? How does your answer compare to the first formant of a schwa?

d. Find the first zero frequency of the /r/.

e. Assume that the first zero is part of a "pole-zero pair." In other words, you can imagine that the first zero splits the nearest oral formant into a pole-zero-pole complex, with the first pole about 200Hz below the zero, and the second pole about 200Hz above the zero. In that case, what is the frequency of F_3 of an /r/?

Problem 3.11

The speed of sound:

a. What is the formula for the speed of sound? Solution: The formula is $c = \sqrt{\eta P_0/\rho_0}$.

b. Identify the variables: Names, units, and values of consonants. Solution: On the left is the speed of sound c, which is a derived quantity, and on the right are physical parameters $\eta \equiv c_p/c_v = 1.4$ [dimentionless], the barometric pressure (e.g., at sea level) $P_0 = 10^5$ [Pa], and the density ρ_0 [kgm/m^3].

c. What is the meaning of ηP_0? Solution: This combination of variables represents the adiabatic compressibility of air. The $\eta = c_p/c_v$ results from holding the temperature constant during the cycle of the wave. Heat diffusion is slow compared to the cycle at acoustic frequencies, such that the thermal energy is trapped in the air.

d. Does P_0 depend on temperature? Explain? Solution: The barometric pressure depends on the weight of the air above us. While its density dependents on temperature, the total weight is a constant, and is therefore independent of temperature. That leaves the question as to why the barometric pressure varies, but that is more about the humidity, winds, and other variables.

e. Does ρ_0 depend on temperature or P_0? Explain. Solution: Yes, as given in class

$$\rho(T, P_0) = 1.29 \frac{273}{273 + C} \frac{P_0}{10^5}, \tag{3.75}$$

where C is the temperature in degrees Celsius.

f. What is the form of the dependence of the speed of sound on temperature? Namely give the formula for $c(T)$, and explain the dependence. Solution: Since $c = sqrt \eta P_0/\rho_0$ and following the state equation for a gas, $\rho_0 \propto 1/T$, we may conclude that $c(T) \propto \sqrt{T}$, where T is in degrees Kelvin. Because the percentage change in degrees Kelvin is much smaller than in degree C, this dependence is relatively small.

Problem 3.12

Basic equations of sound propagation:

a. Write out the 2x2 matrix equation that describes the propagation of 1 dimensional sound waves in a tube having area A: Solution: As shown in class the basic equations are:

$$\frac{d}{dx} \begin{bmatrix} P(x,\omega) \\ U(x,\omega) \end{bmatrix} = - \begin{bmatrix} 0 & \mathcal{Z}(x,s) \\ \mathcal{Y}(x,s) & 0 \end{bmatrix} \begin{bmatrix} P(x,\omega) \\ U(x,\omega) \end{bmatrix}. \tag{3.76}$$

where $\mathcal{Z} = s\rho_0/A$ and $\mathcal{Y} = sA/\eta P_0$, with A the area of the tube.

b. Rewrite these equations as a second order equation in terms of the pressure P, and thereby find the formula for the speed of sound in terms of Z and Y: Solution: If we let P' to mean the partial with respect to space, then

$$P' + ZU = 0 \tag{3.77}$$

and

$$U' + YP = 0. \tag{3.78}$$

Taking the partial wrt x of the first equation, and then using the second, gives

$$P'' + ZU' = P'' - \mathcal{Z}\mathcal{Y}P = 0. \tag{3.79}$$

Since the wave equation is

$$\frac{\partial^2 P}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 P}{\partial x^2} \tag{3.80}$$

by inspection we see that $\omega^2/c^2 = \mathcal{Z}\mathcal{Y}$, which results in the final formula for the speed of sound.

Problem 3.13

Assume that a flash bulb puts out 50 Joules when it is triggered, and it lasts for 20 μs. How much power is delivered while the flash is present, assuming the illumination is constant during the interval?

Problem 3.14

A person is speaking at an intensity of I_0 66 dB-SPL, as measured with a sound level meter at 1 meter.

a. Find the total power in the voice assuming that the level is uniform around the head. Solution: First, 66 dB SPL is equal to a pressure of $0.04 = 20 \times 10^{-6} 10^{66/20}$. This corresponds to an intensity (power/area) of $I_0 \equiv |P^2|/\rho_0 c = 0.04^2/407 = 3.93 \times 10^{-6}$ [Watts/m^2]. The area of the sphere at 1 meter is $4\pi r^2 = 4\pi$. Thus the total speech power is about $16\pi \approx 50$ [μW].

b. Find the total power assuming that the intensity varies as

$$I(\theta, \phi) = I_0 \cos(\theta/2)\cos(\phi/2) \tag{3.81}$$

where θ is the angle in the horizontal plane and ϕ in the vertical plane. where $\theta = 0, \phi = 0$ corresponds to "straight ahead."

Solution: The total power $\mathcal{P}$ is given by the integral of the intensity over the area:

$$\mathcal{P} = I_0 r^2 \times \int_{-\pi/2}^{\pi/2} \cos(\phi/2)\cos(\phi) d\phi \times \int_{-\pi}^{\pi} \cos(\theta/2)\cos(\theta) d\theta.$$

Thus

$$\mathcal{P} = r^2 I_0 \times [2\sin(\theta/2)]_{-\pi}^{\pi} \times \left[\frac{1}{3}\sin(3\phi/2) + \sin(\phi/2)\right]_{-\pi/2}^{\pi/2}$$

Total power is about 60% of part a.
As a check, make sure that when the area is uniform $\mathcal{P} = I_0 A$ with $A = 4\pi r^2 = 4\pi$.

Problem 3.15

Bels:

a. How many millibels [mB] in 1 bel [B]? Solution: 1000 mB = 1 B since mB is a much smaller unit that the bel.

b. Give the formula for the intensity in mB units. Solution: The intensity in mB is $1000\log_{10}(I/I_{ref})$ where $I_{ref} = 10^{-12}$ W/m^2.

c. Give the formula for the sound pressure level in cB (centibel) units. Solution: The sound pressure level in cB is $200\log_{10}(P/P_{ref})$ where $P_{ref} = 20 \times 10^{-6}$ [Pa].

Problem 3.16

Demonstrate that $P_{ref} \equiv 20$ μPa is the same as $I_{ref} \equiv 10^{-12}$ [W/m^2]. Solution: The first is given by the formula $|P_{ref}|^2/\rho c$ and the second is given by I_{ref}. Thus we need to show that numerically these two quantities are similar (almost identical). Since $(20 \times 10^{-6})^2/407 \approx 10^{-12}$, the two references are nearly the same.

Problem 3.17

A bottle has a neck diameter of 1 [cm] and is l =1 cm long. It is connected to the body of the bottle "barrel" which is 5 cm in diameter and L =10 cm long. Treat the barrel as a short piece of transmission line, closed at one end, which looks like a compliance $C = V_{barrel}/\eta P_0$, and the neck which look like a mass $M = \rho_0 l/A_{neck}$. These two impedances are in series, since they both see the same volume velocity (flow).

a. Find the resonant frequency of the bottle. Hint: set the impedance to zero and solve for the resonant frequency in terms of M and C.

b. write out the formula for the resonant frequency in terms of the dimensions of the bottle. Solution: The formula for the Helmholtz resonator was derived in class, where it was shown to be

$$f_0 = \frac{c}{2\pi}\sqrt{A/Vl}, \tag{3.82}$$

where A, l are the area and length of the neck and V is the volume of the bottle.

c. calculate the resonant frequency in Hz for the dimensions given. Solution: From the numbers given $A = \pi.005^2$ [m^2], $l = 0.01$ while $V = \pi 0.025^2 \times 0.1 = 1.9 \times 10^{-4}$ [m^3]. Thus $f_0 = \frac{345}{2\pi}\sqrt{0.01 \times \pi 0.005^2/0.0025} \approx 347.3$ [Hz].

d. Blow into a blottle and measure the resonant frequency by recording the tone, and taking the FFT of the resulting waveform, and finding the frequency (Extra credit).

Version: 0.50; March 30, 2011

Chapter 4

Speech Production

The collection of olfactory, respiratory and digestive apparatus which humans use for speaking is a relatively complex sound-producing system. Its operation has been described qualitatively in the preceding chapter. In this chapter we would like to consider in more detail the acoustical principles underlying speech production. The treatment is not intended to be exhaustive. Rather it is intended to circumscribe the problems of vocal tract analysis and to set forth certain fundamental relations for speech production. In addition, it aims to outline techniques and method for acoustic analysis of the vocal mechanism and to indicate their practical applications. Specialized treatments of a number of these points can be found elsewhere[1]

4.1 The Vocal Tract as an Acoustic System

The operations described qualitatively in the previous chapter can be crudely represented as in Fig. Fig. 4.1. The lungs and associated respiratory muscles are the vocal power supply. For voiced sounds, the expelled air causes the vocal folds to vibrate as a relaxation oscillator, and the air stream is modulated into discrete puffs or pulses. Unvoiced sounds are generated either by passing the air stream through a constriction in the tract, or by making a complete closure, building up pressure behind the closure and abruptly releasing it. In the first case, turbulent flow and incoherent sound are produced. In the second, a brief transient excitation occurs. The physical configuration of the vocal tract is highly variable and is dictated by the positions of the articulators; that is, the jaw, tongue, lips and velum. The latter controls the degree of coupling to the nasal tract.

In general, several major regions figure prominently in speech production. They are: (a) the relatively long cavity formed at the lower back of the throat in the

---

[1] For this purpose G. Fant (Fant [1960]), *Acoustic Theory of Speech Production*, is highly recommended. Besides presenting the acoustical bases for vocal analysis, this volume contains a wealth of data on vocal configurations and their calculated frequency responses. An earlier but still relevant treatise is Chiba and Kajiyama (Chiba and Kajiyama [1941]), *The Vowel; Its Nature and Structure.*

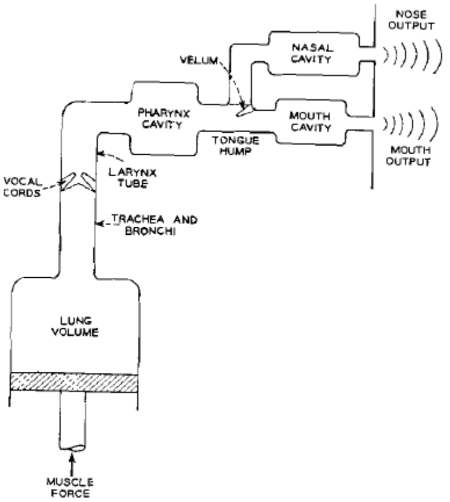


Figure 4.1: Schematic diagram of functional components of the vocal tract

pharynx region; (b) the narrow passage at the place where the tongue is humped; (c) the variable constriction of the velum and the nasal cavity; (d) the relatively large, forward oral cavity; (e) the radiating ports formed by the mouth and nostrils.

Voiced sounds are always excited at the same point in the tract, namely at the vocal folds. Radiation of voiced sounds can take place either from the mouth or nose, or from both. Unvoiced excitation is applied to the acoustic system at the point where turbulent flow or pressure release occurs. This point may range from an anterior position (such as the labio-dental excitation for /f/) to a posterior position (such as the velar excitation for /k/). Unvoiced sounds are normally radiated from the mouth. All sounds generated by the vocal apparatus are characterized by properties of the source of excitation and the acoustic transmission system. To examine these properties, let us first establish some elementary relations for the transmission system, then consider the sound sources, and finally treat the combined operation of sources and system.

The length of the vocal tract (about 17cm for adult males, about 15cm in adult females) is fully comparable to the wavelength of sound in air at audible frequencies. It is therefore not possible to obtain a precise analysis of the tract operation from a lumped-constant approximation of the major acoustic components. Wave motion in the system must be considered for frequencies above about 200Hz. The vocal and nasal tracts constitute lossy tubes of non-uniform cross-sectional area. Wave motion in such tubes is difficult to describe, even for lossless propagation. In fact, exact solutions to the wave equation are available only for two nonuniform geometries, namely for conical and hyperbolic area variations (Morse [1948]). And then only the conical geometry leads to a one-parameter wave.

So long as the greatest cross dimension of the tract is appreciably less than a wavelength (this is usually true for frequencies below about 5000Hz), and so long as the tube does not flare too rapidly (producing internal wave reflections), the acoustic system can be approximated by a one-dimensional wave equation. Such an equation assumes cophasic wave fronts across the cross-section and is sometimes called the Webster equation (Webster [1919]). Its form is

$$\frac{1}{A(x)}\frac{\partial}{\partial x}\left[A(x)\frac{\partial p}{\partial x}\right] = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{4.1}$$

where $A(x)$ is the cross-sectional area normal to the longitudinal dimension, p is the sound pressure (a function of t and x) and c is the sound velocity. In general this equation can only be integrated numerically, and it does not include loss. At least three investigations, however, have made use of this formulation for studying vowel production (Chiba and Kajiyama [1941], Heinz [1962]).

A more tractable approach to the analysis problem (both computationally and conceptually) is to impose a further degree of approximation upon the nonuniform tube. The pipe may be represented in terms of incremental contiguous sections of right circular geometry. The approximation may, for example, be in terms of cylinders, cones, exponential or hyperbolic horns. Although quantizing the area function introduces error, its effect can be made small if the lengths of the approximating sections are kept short compared to a wavelength at the highest frequency of interest. The uniform cylindrical section is particularly easy to treat and will be the one used for the present discussion.

4.2 The Radiation Load at the Mouth and Nostrils

At frequencies where the transverse dimensions of the tract are small compared with a wavelength, the radiating area of the mouth or nose can be assumed to have a velocity distribution that is approximately uniform and cophasic. It can therefore be considered a vibrating surface, all parts of which move in phase. The radiating element is set in a baffle that is the head. To a rough approximation, the baffle is spherical and about 9 cm in radius for an adult. Morse (Morse [1948]) has derived the radiation load on a vibrating piston set in a spherical baffle and shows it to be a function of frequency and the relative sizes of the piston and sphere. The analytical expression for the load is involved and cannot be expressed in closed form. A limiting condition, however, is the case where the radius of the piston becomes small compared with that of the sphere. The radiation load then approaches that of a piston in an infinite, plane baffle. The latter is well known and can be expressed in closed form. In terms of the normalized acoustic impedance

$$z = Z_A\left(\frac{A}{\rho c}\right) = \frac{p}{U}\left(\frac{A}{\rho c}\right)$$

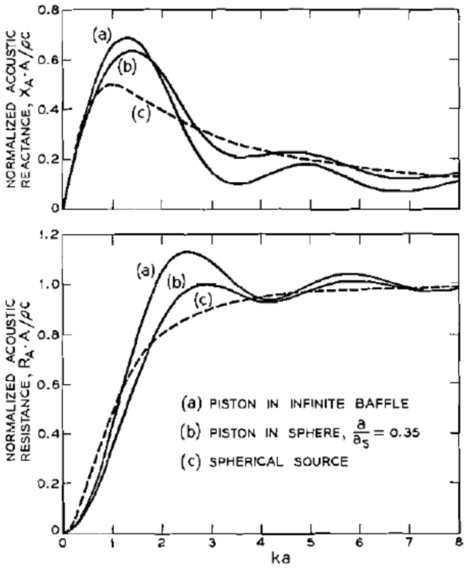


Figure 4.2: Normalized acoustic radiation resistance and reactance for (a) circular piston in all infinite baffle; (b) circular piston in a spherical baffle whose radius is approximately three times that of the piston; (c) pulsating sphere. The radius of the radiator, whether circular or spherical, is a

(that is, per-unit-free-space impedance), it is

$$z_p = \left[1 - \frac{J_1(2ka)}{ka}\right] + \left[\frac{K_1(2ka)}{2(ka)^2}\right], \tag{4.2}$$

where $k = \omega/c$, a is the piston radius, A the piston area, $J_1(x)$ the first order Bessel function, and $K_1(x)$ a related Bessel function given by the series

$$K_1(x) = \frac{2}{\pi}\left[\frac{x^3}{3} - \frac{x^5}{3^2 \cdot 5} + \frac{x^7}{3^2 \cdot 5^2 \cdot 7} \cdots\right].$$

For small values of ka, the first terms of the Bessel functions are the most significant, and the normalized radiation impedance is approximately

$$z_p \approx \frac{(ka)^2}{2} + j\frac{8(ka)}{3\pi}; \quad ka \ll 1 \tag{4.3}$$

This impedance is a resistance proportional to ω^2 in series with an inductance of normalized value $8a/3\pi c$. The parallel circuit equivalent is a resistance of $128/9\pi^2$ in parallel with an inductance of $8a/3\pi c$.

By way of comparison, the normalized acoustic load on a vibrating sphere is also well known and is

$$z_s = \frac{jka}{1 + jka}, \tag{4.4}$$

where a is the radius of the sphere. Note that this is the parallel combination of a unit resistance and an a/c inductance. Again, for small ka,

$$z_s \approx (ka)^2 + j(ka); \quad ka \ll 1. \tag{4.5}$$

Using Morse's results for the spherical baffle, a comparison of the real and imaginary parts of the radiation impedances for the piston-insphere, piston-in-wall, and pulsating sphere is made in Fig. Fig. 4.2. For the former, a piston-to-sphere radius ratio of $a/a_s = 0.35$ is illustrated. The piston-in-wall curves correspond to $a/a_s = 0$. For $ka < l$, one notices that the reactive loads are very nearly the same for all three radiators. The real part for the spherical source is about twice that for the pistons.

These relations can be interpreted in terms of mouth dimensions. Consider typical extreme values of mouth area (smallest and largest) for vowel production. An adult articulating a rounded vowel such as /u/ produces a mouth opening on the order of 0.9 cm^2. For an open vowel such as /ɑ/ an area of 5.0 cm^2 is representative. The radii of circular pistons with these areas are 0.5 cm and 1.3 cm, respectively. For frequencies less than about 5000Hz, these radii place ka less than unity. If the head is approximated as a sphere of 9 cm radius, the ratios of piston-to-sphere radii for the extreme areas are 0.06 and 0.1, respectively. For these dimensions and frequencies, therefore, the radiation load on the mouth is not badly approximated by considering it to be the load on a piston in an infinite wall. The approximation is even better for the nostrils whose radiating area is smaller. For higher frequencies and large mouth areas, the load is more precisely estimated from the piston-insphere relations. Notice, too, that approximating the normalized mouthradiation load as that of a pulsating sphere leads to a radiation resistance that is about twice too high.

4.3 Spreading of Sound about the Head

In making acoustic analyses of the vocal tract one usually determines the volume current delivered to the radiation load at the mouth or nostrils. At these points the sound energy is radiated and spreads spatially. The sound is then received by the ear or by a microphone at some fixed point in space. It consequently is desirable to know the nature of the transmission from the mouth to the given point.

The preceding approximations for the radiation impedances do not necessarily imply how the sound spreads about the head. It is possible for changes in the baffling of a source to make large changes in the spatial distribution of sound and yet produce relatively small changes in the radiation load. For example, the piston-in-wall and piston-insphere were previously shown to be comparable assumptions for the radiation load. Sound radiated by the former is of course confined to the half-space, while that from the latter spreads spherically. The lobe structures are also spatially different.

One might expect that for frequencies where the wavelength is long compared with the head diameter, the head will not greatly influence the field. The spatial

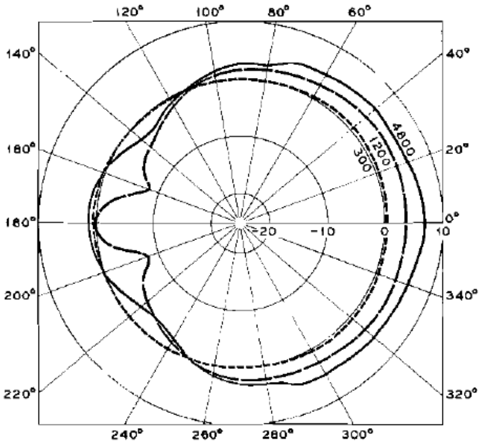


Figure 4.3: Spatial distributions of sound pressure for a small piston in a sphere of 9cm radius. Pressure is expressed in db relative to that produced by a simple spherical source of equal strength

spreading of sound should be much like that produced by a simple spherical source of strength equal to the mouth volume velocity. At high frequencies, however, the diffraction about the head might be expected to influence the field.

A spherical source, pulsating sinusoidally, produces a particle velocity and sound pressure at r distance from its center equal respectively to

$$u(r) = \frac{au_0}{r} \frac{jka}{1+jka} \frac{1+jkr}{jkr} e^{-jk(r-a)},$$

and

$$p(t) = \frac{\rho c a u_0}{r} \frac{jka}{1+jka} e^{-jk(r-a)} \tag{4.6}$$

where a is the radius, u_0 is the velocity magnitude of the surface, and $k = \omega/c$. (Note the third factor in $u(r)$ accounts for the "bass-boost" that is obtained by talking close to a velocity microphone, a favorite artifice of nightclub singers.) If $ka \approx 1$, the source is a so-called simple (point) source, and the sound pressure is

$$p(r) = \frac{j\omega\rho U_0}{4\pi r} e^{-jkr} \tag{4.7}$$

where $U_0 = 4\pi a^2 u_0$ is the source strength or volume velocity. The simple source therefore produces a sound pressure that has spherical symmetry and an amplitude that is proportional to l/r and to ω.

Morse (Morse [1948]) has derived the pressure distribution in the far field of a small vibrating piston set in a spherical baffle. Assuming that the mouth and head are approximately this configuration, with a 9 cm radius roughly appropriate for the sphere, the radiation pattern can be expressed relative to that which would be

Figure 4.4: Life-size mannequin for measuring the relation between the mouth volume velocity and the sound pressure at an external point. The transducer is mounted in the mannequin's head.

produced by a simple source of equal strength located at the same position. When this is done, the result is shown in Fig. Fig. 4.3. If the pressure field were identical to that of a simple spherical source, all the curves would fall on the zero db line of the polar plot. The patterns of Fig. Fig. 4.3 are symmetrical about the axis of the mouth (piston) which lies at zero degrees. One notices that on the mouth axis the high frequencies are emphasized slightly more than the +6 dB/octave variation produced by the simple source (by about another +2 dB/octave for frequencies greater than 300 Hz). Also some lobing occurs, particularly at the rear of the head.

The question can be raised as to how realistic is the spherical approximation of the real head. At least one series of measurements has been carried out to get a partial answer and to estimate spreading of sound about an average life-sized head (Flanagan [1960a]). A sound transducer was fitted into the head of the adult mannequin shown in Fig. Fig. 4.4. The transducer was calibrated to produce a known acoustic volume velocity at the lips of the dummy, and the amplitude and phase of the external pressure field were measured with a microphone. When the amplitudes are expressed relative to the levels which would be produced by a simple source of equal strength

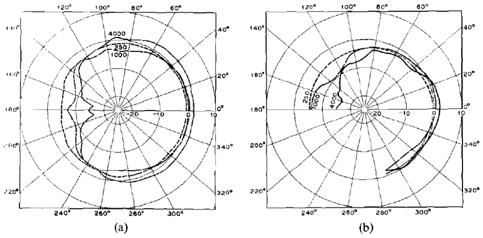


Figure 4.5: Distribution of sound pressure about the head, relative to the distribution for a simple source; (a) horizontal distribution for the mannequin; (b) vertical distribution for the mannequin

located at the mouth, the results for the horizontal and vertical planes through the mouth are shown in Fig. Fig. 4.5.

One notices that for frequencies up to 4000 Hz, the pressures within vertical and horizontal angles of about 60 degrees, centered on the mouth axis, differ from the simple source levels by no more than 3 db. Simultaneous phase measurements show that within this same solid angle, centered on the mouth axis, the phase is within approximately 30 degrees of that for the simple source. Within these limits, then, the function relating the volume velocity through the mouth to the sound pressure in front of the mouth can be approximated as the simple source function of Eq.(Eq. 4.7). Notice that $p(r)/U_0 \sim \omega$, and the relation has a spectral zero at zero frequency.

4.4 The Source for Voiced Sounds

4.4.1 Glottal Excitation

The nature of the vocal tract excitation for voiced sounds has been indicated qualitatively in Figs. Fig. 2.1 through Fig. 2.4. It is possible to be more quantitative about this mechanism and to estimate some of the acoustical properties of the glottal sound source. (The glottis, as pointed out earlier, is the orifice between the vocal folds.) Such estimates are based mainly upon a knowledge of the subglottal pressure, the glottal dimensions, and the time function of glottal area.

TO DO: Provide equations and intuition for the Ishizaka-Flanagan two-mass model of vocal fold vibration (Ishizaka and Flanagan [1972b,a], Flanagan *et al.* [1975]), following up the description in chapter 2.

4.4.2 Sub-Glottal Impedance

The principal physiological components of concern are illustrated schematically in Fig. Fig. 4.6. The diagram represents a front view of the subglottal system. The dimensions are roughly appropriate for an adult male (Judson and Weaver [1942]).

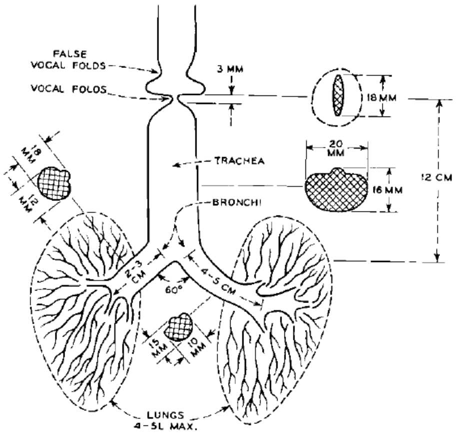


Figure 4.6: Schematic diagram of the human subglottal system

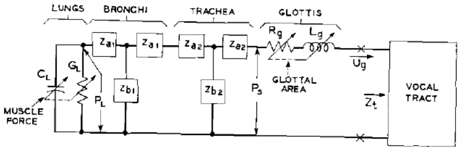


Figure 4.7: An equivalent circuit for the subglottal system

In terms of an electrical network, this system might be thought analogous to the circuit shown in Fig. Fig. 4.7.

A charge of air is drawn into the lungs and stored in their acoustic capacity C_L. The lungs are spongy tissues and exhibit an acoustic loss represented by the conductance G_L. The loss is a function of the state of inflation. The muscles of the rib cage apply force to the lungs, raise the lung pressure P_L, and cause air to be expelled–via the bronchi and trachea–through the relatively small vocal cord orifice. (Recall Fig. Fig. 4.1.) Because of their mass and elastic characteristics, the folds are set vibrating by the local pressure variations in the glottis. The quasiperiodic opening and closing of the folds varies the series impedance $(R_g + jwL_g)$ and modulates the air stream. The air passing into the vocal tract is therefore in the form of discrete puffs or pulses. As air is expelled, the rib-cage muscles contract and tend to maintain a constant lung pressure for a constant vocal effort. The lung capacity is therefore reduced so that the ratio of air charge to capacity remains roughly constant.

The bronchial and tracheal tubes–shown as equivalent T-sections in Fig. Fig. 4.7–are relatively large so that the pressure drop across them is small[2]. The subglottal pressure P, and the lung pressure P_L are therefore nearly the same. The variable-area glottal orifice is the time-varying impedance across which most of the subglottic

---

[2] The branching bronchi are represented as a single tube having a cross-sectional area equal to the sum of the areas of the branches.

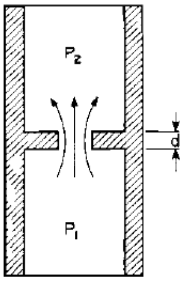


Figure 4.8: Simple orifice approximation to the human glottis,

pressure is expended. The subglottal potential is effectively converted into kinetic energy in Ihe form of the glottal volume velocity pulses, U_g.

TO DO: Describe models of the sub-glottal impedance by (Fant *et al.* [1972]), (Ishizaka *et al.* [1976]), and by (Cranen and Boves [1987]). Provide spectral examples showing sub-glottal formants.

4.4.3 Glottal Impedance

For frequencies less than a couple of thousand Hertz, the main component of the glottal impedance is the resistive term. For many purposes in vocal tract analysis, it is convenient to have a small-signal (ac) equivalent circuit of the glottal resistance; that is, a Thevenin equivalent of the circuit to the left of the X's in Fig. Fig. 4.7. Toward deducing such an equivalent, let us consider the nature of the time-varying glottal impedance and some typical characteristics of glottal area and volume flow.

To make an initial estimate of the glottal impedance, assume first that the ratio of the glottal inertance to resistance is small compared to the period of area variation (that is, the L_g/R_g time constant is small compared with the fundamental period, T). We will show presently the conditions under which this assumption is tenable. For such a case, the glottal volume flow may be considered as a series of consecutively established steady states, and relations for steady flow through an orifice can be used to estimate the glottal resistance.

Flow through the vocal cord orifice in Fig. Fig. 4.6 can be approximated as steady, incompressible flow through the circular orifice shown in Fig. Fig. 4.8. The subglottal and supraglottal pressures are P_1, and P_2, respectively. The particle velocity in the port is u, the orifice area is A and its depth (thickness) is d. If the cross-sectional areas of the adjacent tubes are much larger than A, variations in P_1 and P_2 caused by the flow are small, and the pressures can be assumed sensibly constant. Also, if the dimensions of the orifice are small compared with the wavelength of an acoustic disturbance, and if the mean flow is much smaller than the speed of sound, an acoustic disturbance is known essentially instantaneously throughout the vicinity of the orifice, and incompressibility is a valid assumption. Further, let it be assumed that the velocity distribution over the port is uniform and that there is no viscous dissipation.

Under these conditions, the kinetic energy per-unit-volume possessed by the air in the orifice is developed by the pressure difference $(P_1 - P_2)$ and is

$$(P_1 - P_2) = \frac{\rho u^2}{2}. \tag{4.8}$$

The particle velocity is therefore

$$u = \left[\frac{2(P_1 - P_2)}{\rho}\right]^{1/2} \tag{4.9}$$

We can define an orifice resistance, R_g^*, as the ratio of pressure drop to volume flow

$$R_g^* = \frac{\rho u}{2A} = \frac{\rho U}{2A^2}. \tag{4.10}$$

where $U = uA$ is the volume velocity. In practice, P_2 is essentially atmospheric pressure, so that $(P_1 - P_2) = P_s$ the excess subglottal pressure, and

$$R_g^* = \frac{(2\rho P_s)^{1/2}}{2A}. \tag{4.11}$$

In situations more nearly analogous to glottal operation, the assumptions of uniform velocity distribution across the orifice and negligible viscous losses are not good. The velocity profile is generally not uniform, and the streamlines are not straight and parallel. There is a contraction of the jet a short distance downstream where the distribution is uniform and the streamlines become parallel (vena contracta). The effect is to reduce the effective area of the orifice and to increase R_g^*. Also, the pressure-to-kinetic energy conversion is never accomplished without viscous loss, and the particle velocity is actually somewhat less than that given in (Eq. 4.9). In fact, if the area and flow velocity are sufficiently small, the discharge is actually governed by viscous laws. This can certainly obtain in the glottis where the area of opening can go to zero. Therefore, an expression for orifice resistance–valid also for small velocities and areas–might, as a first approximation, be a linear combination of kinetic and viscous terms

$$R_g = R_v + k\left(\frac{\rho U}{2A^2}\right), \tag{4.12}$$

where R_v is a viscous resistance and k is a real constant. For steady laminar flow, R_v is proportional to the coefficient of viscosity and the length of the conducting passage, and is inversely proportional to a function of area.

To find approximations of the form (Eq. 4.12), Wegel (Wegel [1930]) and van den Berg et al.(van den Berg [1955]) have made steady-flow measurements on models of the human larynx. Both investigations give empirical formulas which agree in order of magnitude. Van den Berg's data are somewhat more extensive and were made on

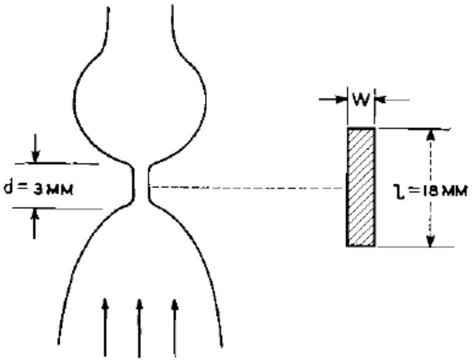


Figure 4.9: Model of the human glottis. (After Berg (van den Berg [1955]))

plaster casts of a normal larynx. The glottis was idealized as a rectangular slit as shown in Fig. Fig. 4.9. The length, l, of the slit was maintained constant at 18 mm, and its depth, d, was maintained at 3 mm. Changes in area were made by changing the width, w. Measurements on the model show the resistance to be approximately

$$R_g = \frac{P_s}{U} = \frac{12\mu d}{lw^3} + 0.875\frac{\rho U}{2(lw)^2}, \tag{4.13}$$

where μ is the coefficient of viscosity. According to van den Berg, (Eq. 4.13) holds within ten per cent for $0.1 \leq w \leq 2.0$mm, for $P_s \leq 64$ cm H_20 at small w, and for $U \ll 2000$ cc/sec at large w. As (Eq. 4.13) implies, values of P, and A specify the volume flow, U.

The glottal area is $A = lw$ so that the viscous (first) term of (Eq. 4.13) is proportional to A^{-3}. The kinetic (second) term is proportional to uA^{-1} or, to the extent that u can be estimated from (Eq. 4.9), it is approximately proportional to $P_s^{1/2}A^{-1}$. Whether the viscous or kinetic term predominates depends upon both A and P_s. They become approximately equal when $(\rho P_s)^{1/2}A^2 = 19.3\mu dl^2$. For typical values of vocal P_s this equality occurs for glottal areas which generally are just a fraction (usually less than $\frac{1}{5}$) of the maximum area. In other words, over most of the open cycle of the vocal folds the glottal resistance is determined by the second term in (Eq. 4.13).

As pointed out previously, (Eq. 4.13) is strictly valid only for steady flow conditions. A relevant question is to what extent might (Eq. 4.13) be applied in computing the glottal flow as a function of time when $A(t)$ and P_s are known. The question is equivalent to inquiring into the influence of the inertance of the glottal air plug. Because the pressure drop across the bronchi and trachea is small, and because P_s is maintained sensibly constant over the duration of several pitch periods by the low-impedance lung reservoir[3], the circuit of Fig. Fig. 4.7 can, for the present purpose,

---

[3] Van den Berg et al. estimate the variation to be less than five per cent of the mean subglottal pressure. P_s was measured by catheters inserted in the trachea and esophagus.

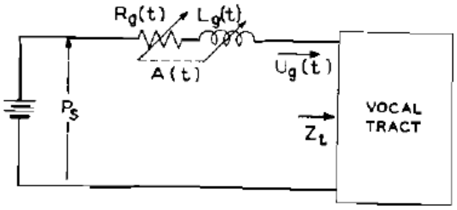


Figure 4.10: Simplified circuit for the glottal source

be simplified to that shown in Fig. Fig. 4.10. Furthermore, it is possible to show that at most frequencies the driving point impedance of the vocal tract, Z_t, is small compared with the glottal impedance. If the idealization $Z_t = 0$ is made, then $U_g(t)$ satisfies

$$U_g(t)R_g(t) + \frac{d}{dt}\left[L_g(t)U_g(t)\right] = P_s \tag{4.14}$$

where Eq. (Eq. 4.13) can be taken as the approximation to $R_g(t)$ and, neglecting end corrections, $L_g(t) = \rho d/A(t)$.

Because R_g is a flow-dependent quantity, Eq. (Eq. 4.14) is a nonlinear, first-order differential equation with nonconstant coefficients. For an arbitrary $A(t)$, it is not easily integrated. However, a simplification in the area function provides some insight into the glottal flow. Consider that $A(t)$ is a step function so that

$$\begin{aligned} A(t) &= A_0; \quad t \geq 0 \\ &= 0; \quad t < O, \text{ and } U_g(0) = 0. \end{aligned}$$

Then dL_g/dt is zero for $t > 0$, and the circuit acts as a flow-dependent resistance in series with a constant inductance. A step of voltage (P_s) is applied at $t = 0$. The behavior of the circuit is therefore described by

$$\frac{dU_g}{dt} = \frac{1}{L_g}\left(P_s - R_g U_g\right). \tag{4.15}$$

At $t = 0$, $U_g(0) = 0$ and

$$\left.\frac{dU_g}{dt}\right|_{t=0} = \frac{P_s}{L_g},$$

so that initially

$$U_g(t) \approx \frac{P_s}{L_g}t \quad \text{(for positive t near zero).}$$

Similarly, at $t = \infty$, $dU_g/dt = 0$ and $U_g(\infty) = P_s/R_g$. The value of $U_g(\infty)$ is the steady-flow value which is conditioned solely by R_g. In this case U_g is the solution of $P_s - U_g R_g = 0$, and is the positive root of a seconddegree polynominal in U_g.

A time constant of a sort can be estimated from these asymptotic values of the flow build-up. Assume that the build-up continues at the initial rate, P_s/L_g, until the

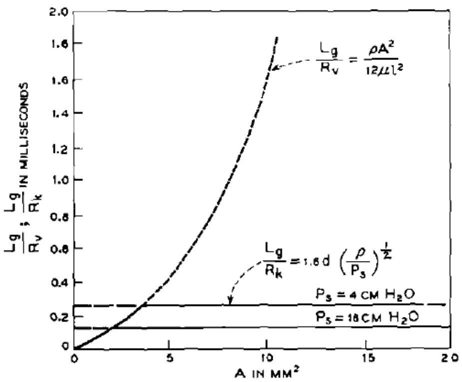


Figure 4.11: Ratios of glottal inertance (L_g) to viscous and kinetic resistance (R_v, R_k) as a function of glottal area (A)

steady-state value $U_g(\infty)$ is achieved. The time, T, necessary to achieve the build-up is then

$$U_g(t) = \frac{P_s}{L_g}T = U_g(\infty) = \frac{P_s}{L_g},$$

or

$$T = \frac{L_g}{R_g}. \tag{4.16}$$

Since R_g is a sum of viscous and kinetic terms R_v and R_k, respectively, the time constant $L_g/(R_v + R_k)$ is smaller than the smaller of L_g/R_v and L_g/R_k. If the step function of area were small, R_v would dominate and the L_g/R_v time constant, which is proportional to A^2, would be more nearly appropriate. If the area step were large, the L_g/R_k constant would apply. In this case, and to the extent that R_v might be neglected (i.e., to the extent that R_g might be approximated as $R_k = 0.875(2\rho P_s)^{1/2}/2A$), the L_g/R_k constant is proportional to $P_s^{-\frac{1}{2}}$ and is independent of A.

On the basis of these assumptions, a plot of the factors L_g/R_v and L_g/R_k is given in Fig. Fig. 4.11. Two values of P_s are shown for L_g/R_k, namely 4 cm H_20 and 16 cm H_20. The first is approximately the minimum (liminal) intensity at which it is possible to utter a vowel. The latter corresponds to a fairly loud utterance or shout. The value of L_g/R_g is therefore less than the solid curves of Fig. Fig. 4.11.

The curves of Fig. Fig. 4.11 show the greatest value of the time constant (i.e., for liminal subglottic pressure) to be of the order of a quarter millisecond. This time might be considered negligible compared with a fundamental vocal cord period an order of magnitude greater, that is, 2.5 msec. The latter corresponds to a fundamental vocal frequency of 400 Hz which is above the average pitch range for an adult male or female voice, but which might be reasonable for a child. To a first order approximation, therefore, the waveform of glottal volume velocity can be estimated from P_s and $A(t)$ simply by applying (Eq. 4.13).

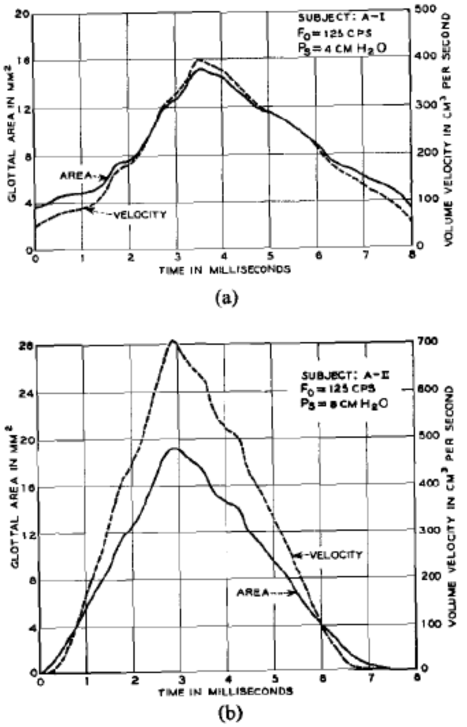


Figure 4.12: Glottal area and computed volume velocity waves for single vocal periods. F_0 is the fundamental frequency: P_s is the subglottal pressure. The subject is an adult male phonating /æ/. (After Flanagan, 1958 (Flanagan [1958]))

Notice also from the preceding results that for $L_g/R_g \approx 0.25$ms (i.e., $P_s \approx 4$cm H_20) the inductive reactance becomes comparable to the resistance for frequencies between 600 and 700 Hz. For $P_s = 16$cm H_20, the critical frequency is about doubled, to around 1300 Hz. This suggests that for frequencies generally greater than about 1000 to 2000 Hz, the glottal impedance may exhibit a significant frequency-proportional term, and the spectrum of the glottal volume flow may reflect the influence of this factor.

If the effects of inertance are neglected, a rough estimate of the glottal volume velocity can be made from the resistance expression (Eq. 4.13). Assuming constant subglottal pressure, the corresponding volume velocity is seen to be proportional to A^3 at small glottal areas and to A at larger areas. Typical volume velocity waves deduced in this manner are shown in Fig. Fig. 4.12 (Flanagan [1958]). The area waves are measured from high speed motion pictures of the glottis (see Fig. Fig. 2.3 in Chapter 2), and the subglottal pressure is estimated from the sound intensity and direct tracheal pressure measurements. The first condition is for the vowel /æ/ uttered at the lowest intensity and pitch possible. The second is for the same sound at a louder intensity and the same pitch. In the first case the glottis never completely

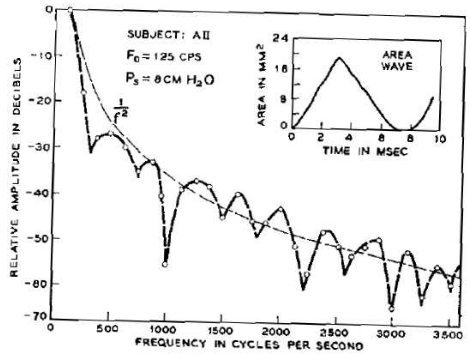


Figure 4.13: Calculated amplitude spectrum for the glottal area wave AII shown in Fig. Fig. 4.12. (After Flanagan, 1961 (Flanagan [1961]))

closes. This is characteristic of weak, voiced utterances. Note that the viscous term in R_g operates to sharpen the leading and trailing edges of the velocity wave. This effect acts to increase the amplitude of the high-frequency components in the glottal spectrum.

The spectrum of the glottal volume flow is generally irregular and is characterized by numerous minima, or spectral zeros. For example, if the wave in Fig. Fig. 4.12b were idealized as a symmetrical triangle, its spectrum would be of the form $(\sin x/x^2)$ with double-order spectral zeros occurring for $\omega = 4n\pi/\tau_0$, where n is an integer and T_0 is the open time of the glottis. If the actual area wave of Fig. Fig. 4.12b is treated as periodic with period 1/125 sec, and its Fourier spectrum computed (most conveniently on a digital computer), the result is shown in Fig. Fig. 4.13 (Flanagan [1961])). The slight asymmetry of the area wave causes the spectral zeros to lie at complex frequencies, so that the spectral minima are neither equally spaced nor as pronounced as for the symmetrical triangle.

4.4.4 Source-Tract Coupling Between Glottis and Vocal Tract

Considering only the resistance R_g given in Eq. (Eq. 4.13), it is possible to approximate an ac or small-signal equivalent source for the glottal source. Such a specification essentially permits the source impedance to be represented by a time-invariant quantity and is useful in performing vocal tract calculations. The Thevenin (or Norton) equivalent generator for the glottis can be obtained in the same manner that the ac equivalent circuit for an electronic amplifier is derived. According to (Eq. 4.13)

$$U_g(t) = f(P_s, A).$$

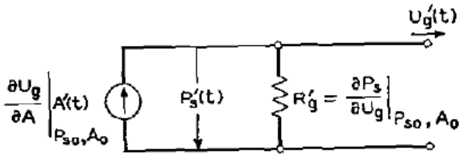


Figure 4.14: Small-signal equivalent circuit for the glottal source. (After Flanagan, 1958 (Flanagan [1958]))

The glottal volume velocity, area and subglottic pressure are unipolar time functions. Each has a varying component superposed upon a mean value. That is,

$$\begin{aligned} U_g(t) &= U_{g0} + U'(t) \\ A(t) &= A_0 + A'(t) \\ P_s(t) &= P_{s0} + P_s'(t). \end{aligned}$$

Expanding $U_g(t)$ as a Taylor series about (P_{s0}, A_0) and taking first terms gives

$$\begin{aligned} U_g(P_s, A) &= U_g(P_{s0}, A_0) + \frac{\partial U_g}{\partial P_s}\bigg|_{P_{s0},A_0} (P_s - P_{s0}) + \frac{\partial U_g}{\partial A}\bigg|_{P_{s0},A_0} (A - A_0) + \cdots, \\ &= U_{g0} + U_g'(t), \end{aligned}$$

and

$$U_g'(t) = \frac{\partial U_g}{\partial P_s}\bigg|_{P_{s0},A_0} P_s' + \frac{\partial U_g}{\partial A}\bigg|_{P_{s0},A_0} A'(t). \tag{4.17}$$

One can interpret (Eq. 4.17) as an ac volume velocity (current) source of value $\partial U_g/\partial A|_{P_{s0},A_0} A'(t)$ with an inherent conductance $\partial U_g/\partial P_s|_{P_{s0},A_0}$. The source delivers the ac volume current $U_g'(t)$ to its terminals. The source configuration is illustrated in Fig. Fig. 4.14. The instantaneous polarity of $P_s'(t)$ is reckoned as the pressure beneath the glottis relative to that above.

The partials in (Eq. 4.17) can be evaluated from (Eq. 4.13). Let

$$R_g' = \frac{\partial P_s}{\partial U_g}\bigg|_{P_{s0},A_0}.$$

Then

$$\frac{\partial P_s}{\partial U_g} = R_g + U_g \frac{\partial R_g}{\partial U_g},$$

and

$$R_g' = (R_v + 2R_k)_{P_{s0},A_0} \tag{4.18}$$

The magnitude of the equivalent velocity source is simply

$$\frac{\partial U_g}{\partial A}\bigg|_{P_{s0},A_0} A'(t) = \left[u + A\frac{\partial u}{\partial A}\right]_{P_{s0},A_0} A'(t).$$

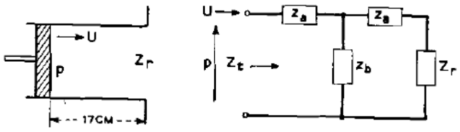


Figure 4.15: Simplified representation of the impedance looking into the vocal tract at the glottis

Neglecting the viscous component of the resistance, Eq. (Eq. 4.9) may be used to approximate u, in which case $\partial u/\partial A = 0$ and

$$\left.\frac{\partial U_g}{\partial A}\right|_{P_{s0},A_0} \approx \left(\frac{2P_{s0}}{\rho}\right)^{1/2} A'(t) \tag{4.19}$$

The approximations (Eq. 4.18) and (Eq. 4.19) therefore suggest that the ac resistance of the glottal source is equal the viscous (first) term of (Eq. 4.13) plus twice the kinetic (second) term, and that the ac volume current source has a waveform similar to the time-varying component of $A(t)$. To consider a typical value of R'_g, take $P_{s0} = 10\text{cmH}_2 0$ and $A_0 = 5$ mm^2. For these commonly encountered values R'_g is computed to be approximately 100 cgs acoustic ohms. This source impedance can be compared with typical values of the acoustic impedance looking into the vocal tract (i.e., the tract driving point impedance). Such a comparison affords an insight into whether the glottal source acts more nearly as a constant current (velocity) generator or a voltage (pressure) source.

The driving point impedance of the tract is highly dependent upon vocal configuration, but it can be easily estimated for the unconstricted shape. Consider the tract as a uniform pipe, 17 cm long and open at the far end. Assuming no nasal coupling, the tract is terminated only by the mouth radiation impedance. The situation is illustrated in Fig. Fig. 4.15.

Using the transmission line relations developed earlier in the chapter, the impedance Z_t looking into the straight pipe is

$$Z_t = Z_0 \frac{Z_r \cosh \gamma l + Z_0 \sinh \gamma l}{Z_0 \cosh \gamma l + Z_r \sinh \gamma l}, \tag{4.20}$$

where $l = 17$cm, and the other quantities have been previously defined. If for a rough estimate the pipe is considered lossless, $\gamma = j\beta$ and (Eq. 4.20) can be written in circular functions

$$Z_t = Z_0 \frac{Z_r \cos \beta l + jZ_0 \sin \beta l}{Z_0 \cos \beta l + jZ_r \sin \beta l}, \tag{4.21}$$

where $Z_0 = \rho c/A$, $\beta = \omega/c$. The maxima of Z_t will occur at frequencies where $l = (2n+1)\lambda/4$, so that $\beta l = (2n+1)\pi/2$ and $\cos \beta l O$. The maxima of Z_t for the lossless pipe are therefore

$$Z_{t_{max}} = Z_0^2/Z_r, \tag{4.22}$$

and the pipe acts as a quarter-wave transformer. The minima, on the other hand, are $Z_{t_{min}} = Z_r$ and the pipe acts as a half-wave transformer.

To estimate $Z_{t_{max}}$ we can use the radiation impedance for the piston in the infinite baffle, developed earlier in the chapter [see Eq. (Eq. 4.3)].

$$Z_r = z_p \frac{\rho c}{A} = \frac{\rho c}{A}\left[\frac{(ka)^2}{2} + j\frac{8}{3\pi}(ka)\right], \tag{4.23}$$

where

$$a = \sqrt{A/\pi}, \quad \text{and} \quad ka \ll 1.$$

As a reasonable area for the unconstricted tract, take $A = 5\text{cm}^2$. The first quarter-wave resonance for the 17cm long pipe occurs at a frequency of about 500 Hz. At this frequency

$$Z_r|_{500Hz} = (0.18 + j0.81), \quad \text{and} \quad Z_{t_{max}}|_{500Hz} = \frac{(\rho c/A)^2}{Z_r} = 86\angle - 77\,\text{deg}$$

cgs acoustic ohms. This driving point impedance is comparable in size to the ac equivalent resistance of the glottal source just determined. As frequency increases, the magnitude of Z_r increases, and the load reflected to the glottis at the quarter-wave resonances becomes smaller. At the second resonance, for example, $Z_r|_{1500Hz} = (1.63 + j2.44)$ and $Z_{t_{max}}|_{1500Hz} = 24\angle - 56\,\text{deg}$ cgs acoustic ohms. The reflected impedance continues to diminish with frequency until at very high frequencies $Z_r = Z_0 = 8.4$ cgs acoustic ohms. Note, too, that at the half-wave resonances of the tract, i.e., $l = n\lambda/2$, the sine terms in (Eq. 4.21) are zero and $Z_t = Z_r$.

The input impedance of the tract is greatest therefore at the frequency of the first quarter-wave resonance (which corresponds to the first formant). At and in the vicinity of this frequency, the driving point impedance (neglecting all losses except radiation) is comparable to the ac resistance of the glottal source. At all other frequencies it is less. For the unconstricted pipe the reflected impedance maxima are capacitive because the radiation load is inductive. To a first approximation, then, the glottal source appears as a constant volume velocity (current) source except at frequencies proximate to the first formant. As previously discussed, the equivalent vocal cord source sends an ac current equal to $u \cdot A'(t)$ into Z_r in parallel with R_g'. So long as constrictions do not become small, changes in the tract configuration generally do not greatly influence the operation of the vocal folds. At and near the frequency of the first formant, however, some interaction of source and tract might be expected, and in fact does occur. Pitch-synchronous variations in the tuning and the damping of the first formant–owing to significant tract-source interaction–can be observed experimentally[4].

---

[4] The acoustic mechanism of vocal-cord vibration and the interactions between source and system are discussed in more detail later. An acoustic oscillator model of the folds is derived in Chapter 11 and a computer simulation of the model is described.

4.4.5 High-Impedance Model of the Glottal Source

TO DO: Describe three successive approximations of the glottal volume velocity waveform: (1) the periodic triangle waveform, (2) the Fant model (LF with discontinuity), and (3) the Liljencrants-Fant model (Fant *et al.* [1986, 1994]). Provide equations, waveforms, and spectra to show the characteristics of glottal volume velocity correctly and incorrectly modeled by each function.

4.4.6 Experimental Studies of Laryngeal Biomechanics

Our knowledge of all speech excitation sources is heavily dependent on the results of mechanical modeling experiments. Parameters of the two-mass model of vocal tract vibration described in Ch. 4 (Ishizaka and Flanagan [1972b], Flanagan *et al.* [1975]) were adjusted in order to match flow measurements acquired from the mechanical model built by van den Berg ((van den Berg *et al.* [1957]); see also (Zantema and P. Doornenbal [1957], Meyer-Eppler [1953], Wegel [1930])).

It is possible that, at the dawn of the twenty-first century, improved computer simulations of turbulence may have finally eliminated the need for mechanical vocal tract models. Computer models developed for simulation of turbulence in vibrating cavities have generated surprising results (Pelorson *et al.* [1994], Huang and Levinson [1999]). Most notably, these studies demonstrate that pulsatile flow from a moving larynx does not re-attach to the vocal tract walls as efficiently as did the flow in van den Berg's model. Flow that fails to re-attach has two interesting consequences. First, pressure does not rise downstream of the glottis (Pelorson *et al.* [1994]). Second, flow continues to be non-laminar all the way between larynx and lips (Huang and Levinson [1999]): the nearly-laminar flow carries vortices along before it, like leaves in the wind, without significantly dissipating their vorticity. The latter finding was measured empirically long before it was successfully modeled; its impact on sound generation is still not well understood.

More recent studies have greatly enhanced our understanding of vocal fold biomechanics by measuring the air flow response of excised canine vocal folds (Alipour-Haghigi and Titze [1983], Alipour and Scherer [1995]), as well as the viscoelastic (Perlman [1985], Alipour-Haghigi and Titze [1985], Perlman *et al.* [1984], Perlman and Titze [1988], Alipour-Haghigi and Titze [1991]) and contractile properties (Alipour-Haghigi and Titze [1987, 1989]) of the vocalis muscle.

4.5 Turbulent Noise Sources

Noise excitation is generated by air moving quickly through a constriction. When air is moving slowly, it moves in a laminar fashion, meaning that the air particle velocity vectors are layered in planes roughly parallel to the vocal tract wall. When the velocity of the air becomes too great, or the constriction width too small, viscous forces tear apart the laminar flow, forcing the jet of air to twist and turn upon itself in a series of eddies and vortices. Each vortex serves as an initial condition for creation of the next vortex, in a kind of highly nonlinear feedback. Because of the nonlinear feedback between successive vortices, there is tremendous variability in the size and angular momentum of successive vortices. Successive vortices are created with diameters more or less randomly selected from a distribution ranging from the micrometer scale to the centimeter scale. Each successive vortex is carried downstream by the air jet, until it strikes against some kind of obstacle downstream from the constriction, and is broken up into yet smaller vortices. As each vortex strikes against obstacles in the vocal tract, the moving air creates local pressure fluctuations on the surface of the obstacle; these pressure fluctuations are pretty random, but the pressure fluctuations created by any single vortex tend to be concentrated at a frequency inversely proportional to the diameter of the vortex. Because the vortex diameters are uniformly distributed over a wide range, the center frequencies of the noise signals are also uniformly distributed over a wide range. The noise source that listeners hear is therefore very similar to "white noise," containing energy at all frequencies. The sound /s/, for example, is produced by forcing air through the narrow constriction between the tongue and the roof of the mouth. If the jet of air leaving the constriction is directed outward, the noise is not very loud; if the jet of air is directed downward against the lower teeth, then vortex energy is very effectively converted into noise, and listeners hear a loud fricative sound. The upper teeth serve this purpose in the production of dental fricatives such as /f/. One fricative consonant, /h/, is produced by turbulent flow generated at the glottis. The excitation mechanism is similar to that for the oral fricatives, except that the nonvibrating vocal folds create the constriction (the glottis during /h/ is open wider than it would be for any vowel, but it is still a narrower constriction than any constriction downstream in the vocal tract). The noise in /h/ may be increased in amplitude if the talker constricts his or her pharynx so that the airstream strikes the epiglottis.

Because it is spatially distributed, the location of the noise source in the tract is difficult to fix precisely. Generally it can be located at the constriction for a short closure, and just anterior to a longer constriction. In terms of a network representation, the noise source and its inherent impedance can be represented as the series elements in Fig. Fig. 4.16. P_s is the sound pressure generated by the turbulent flow and Z_s is the inherent impedance of the source. The series connection of the source can be qualitatively justified by noting that a shunt connection of a low-impedance pressure source would alter the mode structure of the vocal network. Furthermore, experimentally measured mode patterns for consonants appear to correspond to the series connection of the exciting source (Fant [1960]).

Voiced fricative sounds, such as /v/, are produced by simultaneous operation of the glottal and turbulent sources. Because the vibrating vocal folds cause a pulsive flow of air, the turbulent sound generated at the constriction is modulated by the glottal puffs. The turbulent sound is therefore generated as pitch-synchronous bursts

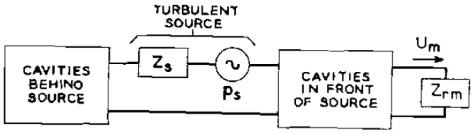


Figure 4.16: Equivalent circuit for noise excitation of the vocal tract

of noise.

It is possible to be a little more quantitative about several aspects or fricative excitation. For example, Meyer-Eppler(Meyer-Eppler [1953]) has carried out measurements on fricative generation in constricted plastic tube models of the vocal tract. He has related these measurements to human production of the fricative consonants /f,s,ʃ/. For these vocal geometries a critical Reynold's number, R_{ee}, apparently exists below which negligible turbulent sound is produced. Meyer-Eppler found that the magnitude of the noise sound pressure P_r—measured at a distance r from the mouth of either the model or the human—is approximately described by

$$P_r = K(R_e^2 - R_{ec}^2), \tag{4.24}$$

where K is a constant, R_e is the dimensionless Reynold's number $R_e = uw\rho/\mu$ and, as before, u is the particle velocity, ρ the air density, μ the coefficient of viscosity and w the effective width of the passage.

TO DO: Provide equations for the dipole turbulence source. Provide a figure showing the mechanism by which it is produced. Provide equations for the effective spectrum, and a figure showing the spectrum (Stevens [1971], Shadle [1985]).

We recall from the earlier discussion (Eq. (Eq. 4.8)) that for turbulent flow at a constriction the pressure drop across the orifice is approximately $P_d = \rho u^2/2 = \rho U^2/2A^2$. Therefore, $R_e^2 = 2\rho(w/\mu)^2 P_d$ and (Eq. 4.24) can be written

$$P_r = (K_1 w^2 P_d - K_2); \quad P_r \gg 0, \tag{4.25}$$

where K_1 and K_2 are constants. This result indicates that, above some threshold value, the fricative sound pressure in front of the mouth is proportional to the pressure drop at the constriction (essentially the excess pressure behind the occlusion) and to the square of the effective width of the passage.

By way of illustrating typical flow velocities associated with consonant production, a constriction area of 0.2 cm^2 and an excess pressure of 10cm $H_2 0$ are not unusual for a fricative like /s/. The particle velocity corresponding to this pressure is $u = (2P_d/\rho)^{\frac{1}{2}} \approx 4100$ cm/sec[5] and the volume flow is $U \approx 820\text{cm}^3/\text{sec}$.

If the constricted vocal passage is progressively opened and the width increased, a constant excess pressure can be maintained behind the constriction only at the expense of increased air flow. The flow must be proportional to the constriction area.

---

[5] Note this velocity is in excess of 0.1 Mach!

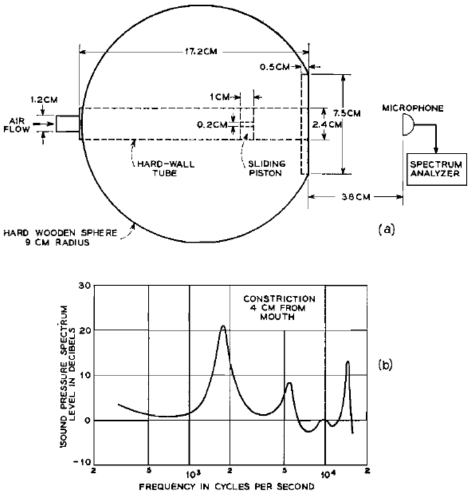


Figure 4.17: (a) Mechanical model of the vocal tract for simulating fricative consonants. (b) Measured sound spectrum for a continuant sound similar to /ʃ/. (After (Heinz [1958]))

The power associated with the flow is essentially $P_d U$ and hence also increases. Since the driving power is derived from the expiratory muscles, their power capabilities determine the maximum flow that can be produced for a given P_d. At some value of constriction area, a further increase in area, and consequently in w, is offset by a diminution of the P_d that can be maintained. The product $w^2 P_d$ in (Eq. 4.25) then begins to decrease and so does the intensity of the fricative sound.

Interest in mechanical analogs continues to the present day. The motivation is mainly that of simulating and measuring characteristics of human speech that are hard to simulate accurately on a computer, e.g., nonlinear aspects of vocal fold vibration, and turbulence in the pharyngeal and oral cavities. For example, one of the difficult parameters to measure in the real vocal tract is the location, intensity, spectrum, and internal impedance of the sound source for unvoiced sounds. One way of gaining knowledge about this source is with a mechanical analog. The technique for making such measurements is shown in Fig. 4.17a (Heinz [1958]).

Some disconnect here

The size of the spherical baffle is taken to represent the human head. A constricted tube in the baffle represents the vocal tract. Air is blown through the constriction to produce turbulence. The sound radiated is measured with a spectrum analyzer. A typical spectrum obtained when the constriction is placed 4cm from the "mouth," is plotted in Fig. 4.17b. The sound is roughly similar to the fricative /ʃ/. Because the constriction size for fricative consonants tends to be small, the spectral resonances are conditioned primarily by the cavities in front of the constriction. The antiresonances occur at frequencies where the impedance looking into the constriction from the mouth side is infinite. (Recall the discussion of Section 4.7.5.) The spectrum of the source is deduced to be relatively flat. Its total power is found to be roughly proportional

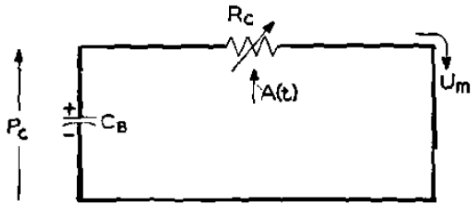


Figure 4.18: Approximate vocal relations for stop consonant production

to the fifth power of the flow velocity.

Recent mechanical modeling experiments have demonstrated the statistics of frication spectra with far more detail (Shadle [1985], Barney *et al.* [1999], Shadle *et al.* [1999]). For example, it is now known that the power spectrum of the frication pressure source tends to be broadly band-pass, with a power spectrum that rises at XXdB/octave below the peak, and falls at XX dB/octave above the peak (PUT FIGURES HERE). (PUT MORE INFORMATION HERE)

4.6 The Source for Transient Excitation

Stop consonants are produced by making a complete closure at an appropriate point (labial, dental or palatal), building up a pressure behind the occlusion, and sharply releasing the pressure by an abrupt opening of the constriction. This excitation is therefore similar to exciting an electrical network with a step function of voltage. The stop explosion is frequently followed by a fricative excitation. This latter element of the stop is similar to a brief fricative continuant of the same articulation.

Voiceless stop consonants contrast with fricatives in that they are more transient. For strongly articulated stops, the glottis is held open so that the subglottal system contributes to the already substantial volume behind the closure (V_B). The respiratory muscles apply a force sufficient to build up the pressure, but do not contract appreciably to force air out during the stop release. The air flow during the initial part of the stop release is mainly turbulent, with laminar streaming obtaining as the flow decays. In voiced stops in word-initial position (for example /d, g/), voicing usually commences following the release, but often (for example, in /b/) can be initiated before the release.

In very crude terms, stop production can be considered analogous to the circuit of Fig. Fig. 4.18. The capacitor C_B is the compliance ($V_B/\rho c^2$) of the cavities back of the closure and is charged to the excess pressure P_c. The resistance R_c is that of the constriction and is, according to the previous discussion [Eq. (Eq. 4.10)], approximately $R_c = \rho U_m/2A^2$. Suppose the constriction area is changed from zero as a step function, that is,

$$\begin{aligned} A(t) &= 0; \quad t < 0 \\ &= A; \quad t \geq 0. \end{aligned}$$

The mouth volume current then satisfies

$$U_m R_c + \frac{1}{C_B}\int_0^t U_m dt = P_c$$

or

$$\frac{\rho U_m^2}{2A^2} + \frac{1}{C_B}\int_0^t U_m dt = P_c, \quad \text{for } U_m > 0$$

and the solution for positive values of U_m is

$$U_m(t) = \left(\frac{2P_c}{\rho}\right)^{\frac{1}{2}} A\left[1 - \frac{At}{C_B(\rho 2P_c)^{\frac{1}{2}}}\right] \tag{4.26}$$

According to (Eq. 4.26) the flow diminishes linearly with time during the initial phases of the stop release. At the indicated rate, the time to deplete the air charge would be

$$t_1 = \frac{C_b(\rho 2P_c)^{\frac{1}{2}}}{A}. \tag{4.27}$$

As the flow velocity becomes small, however, the tendency is toward laminar streaming, and the resistance becomes less velocity dependent [sec first term in Eq. (Eq. 4.13)]. The flow decay then becomes more nearly exponential[6]

---

[6] This can be seen exactly by letting R_c include a constant (viscous) term as well as a flow-dependent term. Although the differential equation is somewhat more complicated, the variables separate, and the solution can be written in terms of U_m and $\ln U_m$.

Let

$$R_c = r_v A^{-3}(t) + r_k A^{-2}(t)\left|U_m\right|,$$

where r_v and r_k are constants involving air density and viscosity [as described in Eq. (Eq. 4.13)]. If the constriction area is changed stepwise from zero to A at time zero, the resulting flow will again be unipolar and now will satisfy

$$(r_k/A^2)U_m^2 + (r_v/A^3)U_m + 1/C_b\int_0^t U_m dt = P_c$$

The variables in this equation are separable and the solution can be obtained by differentiating both sides with respect to time. This yields

$$\frac{r_v}{A^3}\left(\frac{dU_m}{dt}\right) + 2\frac{r_k}{A^2}U_m\frac{dU_m}{dt} + \frac{U_m}{C_b} = 0$$

and

$$\frac{r_v C_B}{A^3}\left(\frac{dU_m}{U_m}\right) + 2\frac{r_k C_B}{A^2}dU_m = -dt.$$

To fix some typical values, consider the production of a voiceless stop such as /t/. According to Fant (Fant [1960]), realistic parameters for articulation of this sound are $P_c = 6$ cm H_20, $V_B = \rho c^2 C_B = 4$ liters (including lungs) and $A = 0.1$ cm^2. Assuming the area changes abruptly, substitution of these values into (Eq. 4.26) and (Eq. 4.27) gives $U_m(0) = 320$ cm^3/sec and $t_1 = 130$ msec. The particle velocity at the beginning of the linear decay is $u_m(0) = 3200$ cm/sec. After 50 msec it has fallen to the value 1300 cm/sec which is about the lower limit suggested by Meyer-Eppler for noise generation. As Fant points out, the amount of air consumed during this time is quite small, on the order of 10 cm^3.

Both Stevens(Stevens [1956]) and Fant(Fant [1960]) emphasize the importance of the open glottis in the production of a strong stop consonant. A closed glottis reduces V_B to something less than 100 cm^3, and the excess pressure which can be produced behind the constriction is typically on the order of 3 cm H_20. For such conditions is it difficult to produce flows sufficient for noise generation. The turbulent noise produced during the stop release is essentially a secondary effect of the excitation. The primary excitation is the impact of the suddenly applied pressure upon the vocal system. As mentioned earlier, this excitation for an abrupt area change is analogous to a step function of voltage applied to an electrical circuit. Such a source is characterized by a spectrum which is proportional to $1/\omega$, or diminishes in amplitude at -6 db/oct.

Integrating termwise give

$$\frac{r_v C_B}{A^3} \ln U_m\Big]_0^t + 2\frac{r_k C_B}{A^2} U_m\Big]_0^t = -t.$$

At $t = 0$, $U_m = U_0$, where U_0 is the positive real root of the quadratic

$$\left(\frac{r_k}{A^2}\right) U_0^2 + \frac{r_v}{A^3} U_0 - P_c = 0.$$

Then

$$\ln\left(\frac{U_m}{U_0}\right) + \frac{2 r_k A}{r_v}\left(U_m - U_0\right) + \frac{tA^3}{r_v C_B} = 0.$$

Note

$$\begin{aligned} \text{for } A \text{ large:} \quad & U_m \approx \left[U_0 - \left(\frac{A^2}{2 r_k C_B}\right) t\right] \\ \text{for } A \text{ small:} \quad & U_m \approx U_0 e^{-\left(\frac{A^3}{r_v C_B}\right) t}. \end{aligned}$$

It also follows that

$$\begin{aligned} \frac{dU_m}{dt} &= \frac{-U_m}{\frac{r_v C_B}{A^3} + \frac{2 r_k C_B}{A^2} U_m} \\ &\approx \frac{-A^2}{2 r_k C_B}, \text{ for large } A \\ &\approx \frac{-U_m A^3}{r_v C_B}, \text{ for small } A. \end{aligned}$$

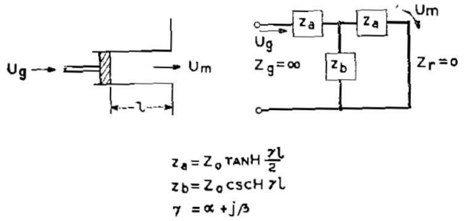


Figure 4.19: Relation between glottal and mouth volume currents for the unconstricted tract. The glottal impedance is assumed infinite and the radiation impedance is zero

TO DO: Compare the calculations above to those of (Massey [1994]), and derive Massey's equivalent transient source.

4.7 Some Characteristics of Vocal Tract Transmission

Some of the fundamental relations developed in the foregoing sections can now be used to put in evidence certain properties of vocal transmission. These characteristics are easiest demonstrated analytically by highly simplifying the tract geometry. Calculations on detailed approximations are more conveniently done with computers. Although our examples generally will be oversimplified, the extensions to more exact descriptions will in most cases be obvious.

As a first step, consider the transmission from glottis to mouth for nonnasal sounds. Further, as an ultimate simplification, consider that the tract is uniform in cross section over its whole length l, is terminated in a radiation load whose magnitude is negligible compared with the characteristic impedance of the tract, and is driven at the glottis from a volume-velocity source whose internal impedance is large compared to the tract input impedance. The simple diagram in Fig. Fig. 4.19 represents this situation. The transmission function relating the mouth and glottal volume currents is then

$$\frac{U_m}{U_g} = \frac{z_b}{z_b + z_a} = \frac{1}{\cosh \gamma l} \tag{4.28}$$

The normal modes (poles) of the transmission are the values of γl which make the denominator zero. These resonances produce spectral varialions in the sound radiated from the mouth. They are

$$\begin{aligned} \cosh \gamma l = &\quad 0 \\ \gamma l = &\quad \pm j(2n+1)\tfrac{\pi}{2}, \qquad n = 0, 1, 2, \ldots \end{aligned} \tag{4.29}$$

The poles therefore occur at complex values of frequency. Letting $j\omega = \sigma + j\omega = s$, the complex frequency, and recalling from (Eq. 3.9) that $\gamma = \alpha + j\beta$ and $\beta \approx \omega/c$ for small losses, the complex pole frequencies may be approximated as

$$s_n \approx -\alpha c \pm j\frac{(2n+1)\pi c}{2l}, \quad n = 0, 1, 2, \ldots ^{7} \tag{4.30}$$

The transmission (Eq. 4.28) can be represented in factored form in terms of the roots of the denominator, namely

$$H(s) = \frac{U_m(s)}{U_g(s)} = \prod_n \frac{s_n s_n^*}{(s - s_n)(s - s_n^*)}, \tag{4.31}$$

where s_n^* is the complex conjugate of s_n, and the numerator is set to satisfy the condition

$$\left.\frac{U_m(j\omega)}{U_g(j\omega)}\right|_{j\omega=0} = \frac{1}{\cosh \alpha l} \approx 1,$$

for small α. The transmission is therefore characterized by an infinite number of complex conjugate poles[8]. The manifestations of these normal modes as spectral peaks in the output sound are called formants. The transmission (Eq. 4.31) exhibits no zeros at finite frequencies. Maxima occur in

$$|H(j\omega)| \quad \text{for } \omega = \pm(2n+1)\frac{\pi}{2}\frac{c}{l}$$

and the resonances have half-power bandwidths in Hertz approximately equal to $\Delta f = \sigma/n = \alpha c/\pi$. For an adult male vocal tract, approximately 17 cm in length, the unconstricted resonant frequencies therefore fall at about $f_1 = 500$ Hz, $f_2 = 1500$ Hz, $f_3 = 2500$ Hz, and continue in $c/2l$ increments.

In the present illustration the only losses taken into account are the classical heat conduction and viscous losses discussed earlier. A calculation of formant bandwidth on this basis alone will consequently be abnormally low. It is nevertheless instructive to note this contribution to the formant damping. Recall from Eq. (Eq. 3.9) that for small losses

$$\alpha \approx \frac{R_a}{2}\sqrt{C_a}L_a + \frac{G_a}{2}\sqrt{L_a}C_a,$$

where R_a, G_a, L_a and C_a have been given previously in Section (3.2.5). At the first-formant frequency for the unconstricted tract (i.e., 500 Hz), and assuming a circular cross-section with typical area 5 cm^2, α is computed to be approximately 5.2×10^{-4}, giving a first-formant bandwidth $\Delta f_1 = 6$ Hz. At the second formant frequency (i.e., 1500 Hz) the same computation gives $\Delta f_2 = 10$ Hz. The losses increase as $f^{\frac{1}{2}}$, and at

---

[7] Actually α is an implicit function of ω [see Eq. (Eq. 3.46)]. However, since its frequency dependence is relatively small, and since usually $\sigma_n \ll \omega_n$, the approximation (Eq. 4.30) is a convenient one.

[8] Rigorous justification of the form (Eq. 4.31) has its basis in function theory (Titchmarsh [1932], Ahlfors [1953]). See Chapter 11, Sec. 11.4 for further discussion of this point.

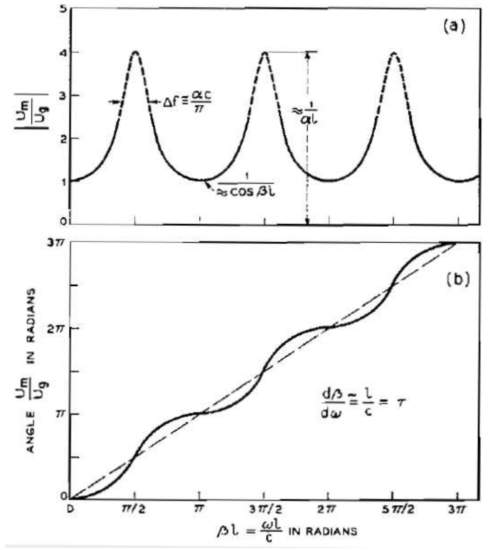


Figure 4.20: Magnitude and phase of the glottis-to-mouth transmission for the vocal tract approximation shown in Fig. Fig. 4.19

the third formant (2500 Hz) give $\Delta f_3 = 13$ Hz. It is also apparent from (Eq. 4.31) that $H(s)$ is a minimum phase function (that is, it has all of its zeros, namely none, in the left half of the s-plane) so that its amplitude and phase responses are uniquely linked (that is, they are Hilbert transforms). Further, the function is completely specified by the s_n's, so that the frequency*and* amplitude of a formant peak in $|H(j\omega)|$ are uniquely described by the pole frequencies. In particular if the formant damping can be considered known and constant, then the amplitudes of the resonant peaks of $|H(j\omega)|$ are implicit in the imaginary parts of the formant frequencies $\omega_1, \omega_2, \ldots,$ (Fant [1956], Flanagan [1957b]). In fact, it follows from (Eq. 4.28) that

$$\begin{aligned} |H(j\omega)|_{\omega=\omega_n} &= \frac{1}{|\cosh(\alpha+j\beta)l|_{\omega=\omega_n}} \\ &= \frac{1}{|j\sinh \alpha l|} \\ &\approx \frac{1}{\alpha l} \end{aligned} \tag{4.32}$$

where $\beta = \omega/c$ and $\omega_n = (2n+1)\pi c/2l$. Notice, too, that the phase angle of $H(j\omega)$ advances n radians in passing a formant frequency ω_n; so the amplitude and phase response of $H(j\omega)$ appear as in Fig. Fig. 4.20. In the same connection, note that for the completely lossless case

$$H(j\omega) = \frac{1}{\cos \frac{\omega l}{c}}.$$

4.7.1 Effect of Radiation Load upon Mode Pattern

If the radiation load on the open end of the tube is taken into account, the equivalent circuit for the tube becomes that shown in Fig. Fig. 4.21. Here A_t is the cross-

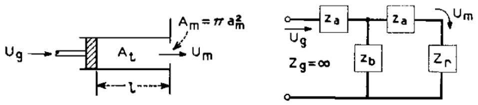


Figure 4.21: Equivalent circuit for the unconstricted vocal tract taking into account the radiation load. The glottal impedance is assumed infinite

sectional area of the tract and A_m is the radiating area of the mouth with equivalent radius a_m. The thickness of the mouth constriction is assumed negligible, the glottal impedance is high, and cross dimensions are small compared with a wavelength. The transmission from glottis to mouth is therefore

$$\frac{U_m}{U_g} = \frac{1}{\cosh \gamma l + \frac{Z_r}{Z_0} \sinh \gamma l},$$

or, more conveniently

$$\frac{U_m}{U_g} = \frac{\cosh \gamma_r l}{\cosh(\gamma + \gamma_r) l}, \tag{4.33}$$

where $\gamma_r l = \tanh^{-1} Z_r/Z_0$. Note that for $Z_r \ll Z_0$, $\cosh \gamma_r l \approx 1$ and for low loss $Z_0 \ \rho c/A_t$.

By the transformation (Eq. 4.33), the radiation impedance is carried into the propagation constant, so that

$$\begin{aligned} (\gamma + \gamma_r) &= \left[\alpha + j\beta + \frac{1}{l} \tan^{-1} \frac{Z_r}{Z_0}\right] \\ &= (\alpha + j\beta + \alpha_r + j\beta_r) = (\alpha' + j\beta') = \gamma'. \end{aligned}$$

If the radiation load is taken as that on a piston in a wall [see Eq. Eq. 4.3 in Sec. 4.2] then

$$Z_r \approx \frac{\rho c}{A_m}\left[\frac{(ka)^2}{2} + j\frac{8ka}{3\pi}\right], \quad ka \ll 1 \tag{4.34}$$

where a equals the mouth radius a_m. Expanding $\tanh^{-l} Z_r/Z_0$ as a series and taking only the first term (i.e., assuming $Z_r \approx Z_0$) gives

$$\gamma_r \approx \frac{1}{l}\frac{A_t}{A_m}\left[\frac{(ka)^2}{2} + j\frac{8ka}{3\pi}\right] \tag{4.35}$$

$$= \alpha_r + j\beta_r.$$

For low loss $\beta \approx \omega/c = k$, so that

$$(\alpha' + j\beta') = \left[\alpha + \frac{A_t}{A_m}\frac{(\beta a)^2}{2l}\right] + j\beta\left[1 + \frac{A_t}{A_m}\frac{8a}{3\pi l}\right]. \tag{4.36}$$

Again the poles of (Eq. 4.33) occur for

$$e^{2\gamma' l} + 1 = 0$$

or

$$\gamma' = \pm j \frac{(2n+1)\pi}{2l}, \quad n = 0, 1, 2, \ldots \tag{4.37}$$

Letting $j\omega \to s = (\sigma + j\omega)$, and remembering that in general $\sigma_n \ll \omega_n$, the poles are approximately

$$s_{nr} \approx \frac{1}{1 + \frac{A_t 8a}{A_m 3\pi l}} \left[-\left(\alpha c + \frac{A_t \omega^2}{2\pi l c} \right) \pm j \frac{(2n+1)\pi c}{2l} \right], \tag{4.38}$$

$$n = 0, 1, 2, \ldots \quad (Z_r \ll Z_0).$$

The general effect of the radiation, therefore, is to decrease the magnitude of the imaginary parts of the pole frequencies and to make their real parts more negative.

For the special case $A_m = A_0$ the modes are

$$s_{nr} \approx \left(\frac{3\pi l}{3\pi l + 8a} \right) \left[-\left(\alpha c + \frac{a^2 \omega^2}{2lc} \right) \pm j \frac{(2n+1)\pi c}{2l} \right]. \tag{4.39}$$

Using the values of the example in the previous section, $A_t = 5$ cm^2, $l = 17$ cm, the spectral resonances (formants) are lowered in frequency by the multiplying factor $3\pi l/(3\pi l + 8a) = 0.94$. The original 500 Hz first formant is lowered to 470 Hz, and the 1500 Hz second formant is lowered to 1410 Hz. The first formant bandwidth is increased to about $\Delta f_1 \approx 0.94(6 + 4) = 9$Hz, and the second formant bandwidth to about $\Delta f_2 \approx 0.94(l0+32) = 40$ Hz. The same computation for the third formant gives $\Delta f_3 \approx 100$ Hz. The latter figures begin to be representative of formant bandwidths measured on real vocal tracts with the glottis closed (House and Stevens [1958], Dunn [1961], van den Berg [1955]). The contributions of the radiation, viscous and heat losses to Δf_1 are seen to be relatively small. Glottal loss and cavity wall vibration generally are more important contributors to the first formant damping.

As (Eq. 4.39) indicates, the contribution of the radiation resistance to the formant damping increases as the square of frequency, while the classical heat conduction and viscous loss cause a to grow as $\omega^{\frac{1}{2}}$. The radiation reactance is inertive and causes the formant frequencies to be lowered. For $A_m = A_t$, Eq. (Eq. 4.38) shows that the radiation reactance has the same effect as lengthening the vocal tract by an amount $(8a/3\pi)$.

4.7.2 Effect of Glottal Impedance upon Mode Pattern

The effect of the equivalent glottal impedance can be considered in much the same manner as the radiation load. To keep the illustration simple, again assume the

Figure 4.22: Equivalent circuit for the unconstricted vocal tract assuming the glottal impedance to be finite and the radiation impedance to be zero

radiation load to be negligible compared with the characteristic impedance of the uniform tract, but take the glottal impedance as finite. This situation is depicted by Fig. Fig. 4.22. Similar to the previous instance, the volume velocity transmission function can be put in the form

$$\begin{aligned}\frac{U_m}{U_g} &= \frac{1}{\frac{z_a}{Z_g}\left(\frac{Z_g}{z_b}+\frac{z_a}{z_b}+1\right)+1+\frac{z_a}{Z_g}} \\ &= \frac{1}{\cosh\gamma l+\frac{Z_0}{Z_g}\sinh\gamma l} \\ &= \frac{\cosh\gamma_g l}{\cosh(\gamma+\gamma_g)l},\end{aligned} \tag{4.40}$$

where $\gamma_g l = \tanh^{-1} Z_0/Z_g$, and the glottal impedance is transformed into the propagation constant. Again taking the first term of the series expansion for $\tanh^{-1} Z_0/Z_g$ (i.e., assumming $Z_g \gg Z_0$) gives

$$(\gamma+\gamma_g) \approx \left(\alpha + j\beta + \frac{1}{l}\frac{Z_0}{Z_g}\right).$$

The equivalent glottal impedance may be approximated as $Z_g = (R'_g + j\omega L_g)$, where R'_g is the ac equivalent resistance determined previously in Eq. (Eq. 4.18), and L_g is the effective inductance of the glottal port. The zeros of the denominator of (Eq. 4.40) are the poles of the transmission, and an argument similar to that used in the preceding section for low losses ($Z_0 \approx \rho c/A_t, \beta \approx \omega/c$) leads to

$$s_{ng} \approx \frac{1}{1-\left(\frac{L_g Z_0 c}{l|Z_g|^2}\right)}\left\{-\left(\alpha c + \frac{R'_g Z_0 c}{l|Z_g|^2}\right) \pm j\frac{(2n+1)\pi c}{2l}\right\}. \tag{4.41}$$

According to (Eq. 4.41), the effect of the finite glottal impedance is to increase the damping of the formant resonances (owing to the glottal loss R'_g) and to increase the formant frequencies by the factor multiplying the bracketed term (owing to the glottal inductance). A sample calculation of the effect can be made. As typical values, take a subglottic pressure (P_s) of 8 cm H_20, a mean glottal area (A_0) of 5mm^2, a glottal orifice thickness (d) of 3 mm, a vocal tract area (A_t) of 5 cm^2 and a tract length (l) of 17 cm. For these conditions the glottal resistance, computed according to Eq. (Eq. 4.18), is $R'_g \approx 91$ cgs acoustic ohms. The glottal inductance is $L_g = \sigma d/A_0 = 6.8 \times 10^{-3}$ cgs units. At about the frequency of the first formant, that is, $\omega \approx \pi c/2l = 2\pi$ (500 Hz), the multiplying factor has a value $1/(1-0.014)$, so that the first formant resonance is increased from its value for the infinite glottal impedance condition by about 1.4%. The effect of the glottal inductance upon formant tuning is greatest for the lowest formant because $|Z_g|$ increases with frequency. The same computation for the second formant (1500 Hz) shows the multiplying factor to be $1/(1-0.010)$. One notices also that the effect of the multiplying term is to shorten the apparent length of the tract to

$$\left(l - \frac{L_g Z_0 c}{|Z_g|^2}\right).$$

The resonant bandwidth for the first formant is computed to be

$$\Delta f_1 = \frac{1}{(1-0.014)}\,[6\text{Hz} + 56\text{Hz}] = 63\text{Hz},$$

which is reasonably representative of first formant bandwidths measured in real speech. The contribution of the glottal loss R_g' to formant damping is greatest for the lowest formant. It diminishes with increasing frequency because $|Z_g|$ grows with frequency. At the second formant frequency, the same calculation gives $\Delta f_2 = 1/(1-0.010)(10\text{Hz}+40\text{Hz}) = 51\text{Hz}$. One recalls, too, that the heat conduction and viscous losses (which specify α) increase as ωt, while the radiation loss increases as ω^2 (for $ka \ll 1$). The lower-formant damping is therefore influenced more by glottal loss, and the higher-formant damping is influenced more by radiation loss.

In this same connection, one is reminded that the glottal resistance and inductance (used here as equivalent constant quantities) are actually time varying. There is consequently a pitch-synchronous modulation of the pole frequencies s_{ng} given in (Eq. 4.41). That is, as the vocal folds open, the damping and resonant frequency of a formant increase, so that with each glottal period the pole frequency traverses a small locus in the complex-frequency plane. This pitch-synchronous change in formant damping and tuning can often be observed experimentally, particularly in inverse filtering of formants. It is most pronounced for the first formant.

4.7.3 Effect of Cavity Wall Vibration

The previous discussion has assumed the walls of the vocal tract to be smooth and rigid. The dissipative elements of concern are then the radiation resistance, the glottal resistance, and the viscous and heat conduction losses at the cavity walls. The human vocal tract is of course not hard-walled, and its surface impedance is not infinite. The yielding walls can consequently contribute to the energy loss in the tract and can influence the mode tuning. We would like to estimate this effect.

The finite impedance of the tract wall constitutes an additional shunt path in the equivalent “T” (or π) section for the pipe (see Fig. Fig. **??**). Because the flesh surrounding the tract is relatively massive and exhibits viscous loss, the additional shunt admittance for the frequency range of interest (i.e., speech frequencies) can be approximated as a per-unit-length reciprocal inductance or inertance ($\Gamma_w = 1/L_w$)

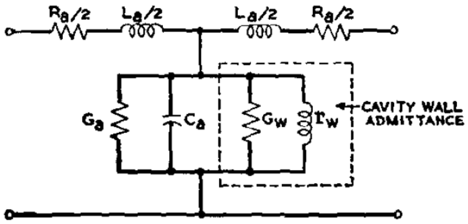


Figure 4.23: Representation of wall impedance in the equivalent T-section for a length l of uniform pipe

and a per-unit-length conductance ($G_w = 1/R_w$) in parallel[9]. The modified equivalent "T" section is shown in Fig. Fig. 4.23.

Let us note the effect of the additional shunt admittance upon the propagation constant for the tube. As before, the basic assumption is that a plane wave is propagating in the pipe and that the sound pressure at any cross section is uniform and cophasic. Recall that

$$\gamma = \alpha + j\beta = \sqrt{yz},$$

where y and z are the per-unit-length shunt admittance and series impedance, respectively. The latter quantities are now

$$z = (R_a + jwL_a)$$
$$y = (G_a + G_w) + j\left(\omega C_a - \frac{\Gamma_w}{\omega}\right). \tag{4.42}$$

Again, most conditions of interest will be relatively small-loss situations for which

$$R_a \ll \omega L_a$$

and

$$(G_a + G_w) \ll \left(\omega C_a - \frac{\Gamma_w}{\omega}\right).$$

Also, in general, the susceptance of the air volume will exceed that of the walls and $\omega C_a \gg \Gamma_w/\omega$. Following the earlier discussion [see Eq. (Eq. 3.9)] the attenuation constant for this situation can be approximated by

$$\alpha \approx \frac{1}{2}R_a\sqrt{\frac{C_a}{L_a}} + \frac{1}{2}(G_a + G_w)\sqrt{\frac{L_a}{C_a}} \tag{4.43}$$

In a like manner, the phase constant is given approximately by

$$\beta \approx \omega\sqrt{L_a\left(C_a - \frac{\Gamma_w}{\omega^2}\right)} = \frac{\omega}{c'}. \tag{4.44}$$

---

[9] For describing the behavior at very low frequencies, a compliance element must also be considered.

The effective sound velocity c' in a pipe with "massive" walls–that is, with negative susceptance–is therefore faster than for free space. The pipe appears shorter and the resonant frequencies are shifted upward. The effect is greatest for the lower frequencies. The same result can be obtained more elegantly in terms of specific wall admittance by writing the wave equation for the cylindrical pipe, noting the radial symmetry and fitting the boundary impedance conditions at the walls (Morse [1948]). In addition to the plane-wave solution, the latter formulation also gives the higher cylindrical modes.

Results (Eq. 4.43) and (Eq. 4.44) therefore show that vibration of the cavity wall contributes an additive component to the attenuation constant, and when the wall is predominantly mass-reactive, its effect is to diminish the phase constant or increase the speed of sound propagation. Following the previous technique [see Eq. (Eq. 4.30)], the natural modes for a uniform tube of this sort are given by

$$\begin{aligned} s_{nw} &= \left[-\alpha c' \pm j \frac{(2n+l)\pi c'}{2l} \right] \\ &= (\sigma_{nw} + j\omega_{nw})\,; \quad n = 0, 1, 2, \ldots \end{aligned} \tag{4.45}$$

To calculate the shunting effect of the walls in the real vocal tract, it is necessary to have some knowledge of the mechanical impedance of the cavity walls. Such measurements are obviously difficult and apparently have not been made. An order-of-magnitude estimate can be made, however, by using mechanical impedance values obtained for other surfaces of the body. At best, such measurements are variable, and the impedance can change appreciably with place. The data do, however, permit us to make some very rough calculations.

One set of measurements (Franke [1951]) has been made for chest, thigh and stomach tissues, and these have been applied previously to estimate the wall effect (House and Stevens [1958]). For frequencies above about 100 Hz, the fleshy areas exhibit resistive and mass reactive components. The specific impedances fall roughly in the range 4000-7000 dyne-sec/cm^3. A typical measurement on the stomach surface gives a specific impedance that is approximately

$$\begin{aligned} z_s &= (r_s + jx_s) = (r_s + j\omega l_s) \\ &= (6500 + j\omega 0.4), \end{aligned} \tag{4.46}$$

for $(2\pi \cdot 200) \leq \omega \leq (2\pi \cdot 1000)$.

This specific series impedance can be put in terms of equivalent parallel resistance and inductance by

$$r_p = \frac{r_s^2 + x_s^2}{r_s} \quad \text{and} \quad jx_p = j\frac{r_s^2 + x_s^2}{x_s}.$$

These specific values (per-unit-area) can be put in terms of per-unitlength of tube by dividing by S, the inner circumference, to give

$$R_w = fracr_s^2 + x_s^2 r_s S \quad \text{and} \quad jX_w = j\frac{r_s^2 + x_s^2}{x_s S}.$$

Therefore,

$$G_w = \frac{r_s S}{r_s^2 + x_s^2} \quad \text{and} \quad -j\frac{\Gamma_w}{\omega} = -j\frac{\omega l_s S}{r_s^2 + x_s^2},$$

where,

$$\Gamma_w = \frac{\omega^2 l_s S}{r_s^2 + x_s^2}. \tag{4.47}$$

Assuming the vocal tract to be unconstricted and to have a uniform cross-sectional area of 5 cm^2 (i.e., $S = 7.9$ cm), we can compute the effect of the wall admittance upon the propagation constant, the formant bandwidth and formant frequency. According to (Eq. 4.43) and (Eq. 4.44), the wall's contribution to α and β is

$$\alpha_w \approx \frac{G_w}{2}\sqrt{\frac{L_a}{C_a}},$$

and

$$\begin{aligned}\beta_w &\approx \omega\sqrt{L_a\left(C_a - \frac{l_s S}{r_s^2 + x_s^2}\right)} \\ &\approx \frac{\omega}{c}\left[1 - \frac{\rho c^2 l_s}{a(r_s^2 + x_s^2)}\right],\end{aligned} \tag{4.48}$$

where the radius of the tube is $a = \sqrt{A/\pi}$, and the bracketed expression is the first two terms in the binomial expansion of the radical.

Substituting the measured values of r_s and l_s and computing α_w, β_w and formant bandwidths at approximately the first three formant frequencies gives[10]

Frequency	α_w	β_w	$\Delta f_w = \frac{\alpha_w c'}{\pi}$
500Hz	4.7×10^{-3}	$\frac{\omega}{c}(1 - 0.011)$	50Hz
1500Hz	3.6×10^{-3}	$\frac{\omega}{c}(1 - 0.008)$	40Hz
2500Hz	2.5×10^{-3}	$\frac{\omega}{c}(1 - 0.006)$	30Hz

The contribution of wall loss to the formant bandwidth is therefore greatest at the lowest formant frequency and diminishes with increasing formant frequency. These computed values, however, when combined with the previous loss contributions actually seem somewhat large. They suggest that the walls of the vocal tract are more rigid than the stomach tissue from which the mechanical impedance estimates were made.

The increase in formant tuning, occasioned by the mass reactance of the cavity walls, is seen to be rather slight. It is of the order of one per cent for the lower formants and, like the damping, diminishes with increasing frequency.

---

[10] Using $c = 3.5 \times 10^4$ cm/sec and $\rho = 1.14X10^{-3}$ gm/cm^3.

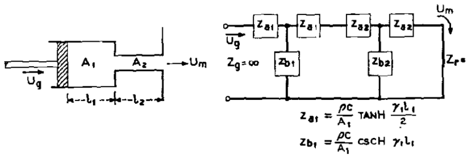


Figure 4.24: Two-tube approximation to the vocal tract. The glottal impedance is assumed infinite and the radiation impedance zero

4.7.4 Two-Tube Approximation of the Vocal Tract

The previous sections utilized a uniform-tube approximation of the vocal tract to put in evidence certain properties. The uniform tube, which displays modes equally spaced in frequency, comes close to a realistic vocal configuration only for the unconstricted schwa sound /ə/. Better insight into the interaction of vocal cavities can be gained by complicating the approximation one step further; namely, by approximating the tract as two uniform, cascaded tubes of different cross section. To keep the discussion tractable and focused mainly upon the transmission properties of the tubes, we again assume the glottal impedance to be high compared with the input impedance of the tract, and the radiation load to be negligible compared with the impedance level at the mouth. This situation is represented in Fig. Fig. 4.24.

For the circuit shown in Fig. Fig. 4.24, the mouth-to-glottis volume current ratio is

$$\frac{U_m}{U_g} = \frac{1}{\left(1 + \frac{z_{a2}}{z_{b2}}\right)\left(1 + \frac{z_{a1}}{z_{b1}} + \frac{z_{a2}}{z_{b1}}\right) + \frac{z_{a2}}{z_{b1}}},$$

which reduces to

$$\frac{U_m}{U_g} = \frac{1}{(\cosh \gamma_1 l_1)(\cosh \gamma_2 l_2)\left(1 + \frac{A_1}{A_2} \tanh \gamma_1 l_1 \tanh \gamma_2 l_2\right)}. \tag{4.49}$$

The poles of (Eq. 4.49) occur for

$$\frac{A_1}{A_2} \tanh \gamma_2 l_2 = -\coth \gamma_1 l_1. \tag{4.50}$$

If the tubes are lossless, the hyperbolic functions reduce to circular functions and all impedances are pure reactances. The normal modes then satisfy

$$\frac{A_1}{A_2} \tan \beta l_2 = \cot \beta l_1 \tag{4.51}$$

Because the vocal tract is relatively low loss, Eq. (Eq. 4.51) provides a simple means for examining the mode pattern of the two-tube approximation. For example, consider the approximations shown in Fig. Fig. 4.25 to the articulatory configurations for four

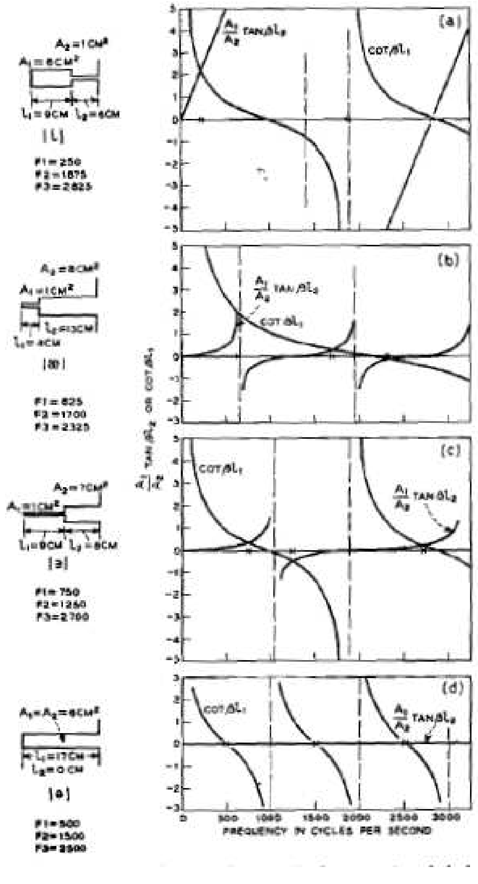


Figure 4.25: Two-tube approximations to the vowels /i,æ,ɑ,ə/ and their undamped mode (formant) patterns

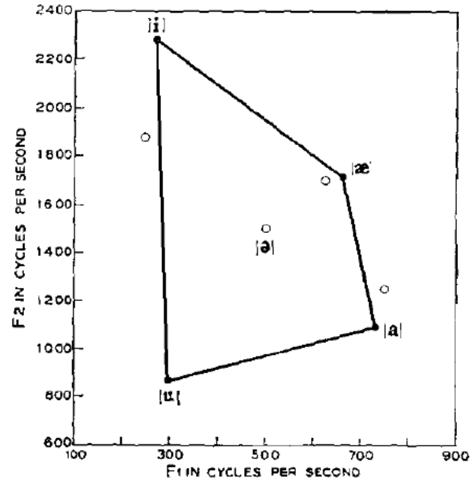


Figure 4.26: First formant ($F1$) versus second formant ($F2$) for several vowels. Solid points are averages from Peterson and Barney's (1952) data for real speech uttered by adult males. Circles are for the two-tube approximation to the vowels shown in Fig. Fig. 4.25

different vowels. The reactance functions of (Eq. 4.51) are plotted for each case, and the pole frequencies are indicated.

One notices that the high front vowel /i/ exhibits the most disparate first and second formants, while the low back vowel /ɑ/ gives rise to the most proximate first and second formants. The neutral vowel /ə/, corresponding to the unconstricted tract, yields formants uniformly spaced 1000 Hz apart. The reactance plots also show that increasing the area ratio (A_1/A_2) of the back-to-front cavities results in a decrease of the first formant frequency. On the classical F_1 vs F_2 plot, the first two modes for the four approximations fall as shown in Fig. Fig. 4.26. The unconstricted /ə/ sound occupies the central position. For comparison, formant data for four vowels–as spoken by adult males–are also plotted (Peterson and Barney [1952]).[11] The lower left corner of the classical vowel plot, the area appropriate to the vowel /u/, has been indicated for completeness. Because of lip rounding, however, the vowel /u/ cannot be approximated in terms of only two tubes.

Eq. (Eq. 4.51) also makes salient an aspect of compensatory articulation. The mode pattern for $l_1 = a$, $l_2 = b$, is exactly the same as for $l_1 = b, l_2 = a$. In other words, so long as the area ratio for the back and front cavities is maintained the same,

---

[11] Most of the vocal tract dimensions used to illustrate acoustic relations in this chapter are appropriate to adult males. Women and children have smaller vocal apparatus. Since the frequencies of the resonant modes are inversely related to the tract length, the vowel formants for women and children are higher than for the men. According to Chiba and Kajiyama (Chiba and Kajiyama [1941]), the young adult female vocal tract is 0.87 as long as the young adult male. The female formants, therefore, should be about 15% higher than those of the male. This situation is also reflected in the measurements of Peterson and Barney.

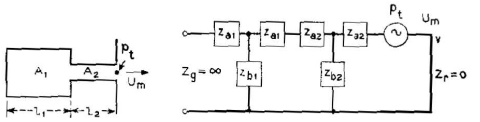


Figure 4.27: Two-tube approximation to the vocal tract with excitation applied forward of the constriction

their lengths may be interchanged without altering the formant frquencies. This is exactly true for the idealized lossless tubes, and is approximately so for practical values of loss. This interchangeability is one freedom available to the ventriloquist. It is also clear from (Eq. 4.51) that if $l_1 = 2l_2$, the infinite values of $\cot \beta l_1$ and $\tan \beta l_2$ are coincident (at $\beta l_2 = \pi/2$) and indicate the second mode. The second formant frequency can therefore be maintained constant by keeping the tube lengths in the ratio of 2:1. The same constancy applies to the third formant if the length ratio is maintained at 3:2.

4.7.5 Excitation by Source Forward in Tract

As pointed out earlier, fricative sounds (except for /h/) are excited by a series pressure source applied at a point forward in the tract. It is pertinent to consider the mouth volume velocity which such an excitation produces.

A previous section showed that for glottal excitation the maxima of glottis-to-mouth transmission occurred at the natural (pole) frequencies of the vocal system, and the transmission exhibited no zeros. If excitation is applied at some other point in the system, without altering the network, the normal modes of the response remain the same. The transmission can, however, exhibit zeros. For the series excitation these zeros must occur at frequencies where the impedance looking back from the source (toward the glottis) is infinite. By way of illustration let us retain the simple two-tube model used previously. Because the turbulent source for voiceless sound is spatially distributed, its exact point of application is difficult to fix. Generally it can be thought to be applied either at or just forward of the point of greatest constriction. The former seems to be more nearly the case for sounds like /ʃ,f,p,k/; the latter for /s,t/. Consider first the case where the source is forward of the constriction. The two-tube circuit is shown in Fig. Fig. 4.27. The back cavity is shown closed, and the impedance of the glottis and larynx tube is considered to be high (compared to the impedance level of the back cavity) even though the glottis may be open. The radiation impedance is again considered small compared with the impedance level at the mouth, and the inherent impedance of the source per se is considered small.

The complex frequency (Laplace) transform of the transmission (U_m/p_t) can be written in the form

$$\frac{U_m(s)}{p_t(s)} = H(s)G(s), \tag{4.52}$$

where $H(s)$ is a given in (Eq. 4.31) and contains all the poles of the system, and $G(s)$ is a function which includes all the zeros and constants appropriate to nonglottal excitation. In this particular case, U_m/p_t is simply the driving point admittance at the lips. It is

$$\frac{U_m}{p_t} = \frac{(z_{b2} + z_{bl} + z_{a1} + z_{a2})}{z_{a2}(z_{b2} + z_{b1} + z_{a1} + z_{a2}) + z_{b2}(z_{b1} + z_{a1} + z_{a2})},$$

which can be put into the form

$$\frac{U_m}{p_t} = \frac{\frac{1}{Z_{01}} \sinh \gamma_1 l_1 \sinh \gamma_2 l_2 \left(\coth \gamma_2 l_2 + \frac{A_2}{A_1} \coth \gamma_1 l_1\right)}{\cosh \gamma_1 l_1 \cosh \gamma_2 l_2 \left[1 + \frac{A_1}{A_2} \tanh \gamma_1 l_1 \tanh \gamma_2 l_2\right]} \tag{4.53}$$

The zeros of transmission occur at frequencies which make the numerator zero, and therefore satisfy

$$\coth \gamma_2 l_2 = -\coth \gamma_1 l_1$$

or

$$\tanh \gamma_1 l_1 = -\tanh \gamma_2 l_2$$

which for lossless conditions reduces to

$$\tan \beta l_1 = -\frac{A_2}{A_1} \tan \beta l_2 \tag{4.54}$$

TO DO: Comment on the zero at zero frequency (obvious in Eq. Eq. 4.54, but not discussed in the sequelae). Demonstrate that the spectra of real fricatives and /h/ is equal to the dipole source of previous sections, pre-emphasized by the zero at zero frequency. Demonstrate that this zero at zero-frequency is present in all turbulent sounds, and that its bandwidth is proportional to the distance between the front of the constriction and the location of the noise source (Stevens [1971], Shadle [1985]).

As an example, let us use (Eq. 4.54) and (Eq. 4.51) to determine the (lossless) zeros and poles of U_m/p_t for an articulatory shape crudely representative of /s/. Take

$$\begin{array}{ll} A_1 = 7\text{cm}^2 & A_2 = 0.2\text{cm}^2 \\ l_1 = 12.5\text{cm} & l_2 = 2.5\text{cm}. \end{array}$$

The pertinent reactance functions are plotted in Fig. Fig. 4.28, and the poles and zeros so determined are listed.

The lower poles and zeros lie relatively close and essentially nullify one another. The first significant uncompensated zero lies in the vicinity of 3400 Hz, with the first uncompensated pole in the neighborhood of 6650 Hz. These two features, as well as the near-cancelling pole-zero pairs, can often be seen in the spectra of real /s/ sounds. For example, Fig. Fig. 4.29 shows two measurements of the natural speech

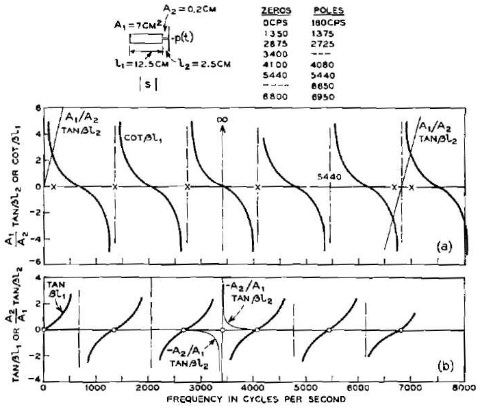


Figure 4.28: Two-tube approximation to the fricative /s/. The undamped pole-zero locations are obtained from the reactance plots

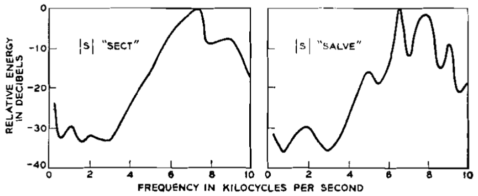


Figure 4.29: Measured spectra for the fricative /s/ in real speech. (After Hughes and Halle (Halle *et al.* [1957]))

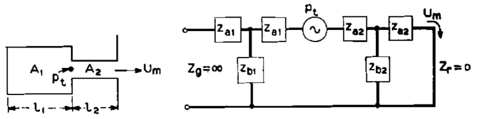


Figure 4.30: Two-tube approximation to the vocal tract with the source of excitation applied at the tube junction

fricative /s/ (Halle *et al.* [1957])). For this speaker, the peak in the vicinity of 6000-7000 Hz would appear to correspond with the uncompensated pole, the dip in the vicinity of 3000 Hz with the zero. The peak and valley alternations at the lower frequencies reflect roughly the effect of pole-zero pairs such as indicated in the reactance diagrams. The measured spectra presumably include the transformation from mouth volume current to pressure at a fixed point in space, as described in Eq. (Eq. 4.7). The spectra therefore include a zero at zero frequency owing to the radiation.

To further examine the influence of source position upon the transmission, suppose the turbulent source is applied more nearly at the junction between the two tubes rather than at the outlet. This situation is crudely representative of sounds like /f/, /k/ or possibly /ʃ/. In /f/, for example, the turbulent flow is produced at the constriction formed by the upper teeth and lower lip. The cavities behind the teeth are large, and the lips forward of the constriction form a short, small-area tube. The circuit for such an arrangement is shown in Fig. Fig. 4.30. The transmission from source to mouth is

$$\frac{U_m}{p_t} = \frac{z_{b2}}{z_{b2}(z_{a1}+z_{a2}+z_{b1}) + z_{a1}(z_{b2}+z_{a1}+z_{a2}+z_{b1})}$$

or

$$\frac{U_m}{p_t} = \frac{\frac{1}{Z_{01}}\sinh\gamma_1 l_1}{\cosh\gamma_1 l_1 \cosh\gamma_2 l_2 \left[1+\frac{A_1}{A_2}\tanh\gamma_1 l_1 \tanh\gamma_2 l_2\right]} \tag{4.55}$$

The system poles are the same as before, but the zeros now occur at

$$\frac{1}{Z_{01}}\sinh\gamma_1 l_1 = 0,$$

or

$$s_m = \left(-\alpha_1 c \pm j\frac{m\pi c}{l_1}\right); \quad m = 0, 1, 2, \ldots \tag{4.56}$$

Again for the lossless case, the zeros occur for $\sin\beta l_1 = 0$, or for frequencies

$$f_m = m\frac{c}{2l_1}\text{Hz} \quad (m = 0, 1, 2, \ldots),$$

where the length of the back cavity is an integral number of half wavelengths. The zeros therefore occur in complex-conjugate pairs except for $m = 0$. The real-axis zero

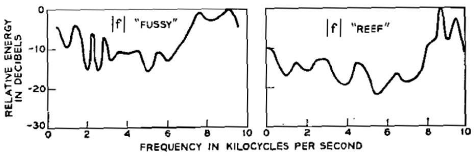


Figure 4.31: Measured spectra for the fricative /f/ in real speech. (After Hughes and Halle (Halle *et al.* [1957]))

arises from the impedance of the back cavity volume at zero frequency. Specifically, for the lossless situation at low frequencies, the numerator of (Eq. 4.55) approaches

$$\lim_{\omega \to 0} \frac{1}{Z_{01}} \sin \beta l_1 \approx \frac{\omega l_1}{Z_{01} c} = \frac{A_1 l_1}{\rho c^2} \omega = \omega C_1, \quad \text{where } C_1 = \frac{V_1}{\rho c^2}$$

is the acoustic compliance of the back cavity. The result (Eq. 4.56) makes clear the reason that a labio-dental fricative such as /f/ exhibits a relatively uniform spectrum (devoid of large maxima and minima) over most of the audible frequency range. A crude approximation to the articulatory configuration for /f/ might be obtained if the parameters of Fig. Fig. 4.30 are taken as follows: $A_l = 7$ cm^2, $A_2 = 0.1$ cm^2, $l_1 = 14$cm, $l_2 = 1$cm. As before the poles occur for $\cot \beta l_1 = A_1/A_2 \tan \beta l_2$. Because of the large value of A_1/A_2 and the small value of l_2, the poles occur very nearly at the frequencies which make $\cot \beta l_1$ infinite; namely

$$f_n \approx n \frac{c}{2l_1}, \quad n = 0, 1, 2, \ldots$$

(The first infinite value of $\tan \beta l_2$ occurs at the frequency $c/4l_2$, in the vicinity of 8500 Hz.) The zeros, according to (Eq. 4.56), occur precisely at the frequencies

$$f_m = m \frac{c}{2l_1}, \quad m = 0, 1, 2, \ldots$$

so that each pole is very nearly cancelled by a zero. The transmission U_m/P_t is therefore relatively constant until frequencies are reached where the value of $A_1/A_2 \tan \beta l_2$ has its second zero. This relative flatness is generally exhibited in the measured spectra of real /f/ sounds such as shown in Fig. Fig. 4.31 (Halle *et al.* [1957]).

4.7.6 Effects of the Nasal Tract

This highly simplified and approximate discussion of vocal transmission has so far neglected the properties of the nasal tract. The nasal tract is called into play for the production of nasal consonants and for nasalizing certain sounds primarily radiated from the mouth. Both of these classes of sounds are voiced. For the nasal consonants, an oral closure is made, the velum is opened and the sound is radiated chiefly from

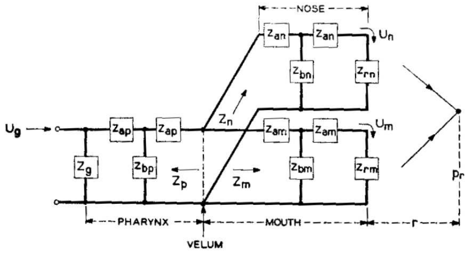


Figure 4.32: An equivalent circuit for the combined vocal and nasal tracts. The pharynx, mouth and nasal cavities are assumed to be uniform tubes.

the nostrils. The blocked oral cavity acts as a side branch resonator. In producing a nasalized vowel, on the other hand, coupling to the nasal tract is introduced by opening the velum while the major radiation of sound continues from the mouth. Some radiation, usually lower in intensity, takes place from the nostrils.

The functioning of the combined vocal and nasal tracts is difficult to treat analytically. The coupled cavities represent a relatively complex system. Precise calculation of their interactions can best be done by analog or digital computer simulation. Nevertheless, it is possible to illustrate computationally certain gross features of the system by making simplifying approximations. More specifically, suppose the pharynx cavity, mouth cavity and nasal cavity are each approximated as uniform tubes. The equivalent network is shown in Fig. Fig. 4.32.

Notice that, in general, the parallel branching of the system at the velum causes zeros of nasal output at frequencies where the driving point impedance (Z_m) of the mouth cavity is zero, and vice versa. At such frequencies, one branch traps all the velar volume flow. In particular for nasal consonants, /m,n,ŋ/, $Z_{rm} = \infty$ and $U_m = 0$. Zeros then occur in the nasal output at frequencies for which $Z_m = 0$ for the closed oral cavity. Nasal consonants and nasalized vowels are generally characterized by resonances which appear somewhat broader, or more highly damped, than those for vowels. Additional loss is contributed by the nasal tract which over a part of its length is partitioned longitudinally. Its inner surface is convoluted, and the cavity exhibits a relatively large ratio of surface area to cross-sectional area. Viscous and heat conduction losses are therefore commensurately larger.

Following the approach used earlier, and with the purpose of indicating the origin of the poles and zeros of a nasal consonant, let us make a crude, simple approximation to the vocal configuration for /m/. Such an approximation is illustrated in Fig. Fig. 4.33. The poles of the nasal output will be determined by the combined pharynx, mouth and nasal cavities, while the side-branch resonator-formed by the closed oral cavity will introduce zeros wherever its input impedance is zero. Considering the system to be lossless, the radiation load to be negligible, and the glottal impedance to be high, the easiest way to estimate the pole frequencies is to find the frequencies where the velar admittance (at the point where the three cavities join) is

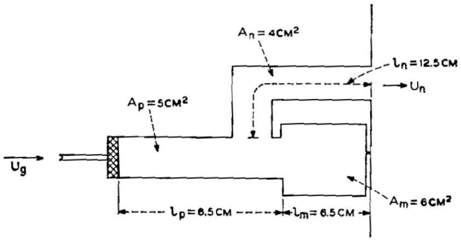


Figure 4.33: A simple approximation to the vocal configuration for the nasal consonant /m/

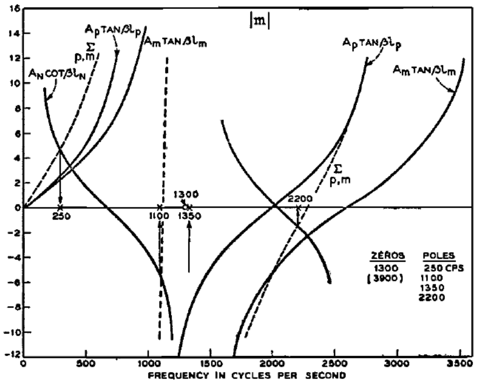


Figure 4.34: Reactance functions and undamped mode pattern for the articulatory approximation to /m/ shown in Fig. Fig. 4.33

zero. This requires

$$\sum_{k=p,m,n} Y_k = 0 = \frac{1}{Z_{0m}} \tan \beta l_m + \frac{1}{Z_{0p}} \tan \beta l_p - \frac{1}{Z_{0n}} \cot \beta l_n \tag{4.57}$$

$$= A_m \tan \beta l_m + A_p \tan \beta l_p - A_n \cot \beta l_n.$$

The zeros of transmission occur for

$$Z_m = 0 = \frac{\rho c}{A_m} \cot \beta l_m$$

or

$$\beta l_m = (2n+1)\frac{\pi}{2}, \quad n = 0, 1, 2, \ldots$$

or

$$f = (2n+1)\frac{c}{4 l_m}. \tag{4.58}$$

The mode pattern determined by relations (Eq. 4.57) and (Eq. 4.58) is shown in Fig. Fig. 4.34. One sees that the first pole of the coupled systems is fairly low, owing to the substantial length of the pharynx and nasal tract and the mouth volume. A

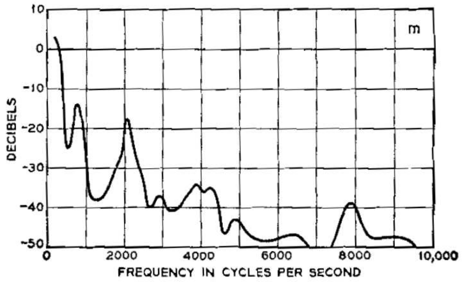


Figure 4.35: Measured spectrum for the nasal consonant /m/ in real speech. (After Fant, 1960)

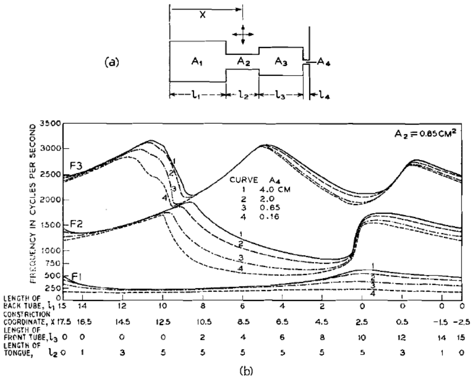


Figure 4.36: Nomogram for the first three undamped modes (F_1, F_2, F_3) of a four-tube approximation to the vocal tract (Data adapted from Fant, 1960). The parameter is the mouth area, A_4. Curves 1, 2, 3 and 4 represent mouth areas of 4, 2, 0.65 and 0.16 cm^2, respectively. Constant quantities are $A_l = A_3 = 8$ cm^2, $l_4 = 1$cm and $A_2 = 0.65$ cm^2. Abscissa lengths are in cm

pole and zero, additional to the poles of the pure vowel articulation, are introduced in the region of 1000 Hz. This mode pattern is roughly representative of all the nasal consonants in that the pharynx and nasal tract have roughly the same shape for all. The first zero falls at approximately 1300 Hz in the present example. For the consonants /n/ and /ŋ/, the oral cavity is progressively shorter, and the zero would be expected to move somewhat higher in frquency. By way of comparison, the measured spectrum of a real /m/ is shown in Fig. Fig. 4.35 (Fant [1960]). In this measured spectrum, the nasal zero appears to be reflected by the relatively broad spectral minimum near 1200 Hz. The larger damping and appreciable diminution of spectral amplitude at the higher frequencies is characteristic of the nasal consonants.

4.7.7 Four-Tube, Three-Parameter Approximation of Vowel Production

To illustrate fundamental relations, the preceding sections have dealt with very simple approximations to the vocal system. Clearly these crude representations are not adequate to describe the gamut of articulatory configurations employed in a language. The approximations can obviously be made better by quantizing the vocal system into more and shorter tube sections. For vowel production in particular, one generally can identify four main features in the tract geometry. These are the back pharynx cavity, the tongue hump constriction, the forward mouth cavity and the lip constriction (see Fig. Fig. 4.1). Approximation of these features by four abutting tubes gives a description of vocal transmission substantially more precise than the two-tube approximation. The first several normal modes of the four-tube model are reasonably good approximations to the lower formants of real vowels. Such a fourtube model is illustrated in Fig. Fig. 4.36a (adapted from (Fant [1960])).

If the glottal impedance is taken as large and the radiation load small, the glottal-to-mouth transmission is

$$\frac{U_m}{U_g} = \frac{1}{\prod_{n=1}^{4} (\cosh \gamma_n l_n)\,(ab+cd)} \tag{4.59}$$

where

$$\begin{aligned} a &= \left(1 + \frac{A_1}{A_2} \tanh \gamma_1 l_1 \tanh \gamma_2 l_2\right) \\ b &= \left(1 + \frac{A_3}{A_4} \tanh \gamma_3 l_3 \tanh \gamma_4 l_3\right) \\ c &= \frac{A_2}{A_3}\left(\tanh \gamma_3 l_3 + \frac{A_3}{A_4} \tanh \gamma_4 l_4\right) \\ d &= \frac{A_1}{A_2}(\tanh \gamma_1 l_1 + \tanh \gamma_2 l_2) \end{aligned}$$

One notices that if $l_3 = l_4 = 0$, Eq. (Eq. 4.59) reduces to the two-tube relations given by Eq. (Eq. 4.49).

To demonstrate how the first several normal modes of such a cavity arrangement depend upon configuration, Fant (Fant [1960]) has worked out detailed nomograms for several combinations of A's and l's. One of these is particularly relevant and essentially depicts the scheme followed by Dunn (Dunn [1950]) in his development of an electrical vocal tract analog. It is reproduced in adapted form in Fig. Fig. 4.36b. The constraints are as follows: $l_1 + l_2 + l_3 = 15$cm; $l_4 = 1$cm; $A_1 = A_3 = 8$ cm^2; $A_2 = 0.65$cm^2; and $l_2 = 5$cm, provided tube 2 is terminated by cavities on both sides. The parameters are the distance from the glottis to the center of the tongue constriction, x, and the mouth area, A_4. For very large and very small values of x,

l_3 and l_1 are zero, respectively, and the length l_2 is varied to satisfy the total length condition. The variation of the first three normal modes for a range of values of the parameters and for one value of the tongue constriction ($A_2 = 0.65$ cm^2) are shown in Fig. Fig. 4.36b.

These data show that a shift of the tongue constriction from a back ($x \approx 3$cm) to a front position ($x \approx 9$cm) is generally associated with a transition from high $F1$-low $F2$ to low $F1$-high $F2$. (This general tendency was also evident in the two-tube models discussed in Section 4.7.4.) Increasing the lip rounding, that is decreasing A_4 (as well as increasing l_4), generally reduces the frequencies of all formants. Although not shown here, decreasing the tongue constriction reduces the frequency variations of the formants with place of constriction. In terms of absolute frequency, the variations in Fl are generally smaller than those of the higher formants. Perceptually, however, the percentage change in formant frequency is more nearly the important quantity. This point will be discussed further in Chapter 9.

Owing to the substantial coupling between the connecting tubes, a particular formant cannot be strictly associated with a particular resonance of a particular vocal cavity. The normal mode pattern is a characteristic of the whole coupled system. Numerous efforts have been made in the literature to relate specific formants to specific vocal cavities, but this can be done exactly only when the constrictions are so small in size that the cavities are, in effect, uncoupled. In instances where the coupling is small, it is possible to loosely associate a given formant with a particular resonator. The treachery of the association, however, can be simply illustrated. If a forward motion of the tongue hump causes a resonant frequency to rise–for example, $F2$ for $3 < x < 9$cm in Fig. Fig. 4.36–the suggestion is that the resonance is mainly influenced by a cavity of diminishing length, in this case the mouth cavity. On the other hand, the same resonance might be caused to rise in frequency by a tongue retraction and a consequent shortening of the pharynx cavity-for example, $F2$ for $16 > x > 13$cm. It is therefore clear that a given formant may be principally dependent upon different cavities at different times. It can change its cavity-mode affiliation with changes in vocal configuration. In fact, its dependence upon the mode of vibration of a particular cavity may vary.

The four-tube approximation to vowel production implies that vowel articulation might be grossly described in terms of three parameters, namely, the distance from the glottis to the tongue-hump constriction, x; the size of the tongue constriction, A_2; and a measure of lip rounding such as the area-to-length ratio for the lip tube, A_4/l_4. This basis notion has long been used qualitatively by phoneticians to describe vowel production. It has been cast into quantitative frameworks by Dunn (Dunn [1950]), Stevens and House (Stevens and House [1955]), Fant (Fant [1960]) and Coker (Coker [1968]), in connection with work on models of the vocal mechanism. As pointed out earlier, Dunn has used the scheme much as represented in Fig. Fig. 4.36, that is, with constant-area tubes approximating the tract adjacent to the constriction. Stevens and House and Fant have extended the scheme by specifying constraints on the taper of the vocal tract in the vicinity of the constriction. Stevens and House use a parabolic function for the area variation, and Fant uses a section of a catenoidal horn (i.e., a hyperbolic area variation). Both use fixed dimensions for the larynx tube and the lower pharynx. In perceptual experiments with synthetic vowels, Stevens and House find that a reasonably unique relation exists between the allowed values of x, A_2 and A_4/l_4 and the first three vowel formants. Although these three parameters provide an adequate description of most nonnasal, nonretroflex, vowel articulations, it is clear that they are not generally sufficient for describing consonant and nasal configurations.

Later work by Coker (Coker [1972]) has aimed at a more detailed and physiologically meaningful description of the vocal area function. Coker's articulatory model is specified by seven, relatively-orthogonal parameters: the $x-y$ position coordinates of the tongue body; the degree and the place of the tongue tip constriction; the mouth area; the lip protrusion; and the degree of velar (nasal) coupling. Each parameter has an associated time constant representative of its vocal feature. This articulatory model has been used as the synthesis element in an automatic system for converting printed text into synthetic speech (Coker *et al.* [1971])[12].

4.7.8 Multitube Approximations and Electrical Analogs of the Vocal Tract

As the number of elemental tubes used to approximate the vocal shape becomes large, the computational complexities increase. One generally resorts to analog or digital aids in solving the network when the number of approximating sections exceeds about four. In early work analog electrical circuitry has proven a useful tool for simulating both vocal and nasal tracts. It has been used extensively by Dunn (Dunn [1950]); Stevens, Fant, and Kasowski (Stevens *et al.* [1953]); Fant (Fant [1960]); Stevens and House (Stevens and House [1955]); and Rosen (Rosen [1958]). The idea is first to approximate the linear properties of the vocal mechanism by a sufficiently large number of tube sections and then to approximate, in terms of lumped-constant electrical elements, the hyperbolic impedances of the equivalent T or π networks shown in Fig. Fig. **??**. At low frequencies the lumped-constant circuit behaves as a distributed transmission line and simulates the one-dimensional acoustic wave propagation in the vocal tract. The number of approximating tube sections used, the approximation of the hyperbolic elements, and the effect of cross modes in the actual vocal tract determine the highest frequency for which the electrical transmission line is an adequate analog.

As shown previously, the elements of the T-section equivalent of the cylindrical tube are

$$z_a = Z_0 \tanh \frac{\gamma l}{2} \quad \text{and} \quad z_b = Z_0 \mathrm{csch} \gamma l.$$

---

[12] See further discussion of this system in Chapters 5 and 11.

Taking first-order approximations to these quantities gives

$$z_a \approx Z_0\left(\frac{\gamma l}{2}\right) \quad \text{and} \quad z_b \approx Z_0\left(\frac{1}{\gamma l}\right)$$

$$z_a \approx Z_0\frac{1}{2}(\alpha_j\beta)l \qquad z_b \approx Z_0\frac{1}{(\alpha+j\beta)l}. \tag{4.60}$$

From the relations developed earlier, $Z_0 = [(R+j\omega L)/(G+j\omega C)]^{\frac{1}{2}}$ and $\gamma = [(R+j\omega L)(G+j\omega C)]^{\frac{1}{2}}$, where R, G, L and C have been given in terms of per-unit-length acoustical quantities in Eq. (Eq. 3.46). The T-elements are therefore approximately

$$z_a = \frac{1}{2}(R+j\omega L)l \quad \text{and} \quad z_b = \frac{1}{(G+j\omega C)l}.$$

In general, the acoustical quantities R_a, L_a, G_a, and C_a [in Eq. (Eq. 3.46)] will not correspond to practical electrical values. It is usually convenient to scale the acoustical and electrical impedance levels so that

$$Z_{0e} = kZ_{0a}$$

or

$$\left[\frac{R_e+j\omega L_e}{G_e+j\omega C_e}\right]^{\frac{1}{2}} = \left[\frac{kR_a+j\omega kL_a}{\frac{G_a}{k}+\frac{j\omega C_a}{k}}\right]^{\frac{1}{2}}. \tag{4.61}$$

By way of indicating the size of a practical scale constant k, consider the low-loss situation where

$$Z_{0e} = \sqrt{\frac{L_e}{C_e}} = kZ_{0a} = k\sqrt{\frac{L_a}{C_a}} = k\left(\frac{\rho c}{A}\right), \tag{4.62}$$

where A is the cross-sectional area of the acoustic tube. A practical value for Z_{0e} is 600 electrical ohms, and a typical value of A is 8 cm^2. Therefore $k = 600/5.3 = 113$, and the mks impedances of the per-unitlength electrical elements are scaled up by 113 times the cgs impedances of the per-unit-length acoustic elements.

Note, too, that $\beta l \approx \omega l/c = \omega l_e\sqrt{L_eC_e} = \omega l_a\sqrt{L_aC_a}$. Since the velocity of sound and the air density in a given length of tube are constant, maintaining the L_eC_e product constant in the electrical line is equivalent to maintaining constant velocity of sound propagation in the simulated pipe. Similarly, changes in the pipe area A are represented by proportional changes in the C_e/L_e ratio.

The electrical simulation is of course applicable to both vocal and nasal tracts. Choice of the elemental cylinder length l, the electrical scale constant k, and a knowledge of the cross-sectional area A along the tract are the only parameters needed to determine the lossless elements of the transmission line. An estimate of tract circumference along its length is needed to compute the viscous and heat conduction losses (R and G). The radiation loads at the mouth and nostrils are obtained hy applying the electrical scale constant to the acoustic radiation impedances obtained earlier in the chapter. It is likewise possible to apply these techniques to the subglottal system and to incorporate it into the electrical simulation. At least four designs of electrical vocal tracts have been developed for studying vocal transmission and for synthesizing speech (Dunn [1950], Stevens *et al.* [1953], Fant [1960], Rosen [1958]). At least one design has been described for the subglottal system (van den Berg [1960]).

The equations used to create electrical circuit simulations of the vocal tract may also be used to implement a vocal tract simulation on a computer. Simulations using the equations described above have been published by (Fant [1960], Mathews and Walker [1962]). Another approach has been to represent the cylindrical sections in terms of the reflection coefficients at their junctions (Jr. and Lochbaum [1962a,b], Mermelstein [1967], Purves *et al.* [1970]). Vocal tract simulation in terms of reflection coefficients is closely related to linear predictive analysis of the speech waveform, and will be considered in considerably more detail in chapter 5.

TO DO: A third method for digital simulation of vocal tract transmission was proposed by Sondhi and Schroeter (Sondhi and Schroeter [1987]), and elaborated by Lin (Lin [1990]). In this method, each of the T sections is represented as a matrix transfer function of the following form.... the relationship of flow and pressure at the lips to flow and pressure at the glottis is therefore given by... the vocal tract transfer function can therefore be computed at any desired number of frequency samples using equations... a vocal tract transfer function computed in this way can be inverse transformed, and convolved with the time-domain waveform of the glottal source, in order to synthesize speech...

4.8 Fundamentals of Speech and Hearing in Analysis-Synthesis Telephony

The preceding sections have set forth certain basic acoustic principles for the vocal mechanism. Not only do these relations concisely describe the physical behavior of the source of speech signals, but they imply a good deal about efficient communication. They suggest possibilities for coding speech information in forms other than merely the transduced pressure wave. The normal mode and excitation relations, for example, indicate a schema on which an analysis-synthesis transmission system might be based. The same can be said for describing the vocal tract by articulatory parameters. Both results reflect constraints peculiar to the speech-producing mechanism.

As yet, however, the properties of hearing and the constraints exhibited by the ear have not entered the discussion. The next chapter proposes to establish certain fundamental properties of the mechanism of hearing–so far as they are known. The exposition will follow a pattern similar to that of the present chapter. The results of both fundamental discussions will then be useful in subsequent consideration of speech analysis and speech synthesis.

4.9 Homework

Problem 4.1

The *intensity* of an acoustic wave is the average product of its acoustic pressure, $p(t)$, times its acoustic air particle velocity, $u(t)$:

$$I = E\left[p(t)u(t)\right]$$

where $p(t)$ is measured in Pascals, and $u(t)$ is measured in m/s. What are the units of intensity? How does intensity relate to power?

Problem 4.2

The back of a loudspeaker is typically encased in a vented cabinet. A “vented” cabinet is a wooden box with the loudspeaker on one side, and a hole on the other side. The simplest model of a vented cabinet is a one-dimensional resonator, with the loudspeaker mounted at $x = -L$, and the vent open at $x = 0$. The boundary conditions are

$$\begin{aligned} U(x=-L,\Omega) &= -U_s(\Omega) \\ P(x=0,\Omega) &= 0 \end{aligned}$$

where $U_s(\Omega)$ is the loudspeaker velocity. Assume that the area of the loudspeaker and the area of the vent are both A. Then the total volume velocity radiating out of the cabinet is

$$Q_s(\Omega) = A(U_s(\Omega) + U(0,\Omega))$$

Find the “cabinet transfer function” $Q_s(\Omega)/U_s(\Omega)$. Sketch $Q_s(\Omega)/U_s(\Omega)$ for $0 \le \Omega \le 2\pi c/L$. Over what range of frequencies is $\frac{Q_s(\Omega)}{U_s(\Omega)} > 0$?

Problem 4.3

Glottal vibration is the result of two kinds of forces: aerodynamic forces, and stiffness and damping of the vocal fold. For now, let us only consider the stiffness of the vocal fold:

$$\frac{d^2x}{dt^2} = -\left(\frac{k}{m}\right)x \tag{4.63}$$

In acoustic terms, Eq. 4.63 governs the undriven, undamped, collision-free vocal fold behavior—undriven because we are ignoring aerodynamic forces, undamped because we are ignoring the viscosity of the tissue, and collision-free because we assume that the two vocal folds are far enough apart to avoid collision. Eq. 4.63 is a good place to start our understanding of vocal fold mechanics, because it isolates the “control knob" that most talkers use, most of the time, to control pitch: the stiffness, k, of the vocal fold. Stiffness of the vocal fold can be increased by stretching it; stiffness can be decreased by shortening the vocal fold.

a. Demonstrate that, with no driving forces and no damping and no collisions, the vocal folds can vibrate forever. Hint: show that $x(t) = A\cos(\omega t - \phi)$ is a solution to Eq. 4.63.

b. The moving part of the vocal fold is a ribbon of tissue about 1cm long, about 3mm deep, and about 1mm wide. This ribbon of tissue has the density of water (1 gram/cm^3). What is its total mass?

c. Suppose that a particular talker speaks at 200Hz. According to the model given in Eq. 4.63, what is the stiffness of her vocal folds? Be sure to tell me what the units are.

d. Assume that the vocal fold displacement, $x(t)$, has an amplitude of 1mm. Plot (by hand or using any program of your choice) one full period of the vocal fold displacement $x(t)$, of the vocal fold velocity dx/dt, and of the stiffness force $f(t) = -kx(t)$.

e. Now suppose that the same talker increases her pitch to 300Hz (an increase in pitch of one musical fifth), without changing her vocal fold mass. What is the new value of her vocal fold stiffness?

Problem 4.4

During production of the vowel /ɑ/, the pharynx is quite narrow (about 1cm^2), while the oral cavity is quite wide (about 8cm^2). Let the boundary between these two parts of the vocal tract be called $x = 0$.

a. Draw a schematic picture of this situation.

b. Pressure $p(x,t)$ (in Pascals) and volume velocity $u(x,t)$ (in liters/second) must be continuous across the boundary, i.e.

$$\begin{aligned} p(0_-, t) &= p(0_+, t) && (4.64) \\ u(0_-, t) &= u(0_+, t) && (4.65) \end{aligned}$$

Re-write Eqs. 4.64 and 4.65 in terms of the forward-going and backward-going waves, whose phasors are p_{1+}, p_{1-}, p_{2+}, and p_{2-}.

c. Show that the outgoing waves from, p_{2+} and p_{1-}, may be written in terms of the incoming waves, p_{2-} and p_{1+}, and in terms of a reflection coefficient γ. Write γ in terms of the front cavity and back cavity areas.

d. Suppose that the glottis is a perfect source, i.e., regardless of what the backward-going wave p_{1-} may be, the forward-going wave is always a perfect cosine $p_{1+} = 1$. Find the forward-going and backward-going waves in the front cavity, p_{2+}, and p_{2-}, as a function of the front cavity length L_f, and the reflection coefficient γ. Assume a zero-pressure termination at the lips.

e. Find the air velocity at the lips, $v(L_f, \omega)$, as a function of L_f, ω, and γ. Assume that $p_{1+} = 1$ at all frequencies.

f. Plot $v(L_f, \omega)$ as a function of ω.

Problem 4.5

Assume a perfectly decoupled back and front cavity, where $A_b \gg A_f$. Assume that $L_b + L_f = 17$cm. Calculate the first three formant frequencies for the following back cavity lengths: $L_b \in \{1, 3, 5, 7, 9, 11, 13, 15\}$cm. Remember to consider the Helmholtz resonance. Plot F_1, F_2, and F_3 (in Hertz) on the same axes, as a function of L_b. This plot is called a "nomogram;" it is considered by many to be a convenient summary of the relationship between vocal tract shape and vowel quality.

Problem 4.6

The vocal tract configuration for an /r/ consonant, in American English, is roughly as follows: the back cavity, behind the tongue tip constriction, has a length of about 15cm, with a relatively large average cross-sectional area (about 10cm^2). The front cavity, between the tongue tip constriction and the lips, has a length of about 4cm, and a cross-sectional area of about 6cm^2. The side branch, under the tongue, has a length of about 4cm, and a cross-sectional area of about 16.5cm^2.

a. Sketch the three-tube model for /r/. Label all areas and lengths of all tube sections, and be sure to show whether each tube is closed or open.

b. Find F_1 of the /r/ configuration, using the low-frequency approximations given in the lecture notes.

c. Find F_1 of the neighboring vowel: set the length of the side branch to 0cm, and then solve the same equations that you solved in part (b). How does your answer compare to part (b)? How does your answer compare to the first formant of a schwa?

d. Find the first zero frequency of the /r/.

e. Assume that the first zero is part of a "pole-zero pair." In other words, you can imagine that the first zero splits the nearest oral formant into a pole-zero-pole complex, with the first pole about 200Hz below the zero, and the second pole about 200Hz above the zero. In that case, what is the frequency of F_3 of an /r/?

Problem 4.7

The speed of sound:

a. What is the formula for the speed of sound? Solution: The formula is $c = \sqrt{\eta P_0/\rho_0}$.

b. Identify the variables: Names, units, and values of consonants. Solution: On the left is the speed of sound c, which is a derived quantity, and on the right are physical parameters $\eta \equiv c_p/c_v = 1.4$ [dimentionless], the barometric pressure (e.g., at sea level) $P_0 = 10^5$ [Pa], and the density ρ_0 [kgm/m^3].

c. What is the meaning of ηP_0? Solution: This combination of variables represents the adiabatic compressibility of air. The $\eta = c_p/c_v$ results from holding the temperature constant during the cycle of the wave. Heat diffusion is slow compared to the cycle at acoustic frequencies, such that the thermal energy is trapped in the air.

d. Does P_0 depend on temperature? Explain? Solution: The barometric pressure depends on the weight of the air above us. While its density dependents on temperature, the total weight is a constant, and is therefore independent of temperature. That leaves the question as to why the barometric pressure varies, but that is more about the humidity, winds, and other variables.

e. Does ρ_0 depend on temperature or P_0? Explain. Solution: Yes, as given in class

$$\rho(T, P_0) = 1.29 \frac{273}{273 + C} \frac{P_0}{10^5}, \tag{4.66}$$

where C is the temperature in degrees Celsius.

f. What is the form of the dependence of the speed of sound on temperature? Namely give the formula for $c(T)$, and explain the dependence. Solution: Since $c = sqrt\eta P_0/\rho_0$ and following the state equation for a gas, $\rho_0 \propto 1/T$, we may conclude that $c(T) \propto \sqrt{T}$, where T is in degrees Kelvin. Because the percentage change in degrees Kelvin is much smaller than in degree C, this dependence is relatively small.

Problem 4.8

Basic equations of sound propagation:

a. Write out the 2x2 matrix equation that describes the propagation of 1 dimensional sound waves in a tube having area A: Solution: As shown in class the basic equations are:

$$\frac{d}{dx}\begin{bmatrix} P(x,\omega) \\ U(x,\omega) \end{bmatrix} = -\begin{bmatrix} 0 & \mathcal{Z}(x,s) \\ \mathcal{Y}(x,s) & 0 \end{bmatrix}\begin{bmatrix} P(x,\omega) \\ U(x,\omega) \end{bmatrix}. \tag{4.67}$$

where $\mathcal{Z} = s\rho_0/A$ and $\mathcal{Y} = sA/\eta P_0$, with A the area of the tube.

b. Rewrite these equations as a second order equation in terms of the pressure P, and thereby find the formula for the speed of sound in terms of Z and Y: Solution: If we let P' to mean the partial with respect to space, then

$$P' + ZU = 0 \tag{4.68}$$

and

$$U' + YP = 0. \tag{4.69}$$

Taking the partial wrt x of the first equation, and then using the second, gives

$$P'' + ZU' = P'' - \mathcal{Z}\mathcal{Y}P = 0. \tag{4.70}$$

Since the wave equation is

$$\frac{\partial^2 P}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 P}{\partial x^2} \tag{4.71}$$

by inspection we see that $\omega^2/c^2 = \mathcal{Z}\mathcal{Y}$, which results in the final formula for the speed of sound.

Problem 4.9

Assume that a flash bulb puts out 50 Joules when it is triggered, and it lasts for 20 μs. How much power is delivered while the flash is present, assuming the illumination is constant during the interval?

Problem 4.10

A person is speaking at an intensity of I_0 66 dB-SPL, as measured with a sound level meter at 1 meter.

a. Find the total power in the voice assuming that the level is uniform around the head. Solution: First, 66 dB SPL is equal to a pressure of $0.04 = 20 \times 10^{-6} 10^{66/20}$. This corresponds to an intensity (power/area) of $I_0 \equiv |P^2|/\rho_0 c = 0.04^2/407 = 3.93 \times 10^{-6}$ [Watts/m^2]. The area of the sphere at 1 meter is $4\pi r^2 = 4\pi$. Thus the total speech power is about $16\pi \approx 50$ [μW].

b. Find the total power assuming that the intensity varies as

$$I(\theta, \phi) = I_0 \cos(\theta/2) \cos(\phi/2) \tag{4.72}$$

where θ is the angle in the horizontal plane and ϕ in the vertical plane. where $\theta = 0, \phi = 0$ corresponds to "straight ahead."

Solution: The total power $\mathcal{P}$ is given by the integral of the intensity over the area:

$$\mathcal{P} = I_0 r^2 \times \int_{-\pi/2}^{\pi/2} \cos(\phi/2) \cos(\phi) d\phi \times \int_{-\pi}^{\pi} \cos(\theta/2) \cos(\theta) d\theta.$$

Thus

$$\mathcal{P} = r^2 I_0 \times [2 \sin(\theta/2)]_{-\pi}^{\pi} \times \left[\frac{1}{3} \sin(3\phi/2) + \sin(\phi/2)\right]_{-\pi/2}^{\pi/2}$$

Total power is about 60% of part a.
As a check, make sure that when the area is uniform $\mathcal{P} = I_0 A$ with $A = 4\pi r^2 = 4\pi$.

Problem 4.11

Bels:

a. How many millibels [mB] in 1 bel [B]? Solution: 1000 mB = 1 B since mB is a much smaller unit that the bel.

b. Give the formula for the intensity in mB units. Solution: The intensity in mB is $1000 \log_{10}(I/I_{ref})$ where $I_{ref} = 10^{-12}$ W/m^2.

c. Give the formula for the sound pressure level in cB (centibel) units. Solution: The sound pressure level in cB is $200 \log_{10}(P/P_{ref})$ where $P_{ref} = 20 \times 10^{-6}$ [Pa].

Problem 4.12

Demonstrate that $P_{ref} \equiv 20$ μPa is the same as $I_{ref} \equiv 10^{-12}$ [W/m^2]. Solution: The first is given by the formula $|P_{ref}|^2/\rho c$ and the second is given by I_{ref}. Thus we need to show that numerically these two quantities are similar (almost identical). Since $(20 \times 10^{-6})^2/407 \approx 10^{-12}$, the two references are nearly the same.

Problem 4.13

A bottle has a neck diameter of 1 [cm] and is l =1 cm long. It is connected to the body of the bottle "barrel" which is 5 cm in diameter and L =10 cm long. Treat the barrel as a short piece of transmission line, closed at one end, which looks like a compliance $C = V_{barrel}/\eta P_0$, and the neck which look like a mass $M = \rho_0 l/A_{neck}$. These two impedances are in series, since they both see the same volume velocity (flow).

a. Find the resonant frequency of the bottle. Hint: set the impedance to zero and solve for the resonant frequency in terms of M and C.

b. write out the formula for the resonant frequency in terms of the dimensions of the bottle. Solution: The formula for the Helmholtz resonator was derived in class, where it was shown to be

$$f_0 = \frac{c}{2\pi}\sqrt{A/Vl}, \tag{4.73}$$

where A, l are the area and length of the neck and V is the volume of the bottle.

c. calculate the resonant frequency in Hz for the dimensions given. Solution: From the numbers given $A = \pi.005^2$ [m^2], $l = 0.01$ while $V = \pi 0.025^2 \times 0.1 = 1.9 \times 10^{-4}$ [m^3]. Thus $f_0 = \frac{345}{2\pi}\sqrt{0.01 \times \pi 0.005^2/0.0025} \approx 347.3$ [Hz].

d. Blow into a blottle and measure the resonant frequency by recording the tone, and taking the FFT of the resulting waveform, and finding the frequency (Extra credit).

Chapter 5

Techniques for Speech Analysis

The earlier discussion suggested that the encoding of speech information might be considered at several stages in the communication chain. On the transmitter side, the configuration and excitation of the vocal tract constitute one description. In the transmission channel, the transduced acoustic waveform is a signal representation commonly encountered. At the receiver, the mechanical motion of the basilar membrane is still another portrayal of the information. Some of these descriptions exhibit properties which might be exploited in communication.

Efforts in speech analysis and synthesis frequently aim at the efficient encoding and transmission of speech information[1]. Here the goal is the transmission of speech information over the smallest channel capacity adequate to satisfy specified perceptual criteria. Acoustical and physiological analyses of the vocal mechanism suggest certain possibilities for efficient description of the signal. Psychological and physiological experiments in hearing also outline certain bounds on perception. Although such analyses may not necessarily lead to totally optimum methods for encoding and transmission, they do bring to focus important physical constraints. Transmission economies beyond this level generally must be sought in linguistic and semantic dependencies.

The discussions in Chapters 2 and 4 set forth certain fundamental relations for the vocal mechanism. Most of the analyses presumed detailed physical knowledge of the tract. In actual communication practice, however, one generally has knowledge only of some transduced version of the acoustic signal. (That is, the speaker does not submit to measurements on his vocal tract.) The acoustic and articulatory parameters of the preceding chapters must therefore be determined from the speech signal if they are to be exploited.

This chapter proposes to discuss certain speech analysis techniques which have been found useful for deriving so-called "information-bearing elements" of speech. Subsequent chapters will consider synthesis of speech from these low information-rate

---

[1] Other motivating objectives are: basic understanding of speech communication, voice control of machines, and voice response from computers.

parameters, perceptual criteria appropriate to the processing of such parameters, and application of analysis, synthesis and perceptual results in complete transmission systems.

5.1 Spectral Analysis of Speech

Frequency-domain representation of speech information appears advantageous from two standpoints. First, acoustic analysis of the vocal mechanism shows that the normal mode or natural frequency concept permits concise description of speech sounds. Second, clear evidence exists that the ear makes a crude frequency analysis at an early stage in its processing. Presumably, then, features salient in frequency analysis are important in production and perception, and consequently hold promise for efficient coding. Experience supports this notion.

Further, the vocal mechanism is a quasi-stationary source of sound. Its excitation and normal modes change with time. Any spectral measure applicable to the speech signal should therefore reflect temporal features of perceptual significance as well as spectral features. Something other then a conventional frequency transform is indicated.

5.1.1 Short-Time Frequency Analysis

The conventional mathematical link between an aperiodic time function $f(t)$ and its complex amplitude-density spectrum $F(\omega)$ is the Fourier transform-pair

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt \tag{5.1}$$

$$f(t) = -\frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{j\omega t} d\omega$$

For the transform to exist, $\int_{\infty}^{\infty} |f(t)| dt$ must be finite. Generally, a continuous speech signal neither satisfies the existence condition nor is known over all time. The signal must consequently be modified so that its transform exists for integration over known (past) values. Further, to reflect significant temporal changes, the integration should extend only over times appropriate to the quasi-steady elements of the speech signal. Essentially what is desired is a running spectrum, with real-time as an independent variable, and in which the spectral computation is made on weighted past values of the signal.

Such a result can be obtained by analyzing a portion of the signal "seen" through a specified time window, or weighting function. The window is chosen to insure that the product of signal and window is Fourier transformable. For practical purposes, the weighting function $h(t)$ usually is the impulse response of a physically-realizable linear system. Then, $h(t) = 0$; for $t < 0$. Generally $h(t)$ is desired to be unipolar

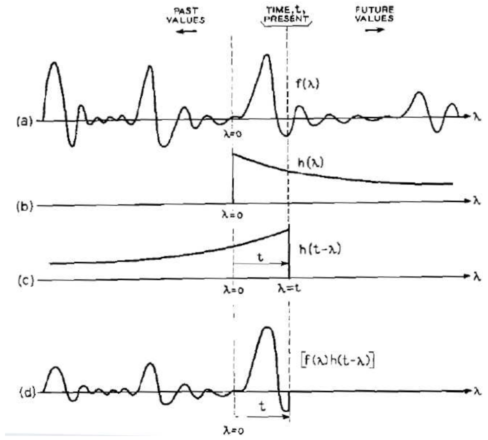


Figure 5.1: Weighting of an on-going signal $f(t)$ by a physically realizable time window $h(t)$. λ is a dummy integration variable for taking the Fourier transform at any instant, t

and is essentially the response of a low-pass filter. The Fourier transform (5.1) can therefore be modified by transforming that part of the signal seen through the window at a given instant of time. The desired operation is

$$F(\omega, t) = \int_{infty}^{t} f(\lambda) h(t-\lambda) e^{-jw\lambda} d\lambda,$$

or,

$$F(\omega, t) = e^{-j\omega t} \int_{0}^{\infty} f(t-\lambda) h(\lambda) e^{j\omega\lambda} d\lambda. \tag{5.2}$$

The signal, with its past values weighted by $h(t)$, is illustrated for a given instant, t, in Fig. 5.1.

The short-time transform, so defined, is the convolution

$$[f(t)e^{-j\omega t} * h(t)], \quad \text{or alternatively,} \quad e^{-j\omega t}[f(t) * h(t)e^{j\omega t}].$$

If the weighting function $h(t)$ is considered to have the dimension $\sec^{-1}$ (i.e., the Fourier transform of $h(t)$ is dimensionless), then $|F(\omega, t)|$ is a short-time amplitude spectrum with the same dimension as the signal. Like the conventional Fourier transform, $F(\omega, t)$ is generally complex with a magnitude and phase, namely $|F(\omega, t)| e^{-j\theta(\omega,t)}$, where $\theta(\omega, t)$ is the short-time phase spectrum. By definition, the inverse relation also holds

$$[f(\lambda)h(t-\lambda)] = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega, t) e^{j\omega\lambda} d\omega.$$

Note that at any time $t = t_1$, the product $[f(\lambda)h(t-\lambda)]$ is determined for all $\lambda \leq t_1$. If the window function $h(t_1 - \lambda)$ is known, then the original function over the interval

$-\infty \leq \lambda \leq t_1$ can be retrieved from the product. For a value of λ equal to t_1

$$[f(t)h(0)] = \frac{1}{2\pi} \int F(\omega, t_1) e^{j\omega t_1} d\omega$$

or in general for $\lambda = t$

$$f(t) = \frac{1}{2\pi h(0)} \int_{-\infty}^{\infty} F(\omega, t) e^{j\omega t} d\omega.$$

The short-time transform is therefore uniquely invertible if one nonzero value of the window function is known. Typically $h(t)$ can be chosen so that $h(0) = 1$ and

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega, t) e^{j\omega t} d\omega$$

which bears a pleasing parallel to the conventional infinite-time Fourier transform.

The inversion implies that $f(t)$ can be determined for the same points in time that $F(\omega, t)$ is known, provided $F(\omega, t)$ is known as a continuous function of frequency. However, in cases where the product function $[f(\lambda)h(t-\lambda)]$ is of finite duration in λ (say owing to a finite duration window) then samples of the waveform $f(t)$ may be recovered exactly from samples in ω of $F(\omega, t)$ (Weinstein [1966]). Discrete-frequency, continuous-time values of the short-time transform, $F(\omega, t)$, are of particular interest and will find applications in later discussions.

5.1.2 Measurement of Short-Time Spectra

We notice that (5.2) can be rewritten

$$F(\omega, t) = -\int_{-\infty}^{t} f(\lambda) \cos \omega\lambda h(t-\lambda) d\lambda - j \int_{-\infty}^{t} f(\lambda) \sin \omega\lambda h(t-\lambda) d\lambda \tag{5.3}$$

$$= \left[a(\omega, t) - jb(\omega, t)\right].$$

Further,

$$|F(\omega, t)| = [F(\omega, t)F^*(\omega, t)] = (a^2 + b^2)^{\frac{1}{2}} \tag{5.4}$$

and

$$\theta(\omega, t) = \tan^{-1} b/a.$$

where $F^*(\omega, t)$ is the complex conjugate of $F(\omega, t)$. Note that $|F(\omega, t)|$ is a scalar, whereas $F(\omega, t)F^*(\omega, t)$ is formally complex, and that $|F(\omega, t)|^2$ is the short-time power spectrum. The measurement of $|F(\omega, t)|$ can therefore be implemented by the operations shown in Fig. 5.2.

The frequency-domain interpretation of these operations is apparent. The heterodyning (or multiplication by $\cos \omega t$ and $\sin \omega t$) shifts (or translates) the spectrum

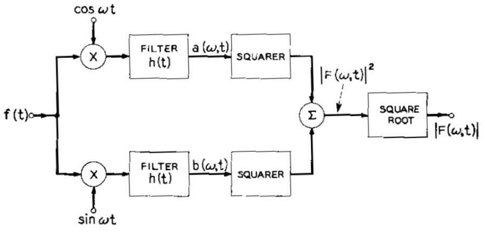


Figure 5.2: A method for measuring the short-time amplitude spectrum $|F(\omega, t)|$

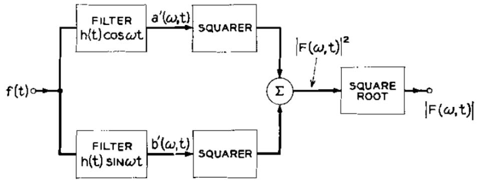


Figure 5.3: Alternative implementation for measuring the short-time amplitude spectrum $|F(\omega, t)|$

of $f(t)$ across the pass-band of filter $h(t)$. The latter is normally a low-pass structure. Frequency components of $f(t)$ lying close to ω produce difference-frequency components inside the low-pass band and yield large outputs from the $h(t)$ filter. Quadrature versions of the shifted signals are squared and added to give the short-time power spectrum $|F(\omega, t)|^2$.

Alternatively, Eq. (5.2) can be written

$$F(\omega,t) = e^{-j\omega t}\left\{\int_0^\infty f(t-\lambda)h(\lambda)\cos\omega\lambda d\lambda + j\int_0^\infty f(t-\lambda)h(\lambda)\sin\omega\lambda d\lambda\right\}$$

$$= \left[a'(\omega,t) + jb'(\omega,t)\right]e^{-j\omega t}. \qquad (5.5)$$

The alternative measurement of $|F(\omega, t)| = [a'^2 + b'^2]^{\frac{1}{2}}$ can therefore be effected by the operations in Fig. 5.3.

Again, in terms of a frequency-domain interpretation, the measurement involves filtering by phase-complementary band-pass filters centered at ω and having bandwidths twice that of the low-pass $h(t)$ function. The outputs are squared and added to produce the short-time power spectrum $|F(\omega, t)|^2$. Both filters have impulse responses whose envelopes are the time window, $h(t)$. As many pairs of filters are required as the number of frequency values for which the spectrum is desired. Notice, too, that for both methods of measurement (i.e., Fig. 5.2 and 5.3) if the input signal $f(t)$ is a unit impulse the short-time amplitude spectrum is simply $h(t)$, the weighting function.

It is common, in experimental practice, to minimize equipment complexity by making an approximation to the measurements indicated in Fig. 5.2 and 5.3. The

Figure 5.4: Practical measurement of the short-time spectrum $|F(\omega, t)|$ by means of a bandpass filter, a rectifier and a smoothing network

desired measurement $|F(\omega, t)| = [a'^2(\omega, t) + b'^2(\omega, t)]^{\frac{1}{2}}$ is essentially the time envelope of either $a'(\omega, t)$ or $b'(\omega, t)$.

The time envelope of a Fourier-transformable function $u(t)$ can be defined as

$$e(t) = \left[u^2(t) + \hat{u}^2(t)\right]^{\frac{1}{2}}, \quad \text{where } \hat{u}(t) = \left[u(t) * \frac{1}{\pi t}\right]$$

is the Hilbert transform of $u(t)$. One can show that $\widehat{u(t)v(t)} = u(t)\hat{v}(t) = \widehat{u(t)}v(t)$, provided the spectra of $u(t)$ and $v(t)$ do not overlap. Making use of these relations, and the possibilities for interchanging orders of integration in the convolutions, one notices that

$$a'(\omega, t) = [f(t) * h(t) cos\omega t] \tag{5.6}$$

$$\begin{aligned}
\hat{a}'(\omega, t) &= \left[a'(\omega, t) * \frac{1}{\pi t}\right] \\
&= f(t) * \left[h(t) \cos \omega t * \frac{1}{\pi t}\right] \\
&= f(t) * [h(t) \sin \omega t] \\
&= b'(\omega, t),
\end{aligned}$$

provided the spectrum of $h(t)$ does not overlap ω. The quantity $|F(\omega, t)|$ is therefore essentially the time envelope of either $a'(\omega, t)$ or $b'(\omega, t)$. The envelope can be approximated electrically by developing the envelope of either filter branch in Fig. 5.3. This is conventionally done by the linear rectification and low-pass filtering indicated in Fig. 5.4. If the impulse response of the low-pass filter is appropriately chosen, the output $|f(t) * p(t)| * q(t)$ approximates $|F(\omega, t)|$.

The measurement method of Fig. 5.4 is precisely the one used in the well-known Sound Spectrograph and in most filter-bank spectrum analyzers. In particular, it is usually the method used to develop the short-time spectrum in vocoders and in several techniques for automatic formant analysis. All of these applications will be discussed in further detail subsequently. As a present example, however, Fig. 5.5 shows successive short-time spectra of a voiced speech sample as produced by a bank of 24 filters. The filters are approximately 150Hz wide, and cover the frequency range 150 to 4000Hz. Each filter is followed by a rectifier and an R-C network. The filter bank is scanned every 10 msec and the short-time spectrum plotted. High-frequency

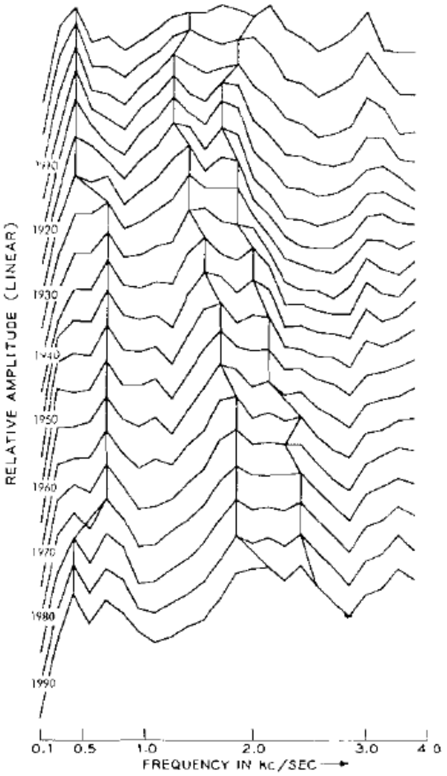


Figure 5.5: Short-time amplitude spectra of speech measured by a bank of 24 band-pass filters. A single filter channel has the configuration shown in Fig. 5.4. The spectral scans are spaced by 10 msec in time. A digital computer was used to plot the spectra and to automatically mark the formant frequencies. (After (Flanagan *et al.* [1962a]))

emphasis is used on the input signal to boost its level in the high-frequency end of the spectrum. The filter-bank output is fed into a digital computer through an analog-to-digital converter, and the spectral scans are plotted automatically by the computer (FLANAGAN, COKER, and BIRD). The lines connecting the peaks represent speech formant frequencies which were automatically determined by computer processing of the short-time spectrum.

5.1.3 Choice of the Weighting Function, $h(t)$

In speech applications, it usually is desirable for the short-time analysis to discriminate vocal properties such as voiced and unvoiced excitation, fundamental frequency, and formant structure. The choice of the analyzing time window $h(t)$ determines the compromise made between temporal and frequency resolution. A time window short in duration corresponds to a broad band-pass filter. It may yield a spectral analysis in which the temporal structure of individual vocal periods is resolved. A window with a duration of several pitch periods, on the other hand, corresponds to a narrower bandpass filter. It may produce an analysis in which individual harmonic spectral components are resolved in frequency.

In order to illustrate applicable orders of magnitude for filter widths and time windows, imagine the analyzing bandpass filter to be ideal (and nonrealizable) with a rectangular amplitude-frequency response and with zero (or exactly linear) phase response. Let the frequencydomain specification be

$$P(\omega)\begin{cases} = 1; & (\omega_0 - \omega_1) \le \omega \le (\omega_0 + \omega_1) \\ = 1; & -(\omega_0 + \omega_1) \le \omega \le -(\omega_0 - \omega_1) \\ = 0; & \text{elsewhere} \end{cases} \tag{5.7}$$

The impulse response of the filter is therefore

$$p(t) = \left(\frac{2\omega_1}{\pi}\right)\left(\frac{\sin\omega_1 t}{\omega_1 t}\right)\cos\omega_0 t \tag{5.8}$$

$$= h(t) * \cos\omega_0 t$$

and the time window for this ideal filter is the $\sin x/x$ envelope of the impulse response. If the time between initial zeros of the envelope is arbitrarily taken as the effective duration, D, of the time window, then $D = 2\pi/\omega_1 = 4\pi/\Delta\omega$, where $\Delta\omega = 2\omega_1$ is the bandwidth of the filter[2]. The D's corresponding to several $\Delta\omega$'s are

---

[2] Sometimes one-half this value is taken as the effective window duration.

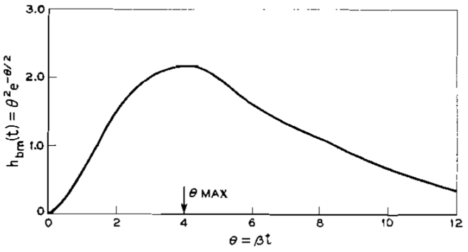


Figure 5.6: The effective time window for short-time frequency analysis by the basilar membrane in the human ear. The weighting function is deduced from the ear model discussed in Chapter IV

Condition	$\Delta\omega/2\pi$ (Hz)	D (msec)
(1)	50	40
(2)	100	20
(3)	250	8

Condition (1) is an analyzing bandwidth commonly used to resolve the harmonic spectral components in voiced portions of speech. For this bandwidth, the duration of the time window spans about four or five pitch periods of a man's voice.

The broad filter condition (3), on the other hand, produces a weighting function comparable in duration with a single pitch period of a man's voice. The time resolution of this analysis therefore resolves amplitude fluctuations whose temporal courses are of the order of a pitch period. Filter conditions analogous to both (1) and (3) are employed in the wellknown Sound Spectrograph which will be discussed in the following section.

The middle condition (2) is a sort of time-frequency compromise for speech. It is a filter width which has been found useful in devices such as vocoders and formant trackers. The short-time spectra already shown in Fig. 5.5 are representative of this resolution.

In passing, it is relevant to estimate the effective time window for the mechanical short-time analysis made by the basilar membrane in the human ear. From the earlier discussion in Chapter 7[3], a reasonably good approximation to the displacement impulse response of the basilar membrane, at a point maximally responsive to radian frequency β, is

$$p(t) = (\beta t)^2 e^{-\beta t/2} \sin \beta t = h_{bm}(t) \sin \beta t \tag{5.9}$$

The time window for the basilar membrane, according to this modeling[4], is the "surge" function plotted in Fig. 5.6. One notices that the time window has a duration

---

[3] See also the "third" model described in (Flanagan [1962a]))

[4] Eq. (5.9) does not include the effects of the middle ear. See Chapter 7 for these details.

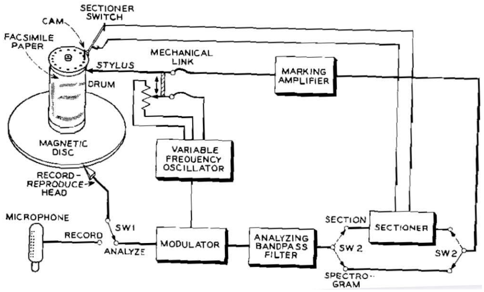


Figure 5.7: Functional diagram of the sound spectrograph

inversely related to β. It has its maximum at $t_{max} = 4/\beta$. If, as a crude estimate, $2t_{max}$ is taken as the effective duration D of the window, then for several membrane places:

$\beta/2\pi$ (Hz)	$d = 2t_{max}$ (msec)
100	12.0
1000	1.2
5000	0.2

For most speech signals, therefore, the mechanical analysis of the ear apparently provides better temporal resolution than spectral resolution. Generally, the only harmonic component resolved mechanically is the fundamental frequency of voiced segments. This result is borne out by observations on the models described in Chapter 7.

5.1.4 The Sound Spectrograph

Spectral analysis of speech came of age, so to speak, with the development of the Sound Spectrograph (Koenig [1946]). This device provides a convenient means for permanently displaying the short-time spectrum of a sizeable duration of signal. Its method of analysis is precisely that shown in Fig. 5.4. Its choice of time windows (see preceding section) is made to highlight important acoustic and perceptual features such as formant structure, voicing, friction, stress and pitch. Many other devices for spectrum analysis have also been developed, but the relative convenience and ease of operation of the sound spectrograph has stimulated its wide acceptance in speech analysis and phonetic science. Because it is such a widely used tool, this section will give a brief description of the device and its principles of operation.

Fig. 5.7 shows a functional diagram of one type of sound spectrograph (commonly known as the Model *D* Sonagraph). With the microphone switch (SW 1) in the *record* position, a speech sample (generally about 2.5 sec in duration) is recorded on a magnetic disc. The microphone switch is turned to *analyze*, and a spectral analysis of the sample is made by playing it repeatedly through a bandpass filter. Upon successive playings the bandpass filter is, in effect, scanned slowly across the

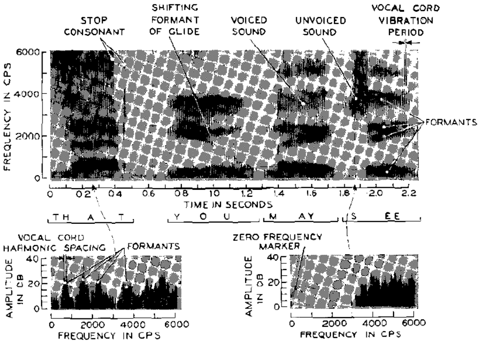


Figure 5.8: (a) Broadband sound spectrogram of the utterance "That you may see." (b) Amplitude vs frequency plots (amplitude sections) taken in the vowel portion of "that" and in the fricative portion of "see." (After (Barney and Dunn [1957]))

frequency band of the signal. The result is therefore equivalent to an analysis by many such filters. For practical reasons it is more convenient to use a fixed bandpass filter and to "slide" the spectrum of the signal past the filter. This is accomplished by modulating the signal onto a high frequency carrier and sliding one sideband of the signal past the fixed bandpass filter. The translation is accomplished by varying the frequency of the carrier. The carrier frequency control is mechanically geared to the magnetic disc so the signal spectrum is progressively analyzed upon repeated rotations of the disc.

With SW 2 in the *spectrogram* position, the output current of the bandpass filter is amplified and passed to a stylus whose vertical motion is geared to the magnetic disc and the carrier control (or to the effective frequency position of the bandpass filter). The stylus is in contact with an electrically sensitive facsimile paper which is fixed to a drum mounted on the same shaft as the magnetic disc. Electrical current from the stylus burns the paper in proportion to the current magnitude. The paper therefore acts as the full-wave rectifier of Fig. 5.4, and the finite size and spreading of the burned trace perform the low-pass filtering. The density of the burned mark is roughly proportional to the logarithm of the current magnitude. Because of the mechanical linkage, the stylus and carrier move slowly across the frequency range of the signal as the magnetic disc rotates, and a time-intensity-frequency plot of the signal is "painted" on the paper.

Two widths of the bandpass filter are conventionally used with the instrument, 300Hz and 45Hz. The time-frequency resolution of the analysis is essentially determined by these widths. As discussed in the preceding section, the wide pass-band provides better temporal resolution of speech events, while the narrow band yields a frequency resolution adequate to resolve harmonic lines in voiced utterances. A typical spectrogram made with the 300Hz wide analyzing filter is shown in the upper diagram of Fig. 5.8. As previously indicated, the abscissa is time, the ordinate is frequency, and darkness of the pattern represents intensity. Several speech features are indicated. Note that the time resolution is such that vertical striations in the voiced portions show the fundamental period of the vocal cords.

The facsimile paper is capable of depicting an intensity range (from lightest gray to darkest black) of only about 12dB (Prestigiacomo [1957]). It often is desirable to examine amplitude spectra over a greater intensity range. A means is therefore provided for making a frequency-versus-amplitude portrayal at any given instant along the time scale. For this operation, SW2 in Fig. 5.7 is put to the *section* position. A cam is placed on the drum periphery at the time of occurrence of the sound whose amplitude section is desired. The functions of the carrier and stylus are as previously described.

The sectioner contains a full-wave rectifier, an R-C integrator and a biased multivibrator. In one version of the apparatus, as the magnetic disc and drum rotate, the cam closes the section switch at the desired instant in the utterance, The value of the short-time spectrum at this instant is effectively "read" and stored on a capacitor in the input circuit of a biased multivibrator. The multi vibrator is held on (i.e., free runs) until the capacitor charge decays to a threshold value. The multivibrator then turns off. During its on-time, it delivers a marking current to the stylus and (because of the exponential decay) the length of the marked trace is proportional to the logarithm of the smoothed output of the analyzing filter. Because the stylus is scanning the frequency scale with the filter, an amplitude (*db*)-versus-frequency plot is painted for the prescribed instant.

Amplitude sections are usually made with the 45Hz (narrow band) filter. Typical sections taken in a vowel and in a fricative are shown in the lower half of Fig. 5.8.

Because the speech sample must be played repeatedly as the analyzing filter scans its band, the time to produce the complete spectrogram is appreciable. Common practice is to shorten the analyzing time by playing back at several times the recording speed. A typical value, for example, is a speed-up of three-to-one. A recorded bandwidth of 100 to 4000Hz is therefore multiplied to 300 to 12000Hz. If the analyzing bandpass filter is centered at, say, 15000Hz, then the carrier oscillator may scan from 15000 to 27000Hz. Depending upon frequency range and technique, one to several minutes may be required to analyze a 2.5 sec speech sample. In the course of the analysis the sample may be played back several hundred times. A common figure for the filter advance is of the order of 20Hz/playback.

The manner in which broadband spectrograms highlight vocal modes, or formants, for various articulatory configurations is illustrated in Fig. 5.9. Articulatory diagrams for four vowels, /i,ə,a,u/ and their corresponding broadband (300Hz) spectrograms are shown. The dark bands indicate the spectral energy concentrations and reflect the vocal modes for a given configuration. (These spectrograms can be compared with the calculated mode patterns for similar vowels in Figs. 4.25 and 4.26 of Chapter 4.)

Typical of the research uses to which this type of spectrographic display has been put is a large-scale study of vowel formant frequencies, amplitudes, and pitches for a number of different speakers ?etersonβarney. The results of this study for 33 men

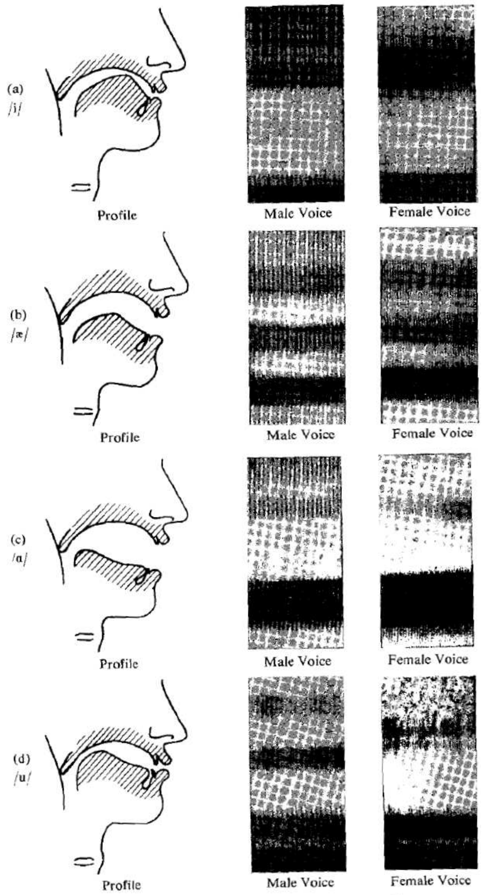


Figure 5.9: Articulatory diagrams and corresponding broad-band spectrograms for the vowels /i,ə,a,u/ as uttered by adult male and female speakers. (After (Potter *et al.* [1947]))

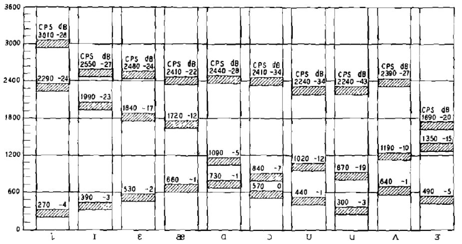


Figure 5.10: Mean formant frequencies and relative amplitudes for 33 men uttering the English vowels in an /h-d/ environment. Relative formant amplitudes are given in dB *re* the first formant of /ɔ/. (After (Peterson and Barney [1952]) as plotted by Haskins Laboratories)

give the mean formant frequencies for the English vowels as plotted in Fig. 5.10. The vowels were uttered in an /h--d/ environment.

Numerous "relatives" of the sound spectrograph–both predecessors and successors–have been designed and used, each usually with a specific purpose in mind. These devices range from scanned filter banks to correlation instruments. In a short space it is not possible to mention many of them. One variation in the spectrographic technique is the socalled "resonagraph" ﬁugginsɐʌ. This device is designed to delineate formant frequencies and to suppress nonformant energy. Another modification displays the time derivative of the spectral amplitude, rather than simply the amplitude ɱeyer-ɛpplerɐɨ, ʁockɱiller. The effect is to emphasize dynamic time changes in the spectrum and to suppress quasi-steady portions. Features such as stop consonants or formant transitions are therefore more sharply delineated.

An even closer relative is the so-called visible speech translator (Dudley and Jr. [1946], Riesz and Schott [1946]) in which the conventional sound spectrogram is painted electronically in real time, either on a moving belt coated with luminescent phosphor, or on a rotating cathode ray tube. A still different variation is the correlatograph (Bennett [1953], Biddulph [1954]) which plots the magnitude of the short-time autocorrelation function of the signal in trace density, the delay parameter on the ordinate, and time along the abscissa. Several schemes for quantizing the intensity dimension of the conventional spectrogram have also been described (Kersta [1948], Prestigiacomo [1957]). The result is to yield a "topological map" of the signal in which intensity gradients are indicated by the closeness of the contour lines.

5.1.5 Short-Time Correlation Functions and Power Spectra

If $x(t)$ is an on-going stationary random signal, its autocorrelation function $\phi(\tau)$ and its power density spectrum $\Phi(\omega)$ are linked by Fourier transforms (Wiener [1949], Lee

[1960], Papoulis [1965]).

$$\begin{aligned}\phi(\tau) &= \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} x(t)x(t+\tau)dt \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi(\omega)e^{j\omega\tau}d\omega\end{aligned}$$

and

$$\Phi(\omega) = \int_{-\infty}^{\infty} \phi(\tau)e^{-j\omega\tau}d\tau. \tag{5.10}$$

[Note that $\phi(0)$ is the mean square value, or average power, of the signal.]

For an aperiodic Fourier-transformable signal, $y(t)$, parallel relations link the autocorrelation function $\psi(\tau)$ and the energy density spectrum $\Psi(\omega)$

$$\begin{aligned}\psi(\tau) &= \int_{-\infty}^{\infty} y(t)y(t+\tau)dt \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \Psi(\omega)e^{j\omega t}d\omega\end{aligned}$$

$$\Psi(\omega) = \int_{-\infty}^{\infty} \psi(\tau)e^{-j\omega\tau}d\tau, \tag{5.11}$$

where

$$\Psi(\omega) = Y(\omega)Y^{*}(\omega), \quad \text{and} \quad Y(\omega) = \int_{-\infty}^{\infty} y(t)e^{-j\omega t}dt$$

[Note that

$$\psi(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Psi(\omega)d\omega$$

is the total energy of the signal.]

In both cases the correlation functions are real and even functions of the delay parameter τ, and the spectra are real and even functions of the frequency ω. All of the transforms can therefore be written as cosine transforms. These transform-pairs suggest the possibility of determining short-time spectral information by means of correlation techniques, provided the latter can be extended to the short-time case.

In the preceding discussion on short-time spectral analysis, the approach was to analyze a Fourier-transformable "piece" of the signal obtained by suitably weighting the past values. The correlation relations for aperiodic functions can be similarly extended to this description of the speech signal. According to the earlier derivations, at any instant t the following transforms are presumed to hold for the speech signal $f(t)$,

$$F(\omega,t) = \int_{-\infty}^{t} f(\lambda)h(t-\lambda)e^{-j\omega\lambda}d\lambda$$

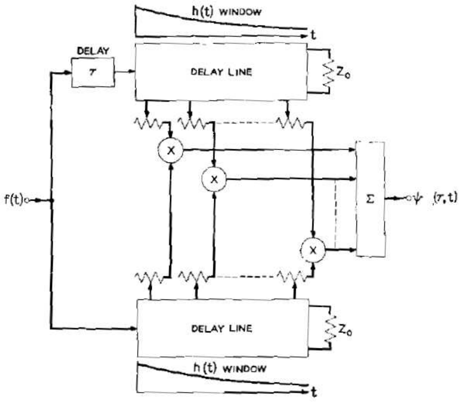


Figure 5.11: Method for the measurement of the short-time correlation function $\psi(\tau, t)$

$$[f(\lambda)h(t-\lambda)] = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega,t)e^{j\omega\lambda}d\omega, \tag{5.12}$$

where $h(t)$ is the weighting function. Then, formally,

$$\psi(\tau,t) = \int_{-\infty}^{t} f(\lambda)h(t-\lambda)f(\lambda+\tau)h(t-\lambda-\tau)d\lambda$$

$$\psi(\tau,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Psi(\omega,t)e^{j\omega\tau}d\omega \tag{5.13}$$

$$\Psi(\omega,t) = [F(\omega,t)F^*(\omega,t)] = \int_{-\infty}^{\infty} \psi(\tau,t)e^{-j\omega\tau}d\tau.$$

Practically, for real time measurement at time t, $f(t+\tau)$ for $\tau > 0$ is not known. [For a fixed over-all delay (comparable to the window duration) τ may be considered to be a differential delay.] However, $\psi(\tau,t)$ is formally an even function of τ. It can therefore be defined in terms of negative τ so that

$$\Psi(\omega,t) = \int_{-\infty}^{\infty} \psi(\tau,t)e^{-j\omega\tau}d\tau = 2\int_{-\infty}^{0} \psi(\tau,t)\cos\omega\tau d\tau \tag{5.14}$$

where $\Psi(\omega,t)$ is also an even function of ω.

Thus a short-time autocorrelation measure, related to the shorttime power spectrum $|F(\omega,t)|^2$ by the aperiodic transform, can be made. Techniques for the measurement of $|F(\omega,t)|^2$ have already been described in Section 5.1.2. Measurement of $\psi(\tau,t)$ for negative τ can be effected by the arrangement shown in Fig. 5.11. The individual output taps from the delay lines are weighted according to $h(t)$. Corresponding points (in the running variable λ) are multiplied, and the integration is

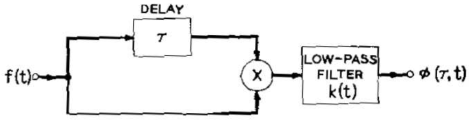


Figure 5.12: Circuit for measuring the running short-time correlation function $\phi(\tau, t)$

approximated as a finite sum[5]. $\psi(\tau, t)$ is therefore a running correlation which is related to $|F(\omega, t)|^2$ or $\Psi(\omega, t)$ by a Fourier transform.

It is also possible to define a short-time correlation function produced by weighting the product of the original signal and the signal delayed (FANO). The defining relation is

$$\phi(\tau, t) = \int_{-\infty}^{t} f(\lambda) f(\lambda + \tau) k(t - \lambda) d\lambda, \tag{5.15}$$

where $k(t) = 0$, $t < 0$ is the weighting function. The measure is easily implemented for $t \leq 0$ by the circuit shown in Fig. 5.12. This technique has been used experimentally to measure correlation functions for speech sounds (Stevens [1950], Kraft [1950], Biddulph [1954]).

In general, no simple transform relation exists between $\phi(\tau, t)$ and a measurable short-time power spectrum. Under the special condition $k(t) = 2\alpha e^{-2\alpha t} = [h(t)]^2$, however, $\phi(\tau, t)$ can be related to $\Psi(\omega, t) = |F(\omega, t)|^2$.

$$\begin{aligned} \psi(\tau, t) &= \int_{-\infty}^{t} f(\lambda) h(t - \lambda) f(\lambda + \tau) h(t - \lambda - \tau) d\lambda \\ &= e^{\alpha\tau} \int_{-\infty}^{t} 2\alpha f(\lambda) f(\lambda + \tau) e^{-2\alpha(t-\lambda)} d\lambda \qquad (5.16) \\ &= e^{\alpha\tau} \phi(\tau, t); \quad \tau \leq 0 \end{aligned}$$

But as previously argued, $\phi(\tau, t)$ is an even function of τ, and if $\phi(\tau, t)$ is defined as an even function, then $\psi(\tau, t) = e^{-\alpha|\tau|} \phi(\tau, t)$ for all τ, or

$$\begin{aligned} \phi(\tau, t) &= e^{\alpha|\tau|} \psi(\tau, t) \\ &= \frac{e^{\alpha|\tau|}}{2\pi} \int_{-\infty}^{\infty} \Psi(\omega, t) e^{j\omega\tau} d\omega, \end{aligned}$$

---

[5] The operations of Fig. 5.11 compute

$$\psi(\tau, t) = \int_{0}^{\infty} f(t - \lambda) h(\lambda) f(t - \lambda - \tau) h(\lambda + \tau) d\lambda,$$

for negative τ, instead of the form given in Eq. (5.13).

and

$$\begin{aligned}\Psi(\omega,t) &= \int_{-\infty}^{\infty} e^{-\alpha|\tau|}\phi(\tau,t)e^{-j\omega\tau}d\tau \\ &= \int_{-\infty}^{\infty} e^{-\alpha|\tau|}\phi(\tau,t)\cos\omega\tau d\tau.\end{aligned}$$

It also follows that

$$\Psi(\omega,t) = \frac{1}{2\pi}\left[\mathcal{F}\left\{e^{-\alpha|\tau|}\right\} * \mathcal{F}\left\{\phi(\tau,t)\right\}\right] \tag{5.17}$$

$$\begin{aligned}&= \frac{1}{2\pi}\left[\left(\frac{2\alpha}{\alpha^2+\omega^2}\right) * \Phi(\omega,t)\right] \\ &= \frac{1}{2\pi}\left[|H(\omega)|^2 * \Phi(\omega,t)\right],\end{aligned}$$

where $\mathcal{F}$ denotes the Fourier transform.

Thus the short-time power spectrum $\Psi(\omega,t)$ is the real convolution of the power spectrum $\Phi(\omega,t)$ with the low-pass energy spectrum $2\alpha/(\alpha^2+\omega^2)$. $\Psi(\omega,t)$ therefore has poorer spectral resolution than the Fourier transform of $\phi(\tau,t)$ [i.e., $\Phi(\omega,t)$]. Note also that for $h(t) = (2\alpha)^{\frac{1}{2}}e^{-\alpha t}$, $|F(\omega,t)|$ is essentially measured by single-resonant circuits with impulse response $[(2\alpha)^{\frac{1}{2}}e^{-\alpha t}\cos\omega t]$ and $[(2\alpha)^{\frac{1}{2}}e^{-\alpha t}\sin\omega t]$ (See Fig. 5.3.)

Weighting functions different from the exponential just discussed do not lead to simple transform relations between $\phi(\tau,t)$ and a power spectrum. Other definitions, however, can be made of measurable correlations and short-time power spectra, and these can be linked by specially defined transforms (Atal [1962]). For example, one can define a short-time spectrum

$$\Omega(\omega,t) = \int_{-\infty}^{\infty} \phi(\tau,t)m(|\tau|)\cos\omega\tau d\tau, \tag{5.18}$$

in which $\phi(\tau,t)$, as given in Eq. (5.15), is defined as an even function of τ (but is measured for delays only) so that,

$$\phi(\tau,t) = \int_{-\infty}^{t} f(\lambda)f(\lambda-|\tau|)n(t-\lambda)d\lambda,$$

where $m(t)$ and $n(t)$ are physically realizable weighting functions and are zero for $t < 0$[6]. $\Omega(\omega,t)$ and $\phi(\tau,t)$ are then linked by the definitions (5.18). $\phi(\tau,t)$ can be

---

[6] If $Q(\omega,t)$ is to be a positive quantity, some further restrictions must be placed on $n(t)$.

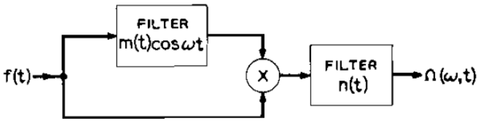


Figure 5.13: Arrangement for measuring the short-time spectrum $Q(\omega, t)$. (After (Atal [1962]))

measured according to Fig. 5.12, and a straightforward measure of $\Omega(\omega, t)$ can also be made. Substituting for $\phi(\tau, t)$ in the definition of $\Omega(\omega, t)$ gives

$$\begin{aligned}\Omega(\omega, t) &= 2\int_{-\infty}^{t} f(\lambda)n(t-\lambda)d\lambda \int_{0}^{\infty} f(\lambda-\tau)m(\tau)\cos\omega\tau d\tau \\ &= 2\left\{n(t) * f(t)\left[f(t) * m(t)\cos\omega t\right]\right\}.\end{aligned} \tag{5.19}$$

The operations indicated in (5.19) are a filtering of the signal $f(t)$ by a (normally bandpass) filter whose impulse response is $[m(t)\cos\omega t]$; a multiplication of this output by the original signal; and a (normally low pass) filtering by a filter whose impulse response is $n(t)$. The measurement is schematized in Fig. 5.13.

For the case $m(t) = n(t) = e^{-\alpha t}$, $\Omega(\omega, t)$ reduces to $\Psi(\omega, t)$. From the definition of $\Omega(\omega, t)$, the inverse relation follows

$$\phi(\tau, t) = \frac{1}{2\pi m(|\tau|)}\int_{-\infty|}^{\infty} \Omega(\omega, t)\cos\omega\tau d\omega. \tag{5.20}$$

The defining relations of Eq. (5.18) also imply that

$$\Omega(\omega, t) = M(\omega) * \Phi(\omega, t) \tag{5.21}$$

where

$$M(\omega) = \int_{-\infty}^{\infty} m(|\tau|)e^{-j\omega\tau}d\tau,$$

and

$$\Phi(\omega, t) = \int_{-\infty}^{\infty} \phi(tau, t)e^{-j\omega\tau}d\tau.$$

This result can be compared with Eq. (5.17), where

$$\begin{aligned}|H(\omega)|^2 &= \int_{-\infty}^{\infty} e^{-\alpha|\tau|}e^{-j\omega\tau}d\tau \\ H(\omega) &= \int_{0}^{\infty} (2\alpha)^{\frac{1}{2}}e^{-\alpha\tau}e^{-j\omega\tau}d\tau = \int_{0}^{\infty} h(\tau)e^{-j\omega\tau}d\tau\end{aligned}$$

Since $\Omega(\omega, t)$ is obtained from $\Phi(\omega, t)$ by convolution with the (low pass) spectrum $M(\omega)$, it has poorer spectral definition than $\Phi(\omega, t)$.

5.1.6 Average Power Spectra

The spectral measuring schemes of the previous discussion use windows which are relatively short in duration to weight past values of the signal. They yield spectra in which brief temporal fluctuations are preserved. A long-term mean value of the spectrum, say $|F(\bar{\omega}, t)|^2$, might also be of interest if average spectral distribution is of more importance than short-time variations. Such an average can be written as

$$\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} F(\omega, t) F^*(\omega, t) dt = |F(\bar{\omega}, t)|^2 = \Psi(\bar{\omega}, t) \tag{5.22}$$

$$= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} dt \int_{-\infty}^{t} f(\lambda) h(t-\lambda) e^{-j\omega\lambda} d\lambda \int_{-\infty}^{t} f(\eta) h(t-\eta) e^{j\omega\eta} d\eta.$$

Changing variables and rearranging

$$|F(\bar{\omega}, t)|^2 = \int_{0}^{\infty} d\lambda h(\lambda) e^{j\omega\lambda} \int_{0}^{\infty} d\eta h(\eta) e^{-j\omega\eta} \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f(t-\lambda) f(t-\eta) dt. \tag{5.23}$$

According to Eqs. (5.10), the latter integral is simply $\phi(\lambda - \eta)$, which is the inverse Fourier transform of $\Phi(\omega)$. That is,

$$\begin{aligned} \phi(\lambda - \eta) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi(\delta) e^{j\delta(\lambda-\eta)} d\delta \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi(\delta) e^{-j\delta(\lambda-\eta)} d\delta, \end{aligned}$$

because $\Phi(\omega)$ is real and even. Then

$$\begin{aligned} |F(\bar{\omega}, t)|^2 &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi(\delta) d\delta \int_{0}^{\infty} h(\lambda) e^{j\lambda(\omega-\delta)} d\lambda \int_{0}^{\infty} h(\eta) e^{-j\eta(\omega-\delta)} d\eta \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi(\delta) H(\omega-\delta) H^*(\omega-\delta) d\delta \end{aligned}$$

$$|F(\bar{\omega}, t)|^2 = \frac{1}{2\pi} \left[\Phi(\omega) * |H(\omega)|^2 \right]. \tag{5.24}$$

Therefore, the long-time average value of the power spectrum $|F(\bar{\omega}, t)|^2$ is the real convolution of the power density spectrum $\Phi(\omega)$ and the energy density spectrum of the time window $h(t)$. The narrower the $|H(\omega)|^2$ spectrum, the more nearly $|F(\bar{\omega}, t)|^2$ represents the power density spectrum $\Phi(\omega)$. A narrow $H(\omega)$ corresponds to a long time window and to narrow bandpass filters in the circuits of Fig. 5.3 and 5.4. In the limit $H(\omega)$ is an impulse at $\omega = 0$, the time window is a unit step function and $|F(\bar{\omega}, t)|^2$ has the same spectral characteristics as $\Phi(\omega)$. For any value of ω, $|F(\bar{\omega}, t)|^2$ is the integral of the power density spectrum "seen" through the aperture $|H(\omega)|^2$

positioned at ω. It is therefore the average power of the signal in the pass band of the filter in Fig. 5.4.

It was previously demonstrated [Eq. (5.17)] that for the special condition $h(t) = [(2\alpha)^{\frac{1}{2}} e^{\alpha t}]$,

$$\Psi(\omega, t) = \frac{1}{2\pi}\left[|H(\omega)|^2 * \Phi(\omega, t)\right].$$

Notice that for this situation, the long-time average is

$$\Psi(\bar{\omega}, t) = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T}\int_{-\infty}^{\infty} e^{-\alpha|\tau|}\phi(\tau, t)\cos\omega\tau d\tau dt \tag{5.25}$$

$$= \int_{-\infty}^{\infty} e^{-\alpha|\tau|}\phi(\bar{\tau}, t)\cos\omega\tau d\tau.$$

Substituting for $\phi(\tau, t)$ from (5.15) and interchanging variables leads to

$$\Psi(\bar{\omega}, t) = \int_{0}^{\infty} e^{-\alpha|\tau|}\phi(\tau)\cos\omega\tau d\tau \int_{0}^{\infty} k(\beta)d\beta. \tag{5.26}$$

Since

$$\int_0^\infty k(t)dt = \int_0^\infty h^2(t)dt = 1$$

then

$$\Psi(\bar{\omega}, t) = \frac{1}{2\pi}\left[|H(\omega)|^2 * \Phi(\omega)\right],$$

which corresponds to the result (5.24).

5.1.7 Measurement of Average Power Spectra for Speech

A number of experimental measurements of the average power spectrum of speech have been made (for example, (Sivian [1929], Dunn and White [1940]). The technique frequently used is essentially the bandpass filter arrangement shown previously in Fig. 5.4, with the exception that a square-law rectifier and a long-time integrator (averager) are used. This arrangement is shown is Fig. 5.14. If the switch closes at time $t = 0$ and remains closed for T sec, the accumulated capacitor voltage is an approximation to $|F(\bar{\omega}, t)|^2$ and is,

$$V_c(T) = \int_0^T a'^2(\omega, \lambda)\frac{1}{RC}e^{-\frac{1}{RC}(T-\lambda)}d\lambda \tag{5.27}$$

If $RC \gg T$, then the exponential is essentially unity for $0 \le \lambda \le T$, and

$$V_c(T) \approx \frac{1}{RC}\int_0^T a'^2(\omega, \lambda)d\lambda \tag{5.28}$$

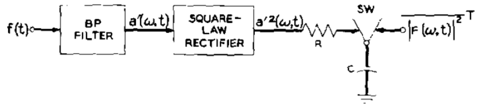


Figure 5.14: Circuit for measuring the long-time average power spectrum of a signal

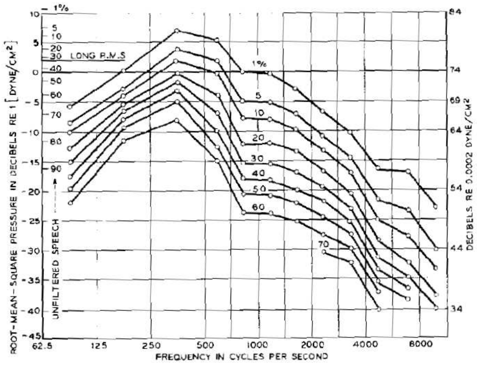


Figure 5.15: Root mean square sound pressures for speech measured in -ll sec intervals 30 cm trom the mouth. The analyzing filter bands are one-half octave wide below 500Hz and one octave wide above 500 Hz. (After (Dunn and White [1940])) The parameter is the percentage of the intervals having levels greater than the ordinate

$$\sim |F(\bar{\omega}, t)|^2$$

The measurement described by (5.28) has been used in one investigation of speech spectra. Bandpass filters with bandwidths one-half octave wide below 500Hz and one octave wide above 500Hz were used. The integration time was $\frac{1}{8}$ sec (Dunn and White [1940]). Distributions of the absolute root-mean-square speech pressure in these bands–measured 30 cm from the mouth of a talker producing continuous conversational speech–are shown in Fig. 5.15. The data are averages for six men. The distribution for the unfiltered speech is shown by the marks on the left ordinate.

If the integration time is made very long, say for more than a minute of continuous speech (all natural pauses between syllables and sentences being included), or if many short-time measurements are averaged, one obtains a long-time power spectrum in which syllabic length variations are completely smoothed out. Assuming that the speech power is uniformly distributed in the octave and half-octave filter bands the measured longtime power density spectrum, $\Phi(\omega)$, for speech is shown in Fig. 5.16. The ordinate here is given in terms of mean-square sound pressure per cycle. In both Fig. 5.15 and 5.16, the detailed formant structure of individual sound is averaged out.

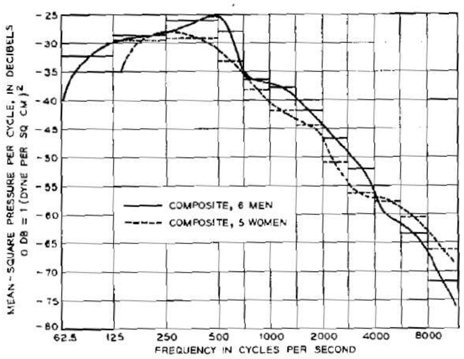


Figure 5.16: Long-time power density spectrum for continuous speech measured 30 cm from the mouth. (After (Dunn and White [1940]))

5.2 Predictive Coding of Speech

For many classes of information signals, including speech, the value of the signal at a given instant is correlated with its values at other instants, and hence represents redundant information. One theory of data compression in digital systems is therefore based upon forming an error signal, e_i, between the samples of an input sequence, s_i, and linear estimates of those samples, $\hat{s}_i$,

$$e_i = (\hat{s}_i - s_i)\,.$$

Generally, the estimate $\hat{s}_i$ of sample s_i is formed as a weighted linear combination of samples from some portion of the input sample sequence, using a linear prediction filter of the form

$$\hat{s}_i = \sum_{k=1}^{p} a_k s_{i-k} \tag{5.29}$$

The estimate $\hat{s}_i$ is called the linear prediction of s_i, and the coefficients a_k are called the linear prediction coefficients (LPC). The coefficients are computed from statistics of the sample sequence in a manner which is optimum in some sense. If the input sample sequence is not stationary, the weighting coefficients must be updated periodically.

In order to transmit a block of M samples to the receiver, it is necessary that the error samples and the weighting coefficients be transmitted to the receiver. Suppose the desired accuracy of the input sample sequence requires "r" bits per sample. By straightforward quantization, it would take $(M \cdot r)$ bits to transmit the block of M samples. However, if the sample sequence is processed through a data compression system, the number of bits needed to transmit the block is hopefully less. Usually the error signal is transmitted at the same rate as the input sample sequence, but the weighting coefficients are transmitted typically at a rate $1/M$ times the input sequence. Suppose the error signal is quantized to q bits and the N weighting coefficients are coded to w bits per coefficient. The number of bits needed the specify the M samples to the receiver is then $(Mq + Nw)$. In order to obtain a saving, it is required that

$$Mq + Nw < Mr$$

or

$$q + \frac{N}{M}w < r.$$

If the sample sequence is highly correlated, the power in the error signal will be significantly less than the power in the input sample sequence. Hence, fewer bits will be required to describe the error samples than the input samples. If $M \gg N$, then the term $\frac{N}{M}w$ becomes negligible and the objective can be achieved.

One such method of data compression is linear prediction (Elias [1955]). Linear prediction has been found to provide significant improvements in picture transmission, speech transmission, and the transmission of telemetry data. A linear predictor forms its estimates of the input samples from past samples in the input sequence. Another method of data compression is linear interpolation. An interpolator forms its estimates of the input samples from both past and future samples in the input sequence.

Linear interpolation has the potential for reducing the power in the error signal beyond that for an equal-order prediction. However, interpolation requires more computation and complex implementation. Also, it looses some of its advantages when the error signal is quantized inside a feedback loop (Haskew [1969]). The present discussion will therefore focus on prediction.

A linear Nth-order predictor estimates the magnitude of the present input sample, s_i, by a linear combination, $\hat{s}_i$, of N weighted past samples.

$$\hat{s}_i = \sum_{j=1}^{N} a_j s_{i-j}, \tag{5.30}$$

where a_j is the weighting coefficient applied to the past sample s_{i-j}.

When the statistics of the input signal are nonstationary (changing as a function of time), the weighting coefficients must be updated periodically. Only the weighting coefficients computed for intervals near the present sample yield accurate estimates of the sample magnitude. In this case, weighting coefficients are updated, for example, every M input samples, where M is usually much larger than the order of the predictor, N. The output of the predictor, the error e_i, is formed by subtracting the estimated value of the present sample from the actual value of the present sample.

$$e_i = s_i - \sum_{j=1}^{N} a_j s_{i-j}. \tag{5.31}$$

The input signal is now described by the output of the predictor (the error signal) and the weighting coefficients. In z-transform notation

$$e(z) = [1 - P(z)]S(z),$$

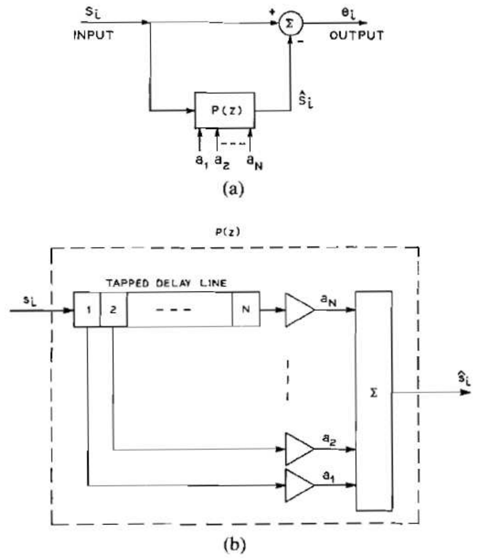


Figure 5.17: Block diagram of linear prediction

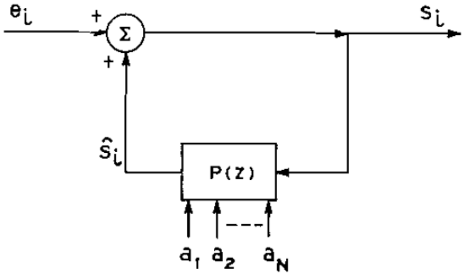


Figure 5.18: Linear prediction receiver

where

$$P(z) = \sum_{j=1}^{N} a_j z^{-j} \tag{5.32}$$

These relations are shown schematically in Fig. 5.17. Recovery of original input signal is obtained from the inverse relation

$$s(z) = e(z)[1 - P(z)]^{-1}, \tag{5.33}$$

and is given by the operations of Fig. 5.18. Typically, however, the transmitted signals, i.e., the e_i and a_i, are quantized, and the receiver has access only to corrupted versions of them.

The criterion by which the a_i are typically determined is a minimization of the power of the error signal (that is, minimization of the square difference between $\hat{s}_i$

and s_i). For M samples the error power is

$$\epsilon^2 = \frac{1}{M}\sum_{j=1}^{M} e_j^2 = \frac{1}{M}\sum_{j=1}^{M}(s_j - \hat{s}_j)^2. \tag{5.34}$$

Substitution for the estimate $\hat{s}_i$ gives

$$\epsilon^2 = \frac{1}{M}\sum_{j=1}^{M}\left[s_j - \sum_{k=1}^{N} a_k s_{j-k}\right]$$

or

$$\epsilon^2 = \frac{1}{M}\sum_{j=1}^{M} s_j^2 - \frac{2}{M}\sum_{j=1}^{M}\sum_{k=1}^{N} a_k s_{j-k} + \frac{1}{M}\sum_{j=1}^{M}\left[\sum_{k=1}^{N} a_k s_{j-k}\right]\left[\sum_{l=1}^{N} a_l s_{j-l}\right] \tag{5.35}$$

Interchanging summations and rearranging terms,

$$\epsilon^2 = \frac{1}{M}\sum_{j=1}^{M} s_j^2 - 2\sum_{k=1}^{N} a_k\left[\frac{1}{M}\sum_{j=1}^{M} s_j s_{j-k}\right] + \sum_{k=1}^{N}\sum_{l=1}^{N} a_k a_l\left[\sum_{j=1}^{M} s_{j-k} s_{j-l}\right] \tag{5.36}$$

Define the signal power σ^2 and its covariance function r_{kl} as

$$\sigma^2 = \frac{1}{M}\sum_{j=1}^{M} s_j^2,$$

and

$$r_{kl} = \frac{1}{M\sigma^2}\sum_{j=1}^{M} s_{j-l} s_{j-k} \tag{5.37}$$

The error power then becomes

$$\epsilon^2 = \sigma^2\left[1 - 2\sum_{k=1}^{N} a_k r_{0k} + \sum_{k=1}^{N}\sum_{l=1}^{N} a_k a_l r_{kl}\right] \tag{5.38}$$

This result can be simplified by matrix notation. Define the column matrix containing the weighting coefficients as

$$A = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} \tag{5.39}$$

Define the column matrix containing the elements r_{0k} as

$$G = \begin{bmatrix} r_{01} \\ r_{02} \\ \vdots \\ r_{0N} \end{bmatrix} \tag{5.40}$$

Define the $(N \times N)$ matrix containing the elements r_{kl} as

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1N} \\ r_{21} & r_{22} & \cdots & r_{2N} \\ \vdots & \vdots & & \vdots \\ r_{N1} & r_{N2} & \cdots & r_{NN} \end{bmatrix} \tag{5.41}$$

Note from the equation for r_{kl} that

$$r_{kl} = r_{lk};$$

hence, R is a symmetric matrix. The error power can then be written as

$$\epsilon^2 = \sigma^2 \left[1 - 2A^T G + A^T R A\right]. \tag{5.42}$$

To optimize the predictor, the column matrix, A, must be selected such that ϵ^2 is a minimum. This is accomplished by taking the derivative of ϵ^2 with respect to A and equating the result to zero.

$$\left.\frac{\partial \epsilon^2}{\partial A}\right|_{A=A_{opt}} = 0,$$

$$\frac{\partial \epsilon^2}{\partial A} = 2G + 2RA = 0.$$

Solving the latter equation for A yields

$$A = R^{-1} G. \tag{5.43}$$

The minimum mean-square value of the error signal for the interval of M samples, ϵ^2, is found by substituting the optimum, A, given by Eq. (5.43) in Eq. (5.42) for ϵ^2 and simplifying. The result is

$$\left\{\epsilon^2\right\}_{min} = \sigma^2 \left(1 - G^T R^{-1} G\right), \tag{5.44}$$

where σ^2 is the mean-square value of the input sequence over the interval of M samples.

The normal equations for autocorrelation can be solved by inverting a Toeplitz matrix:

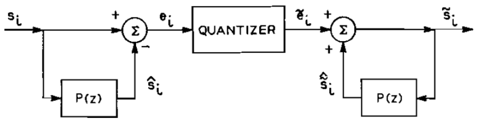


Figure 5.19: Open-loop quantization of a predictor error signal

$$\bar{r} = \bar{\bar{R}}\bar{a} \quad \Rightarrow \quad \bar{a} = \bar{\bar{R}}^{-1}\bar{r} \tag{5.45}$$

$$\bar{r}_p = \begin{bmatrix} R(1) \\ R(2) \\ \dots \\ R(p) \end{bmatrix}, \quad \bar{\bar{R}} = \begin{bmatrix} R(0) & R(1) & \dots & R(p-1) \\ R(1) & R(0) & \dots & R(p-2) \\ \dots & & & \\ R(p-1) & R(p-2) & \dots & R(0) \end{bmatrix}, \quad \bar{a} = \begin{bmatrix} a_1 \\ \dots \\ a_p \end{bmatrix} \tag{5.46}$$

This inversion can be done efficiently using the following recursive algorithm, called the "Levinson-Durbin recursion:"

$$E^{(0)} = R(0) \tag{5.47}$$

$$k_i = \frac{R(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j)}{E^{(i-1)}}, \quad 1 \le i \le p \tag{5.48}$$

$$a_i^{(i)} = k_i \tag{5.49}$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}, \quad 1 \le j \le i-1 \tag{5.50}$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)} \tag{5.51}$$

$$\tag{5.52}$$

For practical digital transmission the error samples and the predictor coefficients are quantized to the fewest possible levels. The receiver of the prediction system uses these data to reconstruct estimates of the sample sequence of the original signal. If care is not exercised, quantizing noise may accumulate in the sample sequence.

This difficulty can be simply illustrated. Consider the "open-loop" quantization of the error signal shown in Fig. 5.19. Let tildas represent quantized versions of the signals. The quantizing noise present in the reconstructed received signal is therefore

$$(s_i - \tilde{s}_i) = (e_i - \tilde{e}_i) + (\hat{s}_i - \hat{\tilde{s}}_i),$$

where

$$\hat{\tilde{s}}_i = \sum_{j=1}^{N} a_j s_{i-j}. \tag{5.53}$$

The quantizing noise in the received signal is not merely the same as the quantizing noise of the error signal, but also includes the quantizing error in the estimate. Since $\hat{\tilde{s}}_i$ is formed from a sum over N past samples the quantizing noise may accumulate.

5.2.1 Choosing the LPC Order

The LPC order needs to be large enough to represent each formant using a complex-conjugate pole pair. There also need to be an extra 2 or 3 poles to represent spectral tilt. With everything together, we have:

$$p \approx 2 \times (\text{Number of Formants}) + (2 \text{ to } 3) \tag{5.54}$$

The number of formants in the spectrum is the Nyquist rate ($F_s/2$), divided by the average spacing between neighboring formants:

$$\text{Number of Formants} = \frac{F_s/2}{\text{average}(F_{n+1} - F_n)} \tag{5.55}$$

The spacing between neighboring formant frequencies is approximately

$$\text{average}(F_{n+1} - F_n) \approx \frac{c}{2l} \tag{5.56}$$

where $c = 35400$cm/s is the speed of sound, and l is the length of the vocal tract. The length of a male vocal tract is close to 17.7cm, so there is approximately one formant per 1000Hz:

$$p \approx \left(\frac{F_s}{1000\text{Hz}}\right) + (2 \text{ to } 3) \tag{5.57}$$

The length of a female vocal tract is close to 14.75cm, so there is approximately one formant per 1200Hz:

$$p \approx \left(\frac{F_s}{1200\text{Hz}}\right) + (2 \text{ to } 3) \tag{5.58}$$

5.2.2 Choosing the LPC Gain

The LPC excitation is defined by the formula:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{5.59}$$

The LPC error is defined by the formula:

$$e(n) \equiv s(n) - \sum_{k=1}^{p} a_k s(n-k) = Gu(n) \tag{5.60}$$

If we define

$$\sum_{k=0}^{N-1} u^2(n) \equiv 1 \tag{5.61}$$

then

$$G^2 = \frac{\sum e^2(n)}{\sum u^2(n)} = E_{min} \tag{5.62}$$

5.2.3 Frequency-Domain Interpretation of LPC

The LPC error is

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{5.63}$$

so the error spectrum is

$$E(z) = S(z)(1 - \sum_{k=1}^{p} a_k z^{-k}) = S(z)A(z) \tag{5.64}$$

Using Parseval's theorem, we get that

$$E_n = \sum_m e^2(m) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |E_n(e^{j\omega})|^2 \, d\omega \tag{5.65}$$

Substituting in the form of $E(z)$, we get

$$E_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} |X_n(e^{j\omega})|^2 |A(e^{j\omega})|^2 \, d\omega = \frac{G^2}{2\pi}\int_{-\pi}^{\pi} \frac{|X_n(e^{j\omega})|^2}{|H(e^{j\omega})|^2} \, d\omega \tag{5.66}$$

Since X is in the numerator inside the integral, any algorithm which minimizes E_n will automatically try to produce an $H(e^{j\omega})$ which does a good job of modeling X at frequencies where X is large. In other words, LPC models spectral peaks better than spectral valleys.

5.2.4 Lattice Filtering

Lattice filtering analyzes or synthesizes speech using ith-order forward prediction error $e^{(i)}(m)$ and backward prediction error $b^{(i)}(m)$:

$$e^{(i)}(m) = s(m) - \sum_{k=1}^{i} a_k s(m-k) = e^{(i-1)}(m) - k_i b^{(i-1)}(m-1) \tag{5.67}$$

$$b^{(i)}(m) = s(m-i) - \sum_{k=1}^{i} a_k s(m+k-i) = b^{(i-1)}(m-1) - k_i e^{(i-1)}(m) \tag{5.68}$$

In LPC synthesis, $s(n)$ and $b^{(p)}(m)$ are calculated recursively from $e^{(p)}(m)$, as shown in figure 5.2.4.

5.2.5 How to Calculate Reflection Coefficients

1. Reflection coefficients can be estimated directly from the speech signal by minimizing an error metric of the form

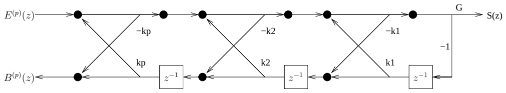


Figure 5.20: LPC synthesis using a lattice filter structure.

$$\tilde{E}^{(i)} = \quad \text{some function of} \;\; e^{(i)}(m), \;\; b^{(i)}(m) \tag{5.69}$$

Direct calculation of reflection coefficients is computationally expensive, but sometimes leads to better estimates of the transfer function (e.g. Burg's method).

2. Any LPC synthesis filter can be implemented as a lattice filter. The relationship between a_i and k_i is given by the Levinson-Durbin recursion.

Why Use Lattice Filter Instead of Direct-Form LPC?

Tests for stability of $1/A(z)$:

- Roots of $A(z)$: $|r_i| < 1$.
- Reflection coefficients: $|k_i| < 1$.
- Direct form coefficients a_k: No simple test.

Equivalence of Lattice and Concatenated-Tube Models

Reflection coefficients in the lattice filter are equivalent to reflection coefficients in a lossless tube model of the vocal tract. There are many ways to make the two structures correspond. One convenient formulation numbers the tube areas A_i backward from the lips:

$$k_i = \frac{A_i - A_{i+1}}{A_i + A_{i+1}}, \qquad A_i = \text{Area of } i\text{th tube section} \tag{5.70}$$

$$A_0 = \infty \qquad \text{(Area of the space beyond the lips — lossless termination)} \tag{5.71}$$

$$A_{p+1} = A_p \qquad \text{(Area of the glottis — lossy termination)} \tag{5.72}$$

The length l of each tube section is determined by the sampling period T and the speed of sound c:

$$T = \frac{2l}{c} \tag{5.73}$$

5.2.6 LPC Distance Measures

Suppose we want to calculate the distance between two all-pole spectra,

$$S_1(\omega) = |X_1(\omega)|^2 \approx \left|\frac{G_1}{A_1(\omega)}\right|^2, \quad S_2(\omega) = |X_2(\omega)|^2 \approx \left|\frac{G_2}{A_2(\omega)}\right|^2 \tag{5.74}$$

We can:

- Calculate the spectral L_2 norm by integrating $\log|G_1/A_1(\omega)|^2 - \log|G_2/A_2(\omega)|^2$ over ω.
- Find the LPC cepstrum, and calculate a weighted cepstral distance.
- Calculate an LPC likelihood distortion.

Itakura-Saito Distortion

Suppose that $S_1(\omega)$ is a random spectrum, produced by filtering a unit-energy noise process $U(\omega)$ through an unknown all-pole filter $A_1(\omega)$:

$$\begin{aligned} S_1(\omega) &= \left|\frac{G_1}{A_1(\omega)}\right|^2 |U(\omega)|^2 && (5.75)\\ E[S_1(\omega)] &= \left|\frac{G_1}{A_1(\omega)}\right|^2 && (5.76)\end{aligned}$$

Suppose we don't know A_1, but we have a spectrum A_2 which might or might not be a good approximation to A_1. One question worth asking is, what is the probability that the signal $x_1(n)$ was generated using filter G_2/A_2? This probability is related to a distance called the Itakura-Saito measure of the distortion between spectra G_1/A_1 and G_2/A_2 (Itakura and Saito [1968, 1970]):

$$\begin{aligned} d_{IS}\left(\frac{G_1^2}{|A_1|^2}, \frac{G_2^2}{|A_2|^2}\right) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} |X_1(\omega)|^2 \frac{|A_2(\omega)|^2}{G_2^2} d\omega - \log\frac{G_1^2}{G_2^2} - 1 && (5.77)\\ &\sim -\log\left(p_{\mathbf{x}_1}\left([x_1(0)\ldots x_1(L-1)] \mid G_2, A_2(\omega)\right)\right) && (5.78)\end{aligned}$$

The first term in the Itakura-Saito distortion is the residual energy of the random signal $x_1(n)$, filtered through the inverse filter $A_2(z)/G_2$:

$$x_1(n) \rightarrow \boxed{A_2(z)/G_2} \rightarrow e_{12}(n) \tag{5.79}$$

$$\begin{aligned}
\frac{1}{2\pi}\int_{-\pi}^{\pi}|E_{12}(\omega)|^2 d\omega &= \frac{1}{2\pi}\int_{-\pi}^{\pi}|X_1(\omega)|^2\frac{|A_2(\omega)|^2}{G_2^2}d\omega && (5.80)\\
&= \frac{1}{G_2^2}\sum_n (x_1(n) - \sum_{k=1}^{p} a_{k,2}x_1(n-k))^2 && (5.81)\\
&= \frac{1}{G_2^2}\left(R_1(0) - 2\sum_{k=1}^{p} a_{k,2}R_1(k) + \sum_{i=1}^{p}\sum_{k=1}^{p} a_{i,2}a_{k,2}R_1(|i-k|)\right) && (5.82)\\
&= \frac{\mathbf{a}_2'\mathbf{R}_{p,1}\mathbf{a}_2}{G_2^2} && (5.83)
\end{aligned}$$

where $\mathbf{a}_2$ is the LPC coefficient vector representing the polynomial $A_2(z)$, and $\mathbf{R}_{p,1}$ is the autocorrelation matrix built out of samples of signal $x_1(n)$. Using this notation, the Itakura-Saito distortion can be written in the easy-to-compute form:

$$\boxed{d_{IS}\left(\frac{G_1^2}{|A_1|^2}, \frac{G_2^2}{|A_2|^2}\right) = \frac{\mathbf{a}_2'\mathbf{R}_{p,1}\mathbf{a}_2}{G_2^2} - \log\frac{G_1^2}{G_2^2} - 1 \quad (5.85)} \qquad (5.85)$$

Characteristics:

- The Itakura-Saito distortion measure is asymmetric:

$$d_{IS}\left(\frac{G_1^2}{|A_1|^2}, \frac{G_2^2}{|A_2|^2}\right) \neq d_{IS}\left(\frac{G_2^2}{|A_2|^2}, \frac{G_1^2}{|A_1|^2}\right) \qquad (5.86)$$

- The Itakura-Saito distance is non-negative ($d_{IS} \geq 0$) and reflexive ($d_{IS}\left(\frac{G_1^2}{|A_1|^2}, \frac{G_2^2}{|A_2|^2}\right) = 0$ if and only if $G_1/A_1 = G_2/A_2$).

Likelihood-Ratio Distortion, Itakura Distortion

Many times, we don't care about differences in the spectral energy of S_1 and S_2. The *likelihood-ratio* distortion measure is obtained by normalizing both S_1 and S_2 to unit energy, and then calculating an Itakura-Saito distortion:

$$\begin{aligned}
d_{LR}\left(\frac{1}{|A_1|^2}, \frac{1}{|A_2|^2}\right) &= d_{IS}\left(\frac{1}{|A_2|^2}, \frac{1}{|A_1|^2}\right) && (5.87)\\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi}|A_2(\omega)|^2\frac{S_1(\omega)}{G_1^2}d\omega - 1 && (5.88)\\
&= \frac{\mathbf{a}_2'\mathbf{R}_{p,1}\mathbf{a}_2}{G_1^2} - 1 && (5.89)\\
& && (5.90)
\end{aligned}$$

A similar distortion measure called the Itakura distortion is often used instead of the likelihood-ratio distortion:

$$\begin{aligned}
d_I\left(\frac{1}{|A_1|^2}, \frac{1}{|A_2|^2}\right) &\equiv \log\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} |A_2(\omega)|^2 \frac{S_1(\omega)}{G_1^2} d\omega\right) && (5.91)\\
&= \log\left(\frac{\mathbf{a}_2'\mathbf{R}_{p,1}\mathbf{a}_2}{G_1^2}\right) && (5.92)
\end{aligned}$$

$x(n)$ and $\hat{x}(n)$:

$$\left.\begin{array}{l} X(z) = R(z)T(z)P(z) \\ \hat{X}(z) = \hat{R}(z) + \hat{T}(z) + \hat{P}(z) \end{array}\right\} \hat{X}(z) = \log(X(z)) \qquad (5.93)$$

5.3 Homomorphic Analysis

In a further approach toward exploiting the source-system distinction in the speech signal, a processing technique called homomorphic filtering has been applied to vocoder design (Oppenheim [1969], Oppenheim *et al.* [1968], Oppenheim [1971]). The approach is based on the observation that the mouth output pressure is approximately the linear convolution of the vocal excitation signal and the impulse response of the vocal tract. Homomorphic filtering is applied to deconvolve the components and provide for their individual processing and description.

Homomorphic filtering is a generic term applying to a class of systems in which a signal-complex is transformed into a form where the principles of a linear filtering may be applied (Oppenheim [1969]). In the case of a speech signal, whose spectrum is approximately the product of the excitation spectrum and the vocal-tract transmission, a logarithm operation produces an additive combination of the source and system components.

Linear systems are homomorphic to addition:

$$L[x_1(n) + x_2(n)] = L[x_1(n)] + L[x_2(n)] \qquad (5.94)$$

Linear filtering is useful for analyzing a signal with two additive components, e.g. $y(n) = x(n) + \epsilon(n)$. In speech, we are often more interested in "convolutional components." For example, the speech signal can be modeled as the convolution of a source function $p(n)$, a transfer function $t(n)$, and a radiation function $r(n)$:

$$x(n) = r(n) * (t(n) * p(n)) \qquad (5.95)$$

In order to analyze $x(n)$, we want a nonlinear "filtering" system which is "homomorphic to convolution," that is,

$$H[t(n) * p(n)] = H[t(n)] * H[p(n)] \qquad (5.96)$$

The system $H[\bullet]$ can be written as the series connection of a transformation $D[\bullet]$, a linear system $L[\bullet]$, and the inverse transformation $D^{-1}[\bullet]$:

$$H[t(n) * p(n)] = D^{-1}\left[L\left[D[t(n) * p(n)]\right]\right] \tag{5.97}$$

where $D[\bullet]$ is the transformation which converts convolution into addition:

$$D[t(n) * p(n)] = D[t(n)] + D[p(n)] \tag{5.98}$$

$D[x(n)]$ can be written as $D[x(n)] = \hat{x}(n)$, where $\hat{x}(n)$ is defined to be the *complex cepstrum* of $x(n)$. The form of the complex cepstrum is obvious if one considers the z transforms of $x(n)$ and $\hat{x}(n)$:

$$\left.\begin{array}{l} X(z) = R(z)T(z)P(z) \\ \hat{X}(z) = \hat{R}(z) + \hat{T}(z) + \hat{P}(z) \end{array}\right\} \hat{X}(z) = \log(X(z)) \tag{5.99}$$

5.3.1 Complex Cepstrum

The "complex cepstrum" of $x[n]$ is a real-valued sequence, $\hat{x}[n]$, containing sufficient information to reconstruct the complex Fourier transform $X(e^{j\omega})$. Specifically,

$$\boxed{\hat{x}(n) = \frac{1}{2\pi}\int_0^{2\pi} \log(X(e^{j\omega}))e^{j\omega n}d\omega} \tag{5.100}$$

$$\hat{X}(e^{j\omega}) = \log(X(e^{j\omega})) = \log|X(e^{j\omega})| + j\widehat{\arg}(X(e^{j\omega})) \tag{5.101}$$

The complex cepstrum $\hat{x}(n)$ is only defined if $\log(X(z))$ is a valid Z transform, uniformly defined on the unit circle. $\hat{x}[n]$ is real-valued if and only if the time-domain sequence $x[n]$ is also real-valued: if $x(n)$ is real, then $\log|X(e^{j\omega})|$ is even, and $\widehat{\arg}(X(e^{j\omega}))$ is odd, and therefore $\hat{x}(n)$ is real.

The function $\widehat{\arg}(X(e^{j\omega})$ is the "unwrapped phase" of X. Recall that the principal argument, $\arg(X(e^{j\omega}))$, is only defined over the range of $(-\pi, \pi]$. Such a constraint is not appropriate for the definition of cepstrum, because we require that the sum of two cepstra should still be a valid cepstrum:

$$\widehat{\arg}(X(e^{j\omega})) = \widehat{\arg}(R(e^{j\omega})) + \widehat{\arg}(T(e^{j\omega})) + \widehat{\arg}(P(e^{j\omega})) \tag{5.102}$$

This requirement can be met by adding integer multiples of 2π to the principal argument, as necessary, in order to produce a continuous, odd function of ω; this process is known as "unwrapping" the phase (the argument is only odd if $x(n)$ is real).

The argument of the cepstrum, n, is sometimes called the "quefrency," especially in echo analysis applications (Bogert *et al.* [1962]). In speech analysis, n is often called the cepstral "lag."

5.3.2 Cepstrum

The cepstrum (sometimes called the "magnitude cepstrum") of $x[n]$ is a sequence, $c[n]$, with sufficient information to reconstruct the magnitude but not the phase of the Fourier transform:

$$\boxed{c(n) = \tfrac{1}{2\pi}\int_0^{2\pi} \log|X(e^{j\omega})|e^{j\omega n}d\omega} \tag{5.103}$$

By algebraic manipulation of Eq. 5.103, it is possible to show that the magnitude cepstrum is the even part of the complex cepstrum:

$$\frac{\hat{x}(n)+\hat{x}(-n)}{2} = \frac{1}{2\pi}\int_0^{2\pi} \log|X(e^{j\omega})|e^{j\omega n}d\omega = c(n) \tag{5.104}$$

Example 5.3.1 Relationship Between Magnitude Cepstrum and Complex Cepstrum

$$x(n) = \delta(n) - \alpha\delta(n-N), \qquad |\alpha| < 1 \tag{5.105}$$

$$X(z) = 1 - \alpha z^{-N} \tag{5.106}$$

$$\hat{X}(z) = \log(1-\alpha z^{-N}) = -\sum_{r=1}^{N} \frac{\alpha^r z^{-rN}}{r} \quad \text{if } |\alpha z^{-N}| < 1 \tag{5.107}$$

$$\boxed{\hat{x}(n) = -\textstyle\sum_{r=1}^{\infty} \frac{\alpha^r}{r}\delta(n-rN)} \tag{5.108}$$

$$\boxed{c(n) = (1/2)(\hat{x}(n)+\hat{x}(-n)) = -\textstyle\sum_{r=1}^{\infty} \frac{\alpha^r}{2r}(\delta(n-rN)+\delta(n+rN))} \tag{5.109}$$

5.3.3 Signals with Rational Spectrum

Consider the class of "rational signals," that is, signals whose Z-transform can be written as:

$$X(z) = G\frac{\prod_{k=1}^{N_a}(1-a_k z^{-1})\prod_{k=1}^{N_b}(1-b_k z)}{\prod_{k=1}^{N_c}(1-c_k z^{-1})\prod_{k=1}^{N_d}(1-d_k z)}, \qquad |a_k|, |b_k|, |c_k|, |d_k| < 1 \tag{5.110}$$

For this class of signals, all stable minimum phase signals (all signals with $N_b = 0, N_d = 0$) are also causal, and all stable maximum phase signals ($N_a = 0, N_c = 0$) are also anti-causal. The linear predictive filter ($N_a = N_b = N_d = 0$), for example, is a minimum-phase rational signal.

For signals in this class, $\hat{x}(n)$ is an infinite length signal, even if $x(n)$ is finite in length (the only exception is $x(n) = G\delta(n)$):

$$\hat{x}(n) = \begin{cases} -\sum_{k=1}^{N_a} \frac{a_k^n}{n} + \sum_{k=1}^{N_c} \frac{c_k^n}{n} & n > 0 \\ \log(G) & n = 0 \\ \sum_{k=1}^{N_b} \frac{b_k^{-n}}{n} - \sum_{k=1}^{N_d} \frac{d_k^{-n}}{n} & n < 0 \end{cases} \tag{5.111}$$

Although it is infinite in length, $\hat{x}(n)$ is always largest for small values of $|n|$; as $|n|$ increases, $\hat{x}[n]$ decays exponentially fast. Minimum-phase, causal sequences have causal $\hat{x}(n)$. Maximum-phase, anti-causal sequences have anti-causal $\hat{x}(n)$.

Although $\hat{x}[n]$ is infinite in length, only $N_a + N_b + N_c + N_d + 1$ of its samples are independently specified. All of the other samples can be computed recursively, as follows:

$$\hat{X}(z) = \log(X(z)) \tag{5.112}$$

$$\frac{d}{dz}\hat{X}(z) = \frac{1}{X(z)}\frac{d}{dz}X(z) \tag{5.113}$$

$$\left[-z\frac{d}{dz}\hat{X}(z)\right] X(z) = \left[-z\frac{d}{dz}X(z)\right] \tag{5.114}$$

$$n\hat{x}(n) * x(n) = nx(n) \tag{5.115}$$

$$\sum_{k=-\infty}^{\infty} k\hat{x}(k)x(n-k) = nx(n) \tag{5.116}$$

For $n \neq 0$, this yields

$$\boxed{\sum_{k=-\infty}^{\infty} \frac{k}{n}\hat{x}(k)x(n-k) = x(n)} \tag{5.117}$$

If $x(n)$ is minimum-phase and causal, the summation in the above equation is only non-zero for $0 \leq k \leq n$, yielding the following recursion for $\hat{x}(n)$:

$$\hat{x}(n) = \frac{x(n)}{x(0)} - \sum_{k=0}^{n-1} \frac{kx(n-k)}{nx(0)}\hat{x}(k), \quad n > 0 \tag{5.118}$$

If $x(n)$ is maximum-phase and anti-causal, the summation is only non-zero for $n \leq k \leq 0$, yielding the following formula for $\hat{x}(n)$:

$$\hat{x}(n) = \frac{x(n)}{x(0)} - \sum_{k=n+1}^{0} \frac{kx(n-k)}{nx(0)}\hat{x}(k), \quad n < 0 \tag{5.119}$$

The low-quefrency part of the cepstrum of a speech signal can be estimated from its LPC coefficients:

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{5.120}$$

The LPC cepstrum $\hat{h}(m)$ is the inverse transform of $\log H(z)$:

$$\begin{aligned}
\hat{h}(m) &= \mathcal{Z}^{-1}\left(\log H(z)\right) && (5.121)\\
&= \mathcal{Z}^{-1}\left(\log G - \log A(z)\right) && (5.122)\\
&= \log G\delta(n) - \hat{a}(m) && (5.123)\\
&= \begin{cases} 0 & n<0 \\ \log(G) & n=0 \\ \hat{a}(m) & n>0 \end{cases} && (5.124)
\end{aligned}$$

Since $A(z)$ is minimum-phase, $\hat{a}(m)$ is causal, and therefore $\hat{h}(m)$ is also a causal sequence. The form of $\hat{a}(m)$ can be computed as a special case of Eq. 5.118, yielding:

$$\hat{h}(n) = \begin{cases} 0 & n<0 \\ \log(G) & n=0 \\ a_n + \sum_{k=1}^{n-1} \frac{k}{n} a_{n-k}\hat{h}(k) & p \geq n > 0 \end{cases} \tag{5.125}$$

Notice that the first $p+1$ cepstral coefficients ($0 \leq n \leq p$) contain a complete description of the transfer function; $\hat{h}(n)$ for larger n can be computed recursively from the first $p+1$ values of $\hat{h}(n)$.

5.3.4 Liftering

Remember that windowing in time equals convolution in frequency. Suppose that $w[m]$ is a windowing sequence with spectrum $W(\omega)$; then

$$\mathcal{F}\left\{c[m]w[m]\right\} = \log S(\omega) * W(\omega) \tag{5.126}$$

The L_2 distance between two log-power spectra $S_1(\omega)$ and $S_2(\omega)$ is defined to be

$$(d_2)^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|\log S_1(\omega) - \log S_2(\omega)\right|^2 d\omega \tag{5.127}$$

The distance between two spectra, as shown in Eq. 5.127, is rarely useful. The fine structure of the spectrum may convey information about voice quality or pitch, but most linguistic information is conveyed by the vocal tract transfer function. A pretty good estimate of the vocal tract transfer function can be computed by smoothing the power spectra, thus

$$(\hat{d}_2)^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|W(\omega) * \log S_1(\omega) - W(\omega) * \log S_2(\omega)\right|^2 d\omega \tag{5.128}$$

Using Parseval's theorem, the smoothed spectral distance $\hat{d}_2$ may be efficiently computed as

$$\hat{d}_2^2 = \sum_{m=-\infty}^{\infty} w[m]^2 (c_1[m] - c_2[m])^2 \tag{5.129}$$

The magnitude cepstrum is a real-valued, even sequence, $c_1[m] = c_1[-m]$, therefore it is not necessary to include both positive and negative values of m in the summation of Eq. 5.129. Suppose we define the even part of $w(m)$ to be $\tilde{w}(m)$:

$$\tilde{W}(\omega) = \Re\left\{W(\omega)\right\}, \qquad \tilde{w}(m) = \begin{cases} w(m)/2 & m > 0 \\ w(-m)/2 & m < 0 \end{cases} \tag{5.130}$$

If $w(m)$ is delayed causal, we can take advantage of the even symmetry of $c_1(m)$ to express $\hat{d}_2^2$ as a two-sided sum:

$$\begin{aligned} \hat{d}_2^2 &= 2\sum_{m=-L}^{L} \tilde{w}^2(m)(c_1(m) - c_2(m))^2 && (5.131) \\ &= \frac{1}{\pi}\int_{-\pi}^{\pi} |(\log S_1(\omega) * \tilde{W}(\omega)) - (\log S_2(\omega) * \tilde{W}(\omega))|^2 d\omega && (5.132) \\ & && (5.133) \end{aligned}$$

So a weighted cepstral distance is similar to the following L_2 norm:

- Smooth $\log S_1(\omega)$ and $\log S_2(\omega)$ using the smoothing spectrum $\tilde{W}(\omega) = \Re\left\{W(\omega)\right\}$.
- Calculate the L_2 distortion measure between the two smoothed log spectra.

Example 5.3.2 Liftering

If $w(m)$ is a causal rectangular window covering samples 1 through L, then $\tilde{w}(m)$ is an even window of length $2L + 1$:

$$w(m) = \begin{cases} 1 & m = 1, \ldots, L \\ 0 & \text{else} \end{cases}, \qquad \tilde{w}(m) = \begin{cases} 1/2 & m = -L, \ldots, -1, 1, \ldots, L \\ 0 & \text{else} \end{cases} \tag{5.134}$$

$\tilde{w}(m)$ is just a rectangular window of length $2L + 1$, minus the impulse $\delta(n)$. The spectrum is therefore:

$$\begin{aligned} \tilde{W}(\omega) &= \frac{\sin\frac{\omega(2L+1)}{2}}{2\sin\frac{\omega}{2}} - \frac{1}{2} \approx \frac{\sin\frac{\omega(2L+1)}{2}}{2\sin\frac{\omega}{2}} && (5.135) \\ & && (5.136) \end{aligned}$$

where the approximation holds for large L.

5.4 Spectral and Cepstral Derivatives

Remember that the log-power-STFT is a function of both time and frequency. Its inverse transform, the power cepstrum, is therefore a function of both signal time and cepstral lag:

$$c_t(m) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log S_t(\omega) e^{j\omega m} d\omega, \qquad S_t(\omega) \equiv |X_t(\omega)|^2 \tag{5.137}$$

Suppose we are interested in the time-derivative of the log power spectrum. This can be computed by taking the time-derivative of the cepstrum:

$$\frac{\partial c_t(m)}{\partial t} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\partial \log S_t(\omega)}{\partial t} e^{j\omega m} d\omega \tag{5.138}$$

Suppose that the rate of change of the log power spectrum is governed by a Gaussian distribution:

$$\frac{\partial \log S_t(\omega)}{\partial t} \sim \mathcal{N}(S, \mu_S, U_S) \tag{5.139}$$

then the cepstral derivative is a weighted sum of Gaussians, and is therefore itself a Gaussian random variable:

$$\frac{\partial c_t(m)}{\partial t} \sim \mathcal{N}(o, \mu, U) \tag{5.140}$$

5.4.1 Derivative Estimators

One of the most straightforward ways to model the rate of spectral change is using the first difference between successive cepstral frames:

$$\frac{\partial c_t[m]}{\partial t} \approx c_t(m) - c_{t-\delta}(m) \tag{5.141}$$

where the offset $\delta \approx$ 10ms is adjusted to model short-term spectral changes. It is possible to use different windowing functions at each delay:

$$o_t = [\ldots,\, c_t(m)w_1(m), \ldots,\, \left(\frac{c_{t+\delta}(m) - c_{t-\delta}(m)}{2}\right) w_2(m), \ldots,\, \left(\frac{c_{t+\Delta}(m) - c_{t-\Delta}(m)}{2}\right) w_3(m), \ldots] \tag{5.142}$$

In particular, the spectral energy, $c_t(0)$, is a function of the recording level, so that taken by itself, it tells you nothing at all about the phoneme. However, the rate of change of spectral energy can tell you a great deal about the phoneme, so most analysis systems include $c_{t+\delta}(0) - c_{t-\delta}(0)$ in the observation vector, even if they don't include any other dynamic information.

Likewise, $c_t(1)$ contains information about the spectral slope, which depends on such extraneous factors as the type of microphone and the speaker identity. The rate of change of the spectral slope, however, contains a great deal of phonetic information, so it is useful to include $c_{t+\delta}(1) - c_{t-\delta}(1)$ in your observation vector.

The simple first cepstral difference discussed above is a noisy estimate of the cepstral derivative. Instead of using a short-term and long-term cepstral difference, it is possible to compute parametric estimates of the first and second cepstral derivatives. The first and second cepstral derivatives are calculated by fitting a quadratic curve to the cepstral trajectory, in order to minimize the error

$$E = \sum_{t=-M}^{M} [c_t(m) - (h_1(m) + h_2(m)t - h_3(m)t^2)]^2 \tag{5.143}$$

where the window size M is comparable to the long-term cepstral difference window Δ.

The book gives formulas for the optimum h_1, h_2, and h_3 in terms of the cepstral coefficients. Once these coefficients have been computed, the cepstral derivative estimates are

$$\frac{\partial c_t(m)}{\partial t} \approx h_2(m), \qquad \frac{\partial^2 c_t(m)}{\partial t^2} \approx 2h_3(m) \tag{5.144}$$

5.4.2 Modulation Filtering

A spectral derivative estimate can be viewed as a high-pass filter of the log-power spectrum or cepstrum. For example, a short-term cepstral difference can be written as

$$d_t(m) = c_{t+1}(m) - c_{t-1}(m), \qquad D_z(m) = H(z)C_z(m), \quad H(z) = z(1 - z^{-2}) \tag{5.145}$$

- One of the biggest advantages of high-pass filtering $D_z(m)$ is that it removes the relatively constant effects of microphone and room tone. Thus, for example, $d_t(0)$ and $d_t(1)$ are more useful than $c_t(0)$ and $c_t(1)$, because constant offsets in the spectral energy and spectral slope caused by variations in recording conditions have been factored out.

- The biggest disadvantage of the high-pass filter is that it emphasizes rapid spectral changes which may not be very important perceptually. In fact, psychophysical research suggests that humans are most sensitive to spectral changes which occur at a rate of about 4-6 cycles per second (about the rate of syllable production in normal speech), and that sensitivity to spectral change drops off at higher frequencies. The high-pass filter in equation 5.145 rises at 6dB per octave all the way up to half the frame rate, e.g. if the frame length is 10ms, $H(z)$ gives the most emphasis to changes at a rate of 50 cycles per second. This does not reflect human hearing very well, which is part of the reason why the observation vector o_t must contain samples of $c_t(m)$ as well as samples of $d_t(m)$.

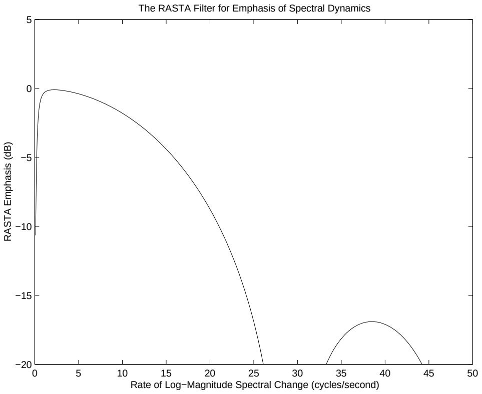


Figure 5.21: In the RASTA method, frame-to-frame variations in a spectral estimate are smoothed using a filter like the one shown here.

The RASTA (RelAtive SpecTrAl) method replaces the high-pass filter in equation 5.145 with a band-pass filter (Hermansky and Morgan [1994]). The band-pass filter has the following characteristics:

- A very sharp zero at zero frequency, to remove the effect of recording conditions.
- A relatively flat pass-band from 2 to 6 Hertz, which allows RASTA-filtered coefficients $r_t(m)$ to be used in place of the original coefficients $c_t(m)$.
- A slow roll-off above about 6 Hz, which de-emphasizes rapid spectral changes which are mostly inaudible to human listeners.

The original RASTA filter is as follows, but any filter with the characteristics above could be used just as well:

$$H(z) = \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{10z^{-2}(1 - 0.98z^{-1})} \tag{5.146}$$

The RASTA technique fulfills one of the original purposes of the delta-cepstrum (removing the influence of recording conditions), but the other condition is not fulfilled. Since $r_t(m)$ has a flat pass-band, it tends to model relatively steady-state spectra, not spectral change; for example, the difference in spectral rate of change between a /b/ and a /w/ is not captured by $r_t(m)$! Therefore, it seems reasonable to use an observation vector which contains RASTA-cepstra and delta-RASTA cepstra, e.g.

$$o_t = [\ldots,\ r_t(m), \ldots,\ \frac{r_{t+\delta}(m) - r_{t-\delta}(m)}{2}, \ldots] \tag{5.147}$$

Unfortunately, the RASTA technique is relatively new, and it is not yet clear whether including delta-RASTA in the observation vector reduces recognition error or not.

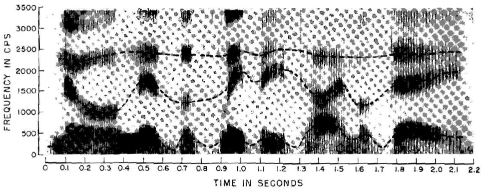


Figure 5.22: Sound spectrogram showing idealized tracks for the first three speech formants

5.5 Formant Analysis of Speech

Formant analysis of speech can be considered a special case of spectral analysis. The objective is to determine the complex natural frequencies of the vocal mechanism as they change temporally. The changes are, of course, conditioned by the articulatory deformations of the vocal tract. One approach to such analysis is to consider how the modes are exhibited in the short-time spectrum of the signal. As an initial illustration, the temporal courses of the first three speech formants are traced in an idealized form on the spectrogram of Fig. 5.22. Often, for bandwidth compression application, an automatic, real-time determination of these data is desired.

As certain of the results in Chapter 4 imply, the damping or dissipation characteristics of the vocal system are relatively constant and predictable, especially over the frequency range of a given formant. Generally, therefore, more interest attaches to the temporal variations of the imaginary parts of the complex formant frequencies than to the real parts. Nevertheless, an adequate knowledge of the real parts, or of the formant bandwidths, is important both perceptually and in spectral analysis procedures.

The "system function" approach to speech analysis, as discussed in Chapter 4, aims at a specification of the signal in terms of a transmission function and an excitation function. If the vocal configuration is known, the mode pattern can be computed, and the output response to a given excitation can be obtained. In automatic analysis for encoding and transmission purposes, the reverse situation generally exists. One has available only the acoustic signal and desires to analyze it in terms of the properties of the source and the modes of the system. One main difficulty is in not knowing how to separate uniquely the source and the system.

The normal modes of the vocal system move continuously with time, but they may not, for example, always be clearly manifest in a shorttime spectrum of the signal. A particular pole may be momentarily obscured or suppressed by a source zero or by a system zero arising from a side-branch element (such as the nasal cavity). The short-time spectrum generally exhibits the prominent modes, but it is often difficult to say with assurance where the low-amplitude poles or significant polezero pairs might lie.

Further complicating the situation is the fact that the output speech signal is generally not a minimum-phase function (that is, it may not have all its zeros in the left half of the complex frequency plane). If it were, its phase spectrum would be implied by its amplitude spectrum. The vocal-tract transmission is, of course, minimum phase for all conditions where radiation takes place from only one point, i.e., mouth or nostril. For simultaneous radiation from these points it is not. It can be shown that the glottal source, provided the volume velocity wave is zero at some time during its period, possesses only finite-frequency zeros and no poles (Mathews and Walker [1962]). Further, it can be shown that the zeros can lie in either the right or left half planes, or in both (Dunn *et al.* [1962]). These factors conspire to make accurate automatic formant analysis a difficult problem. The present section outlines a number techniques for the automatic measurement of formant frequency and formant bandwidth, and indicates the performance they achieve.

5.5.1 Formant-Frequency Extraction

In its simplest visualization, the voiced excitation of a vocal resonance is analogous to the excitation of a single-tuned circuit by brief, periodic pulses. The output is a damped sinusoid repeated at the pulse rate. The envelope of the amplitude spectrum has a maximum at a frequency equal essentially to the imaginary part of the complex pole frequency. The formant frequency might be measured either by measuring the axis-crossing rate of the time waveform, or by measuring the frequency of the peak in the spectral envelope. If the bandwidth of the resonance is relatively small, the first moment of the amplitude spectrum,

$$\bar{f} = \frac{\int f A(f) df}{\int A(f) df}$$

might also be a reasonable estimate of the imaginary part of the pole frequency.

The resonances of the vocal tract are, of course, multiple. The output time waveform is therefore a superposition of damped sinusoids and the amplitude spectrum generally exhibits multiple peaks. If the individual resonances can be suitably isolated, say by appropriate filtering, the axis-crossing measures, the spectral maxima and the moments might all be useful indications of formant frequency. If, on the other hand, the more subtle properties of the source and the system are to he accounted for–say the spectral zeros produced by the glottal source or by a sidebranch resonator–a more sophisticated measure of the normal modes generally is necessary. One such approach is the detailed fitting of an hypothesized spectral model to the real speech spectrum. For analyses of this type, it is often advantageous to employ the storage and rapid logical operations of a digital computer.

Axis-Crossing Measures of Formant Frequency

One of the earliest attempts at automatic tracking of formant frequencies was an average zero-crossing count (Peterson [1951]). The idea was to take the average density

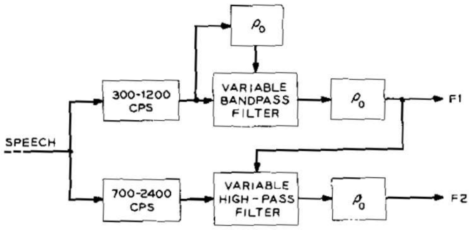


Figure 5.23: Automatic formant measurement by zero-crossing count and adjustable prefiltering. (After (Chang [1956]))

of zero-crossings of the speech wave and of its time derivative as approximations to the first and second formants, respectively. The reasoning was that in the unfiltered, voiced speech the first formant is the most prominent spectral component. It consequently is expected to have the strongest influence upon the axis-crossing rate. In the differentiated signal, on the other hand, the first formant is de-emphasized and the second formant is dominant. The results of these measures, however, were found to be poor, and the conclusion was, that the method did not give acceptable precision.

A number of refinements of the zero-crossing technique have been made. In one (Munson and Montgomery [1950], Davis *et al.* [1952]), the speech signal is pre-filtered into frequency ranges appropriate to individual formants. The axis-crossing rate and the amplitude are measured for the signal in each of the bands. A remaining disadvantage, however, is that the method is still subject to the overlapping of the formant frequency ranges.

A more elaborate implementation of the same basic idea, but with a feature designed to minimize deleterious overlap, has also been made (Chang [1956]). The notion is to employ an iterative measure of the average rate of zero-crossing in a given frequency range and to successively narrow the frequency range on the basis of the measured rate. The expectation is for rapid convergence. Fig. 5.23 illustrates the method. The signal is pre-filtered by fixed filters into ranges roughly appropriate to the first two formants. An axis-crossing measure, ρ_0, of the lower band is made and its value is used to tune automatically a narrower, variable band-pass filter. The axis-crossing output of this filter is, in turn, taken as an indication of the first formant frequency (F1). Its value is used to adjust the cut-off frequency of a variable HP filter. The average axis-crossing output of the latter is taken as an estimate of the second formant frequency (F2).

If the spectral distribution of the signal is continuous, as in the case of unvoiced sounds, the average axis-crossing rate for a given spectral element is approximately twice the first moment of the spectral piece (Chang [1956]). However, other more direct methods for measuring spectral moments have been considered.

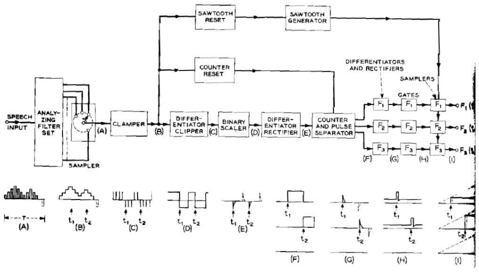


Figure 5.24: Spectrum scanning method for automatic extraction of formant frequencies (After (Flanagan [1956a]))

Spectral Moments

The n-th moment of an amplitude spectrum $A(\omega)$ is $M_n = \int \omega^n A(\omega) d\omega$, where ω is the radian frequency. If a suitable pre-filtering or partitioning of the spectrum can be made, then a formant frequency can be approximated by

$$\bar{\omega} = \frac{M_1}{M_0} \approx \frac{\sum_i \omega_i A(\omega_i)}{\sum_i A(\omega_i)}$$

A number of formant measures based upon this principle have been examined (Potter and Steinberg [1950], Gabor [1952], Atal and Schroeder [1956], Campanella *et al.* [1962]). The spectral partitioning problem remains of considerable importance in the accuracy of these methods. However, certain moment ratios have been found useful in separating the frequency ranges occupied by formants (Suzuki *et al.* [1963]). Another difficulty in moment techniques is the asymmetry or skewness of the spectral resonances. The measured formant frequency may be weighted toward the "heavier" side of the spectrum, rather than placed at the spectral peak.

Spectrum Scanning and Peak-Picking Methods

Another approach to real-time automatic formant tracking is simply the detection and measurement of prominences in the short-time amplitude spectrum. At least two methods of this type have been designed and implemented (Flanagan [1956a]). One is based upon locating points of zero slope in the spectral envelope, and the other is the detection of local spectral maxima by magnitude comparison. In the first–illustrated in Fig. 5.24-a short-time ampliutde spectrum is first produced by a set of bandpass filters, rectifiers, and integrators. The analysis is precisely as described earlier in Section 5.1.2. The outputs of the filter channels are scanned rapidly (on the order of 100 times per second) by a sample-and-hold circuit. This produces a time function which is a step-wise representation of the short-time spectrum at a number (36 in this instance) of frequency values. For each scan, the time function is differentiated and binary-scaled to produce pulses marking the maxima of the spectrum. The marking pulses are directed into separate channels by a counter where they sample a sweep

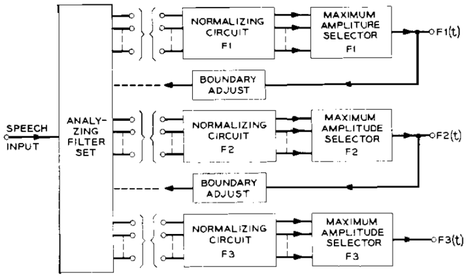


Figure 5.25: Peak-picking method for automatic tracking of speech formants. (After FLANAGAN, 1956a)

voltage produced at the scanning rate. The sampled voltages are proportional to the frequencies of the respective spectral maxima and are held during the remainder of the scan. The resulting stepwise voltages are subsequently smoothed by low-pass filtering.

The second method segments the short-time spectrum into frequency ranges that ideally contain a single formant. The frequency of the spectral maximum within each segment is then measured. The operation is illustrated in Fig. 5.25. In the simplest form the segment boundaries are fixed. However, additional control circuitry can automatically adjust the boundaries so that the frequency range of a given segment is contingent upon the frequency of the next lower formant. The normalizing circuit “clamps” the spectral segment either in terms of its peak value or its mean value. This common-mode rejection enables the following peak-selecting circuitry to operate over a wide range of amplitudes. The maxima of each segment are selected at a rapid rate–for example, 100 times per second–and a voltage proportional to the frequency of the selected channel is delivered to the output. The selections can be time-phased so that the boundary adjustments of the spectral segments are made sequentially and are set according to the measured position of the next lower formant. A number of improvements on the basic method have been made by providing frequency interpolation (Shearme [1959]), more sophisticated logic for adjusting the segment boundaries (Holes and Kelly [1960]), and greater dynamic range for the peak selectors (Stead and Jones [1961]). The objective in all these designs has been the realization of a real-time, practicable hardware device for direct application in a transmission system.

A typical output from the device of Fig. 5.25, using fixed boundaries, is shown in Fig. 5.26. It is clear that the operation is far from perfect. In this example a large third formant error occurs in the /r/ of “rain.” Automatic control of the F2-F3 boundary, however, eliminates this error. As a rough indication of the performance, one evaluation shows that its output follows F1 of vowels within ±150Hz greater than 93% of the time, and F2 within ±200Hz greater than 91% of the time (Flanagan [1956a]). Although one desires greater precision, this method–because of its simplicity and facility for real-time analysis–has proved useful in several investigations of complete formant-vocoder systems (Flanagan and House [1956], Stead and

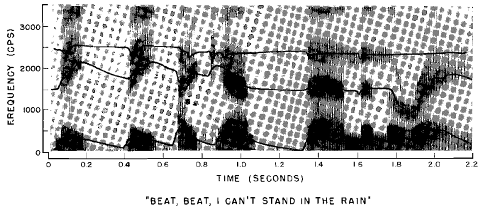


Figure 5.26: Formant outputs from the tracking device shown in Fig. 5.25. In this instance the boundaries of the spectral segments are fixed

Jones [1961], Shearme *et al.* [1962]).

Digital Computer Methods for Formant Extraction

The development of digital computers has enabled application of more sophisticated strategies to speech processing. The more esoteric processings are made possible by the ability of the computer to store and rapidly manipulate large quantities of numerical data. A given data sample can be held in the machine while complex tests and measures are applied to analyze a particular feature and make a decision. This advantage extends not only to formant tracking, but to all phases of speech processing. The relations between sampled-data systems and continuous systems (see, for example, (Ragazzini and Franklin [1958])) permit simulation of complete transmission systems within the digital computer. This is a topic in itself, and we will return to it in a later chapter.

The digital analyses which have been made for speech formants have been primarily in terms of operations on the spectrum. The spectrum either is sampled and read into the computer from an external filter bank, or is computed from a sampled and quantized version of the speech waveform. One approach along the latter line has been a pitch-synchronous analysis of voiced sounds (Mathews and Walker [1962]). Individual pitch periods are determined by visual inspection of the speech oscillogram. The computer then calculates the Fourier series for each pitch period as though that period were one of an exactly periodic signal. The envelope of the calculated spectrum is then fitted by a synthetic spectrum in successive approximations and according to a weighted least-square error criterion. A pole-zero model for the vocal tract and the glottal source, based upon acoustic relations for the vocal tract (see Chapter 4), produces the synthetic spectrum.

The fitting procedure is initiated by guessing a set of poles and zeros appropriate to the calculated real spectrum. The computer then successively increments the frequency and damping of each individual pole and zero to minimize the weighted mean-square error (in log-amplitude measure). After about 10 to 20 complete cycles, a close fit to the speech spectrum can be obtained. Typical rms log-amplitude errors range from about 1.5 to 2.5 db. A typical result of the fitting procedure is shown in Fig. 5.27. The measured formant frequencies and bandwidths are then taken as the

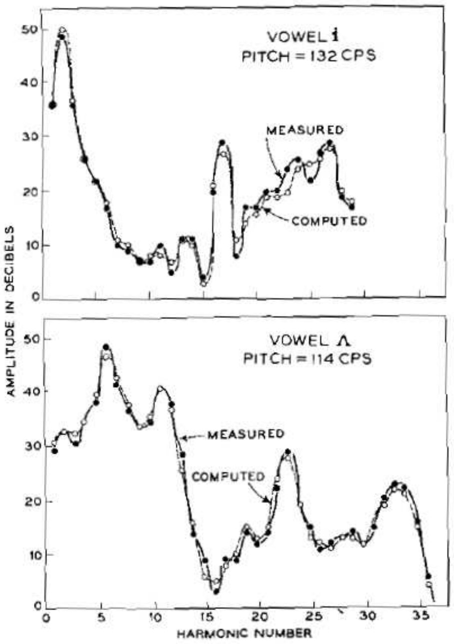


Figure 5.27: Spectral fit computed for one pitch period of a voiced sound. (After (Mathews and Walker [1962]))

frequencies and bandwidths of the best fitting spectral model.

A computer system for non-pitch-synchronous formant analysis, in which spectral data are produced external to the computer, can also be summarized (Hughes [1958, 1961]). A bank of 35 contiguous bandpass filters with rectifiers and integrators produces a short-time spectrum of the running speech. The filter outputs are scanned at a rapid rate (180 sec^{-1}) to produce a framed time function which represents successive spectral sections (essentially the same as that shown in Fig. 5.5). This time function is sampled every 154μsec and quantized to 11 bits by an analog-to-digital converter. A certain amount of the data is then held in the computer storage for processing.

One analysis procedure for which the computer is programmed (1) locates the fricative sounds in a word and classifies them; (2) locates the first and second formants in voiced segments; and (3) calculates the overall sound level. The formant tracking procedure is basically a peak-picking scheme similar to that shown previously in Fig. 5.25. However, a number of detailed, programmed constraints are included to exploit vocal tract characteristics and limitations. In principle, the procedure for a given spectral scan is as follows. Find the peak filter in the frequency range appropriate to the first formant. Store the frequency and amplitude values of this channel. On the basis of the F1 location, adjust the frequency range for locating F2. Locate the peak filter in the adjusted F2 range and store its frequency and amplitude values. Finally, examine the next spectral scan and find F1 and F2, subject to continuity constraints with previously determined values. Large, abrupt changes in F1 and F2 of

"Hawaii" (a)

"Yowie" (b)

Figure 5.28: Tracks for the first and second formant frequencies obtained from a computer-analysis of real-time spectra. The speech samples are (a) "Hawaii" and (b) "Yowie" uttered by a man. (After (Hughes [1958]))

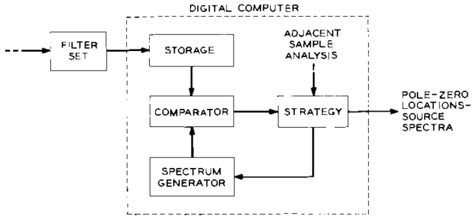


Figure 5.29: Computer procedure for formant location by the" analysis-by-synthesis" method. (After (Bell *et al.* [1961]))

small time duration are ignored. Typical results, described as "good" and "average" from this procedure are shown in 5.28.

A real-time spectral input to a computer has also been applied in a spectral-fitting technique for formant location (Bell *et al.* [1961]). The procedure–termed "analysis-by-synthesis" by its originators–is illustrated in Fig. 5.29. As before, a filter bank produces a short-time spectrum which is read into the digital computer via an analog-to-digital converter. Inside the computer, speech-like spectra are generated from a pole-zero model of the vocal tract and its excitation. (The filter bank characteristics are also applied to the synthetic spectra.) As in the pitch-synchronous analysis, the model is based upon the acoustical principles discussed in Chapter 4. The real and synthetic spectra at a given instant are compared, and a weighted square error is computed. The nature of the comparison is illustrated in Fig. 5.30. The effect of an error in formant frequency is indicated by Fig. 5.30a. An error in formant bandwidth is illustrated in Fig. 5.30b.

On the basis of error computations for the immediate and for adjacent spectral samples, a programmed automatic control strategy determines the procedure for adjusting the pole-zero positions of the fitting synthetic spectrum to reduce the weighted error. When a minimum-error fit is obtained, the computer automatically stores the pole-zero locations of the vocal tract model and the source characteristics chosen for that spectrum. Five operations are carried out by the computer: (1) storage of real

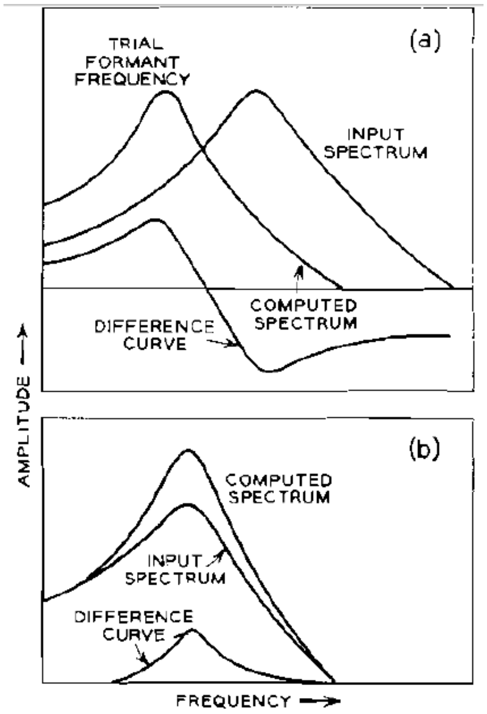


Figure 5.30: Idealized illustration of formant location by the "analysis-by-synthesis" method shown in Fig. 5.29

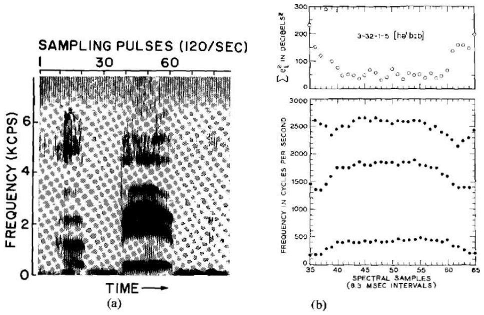


Figure 5.31: Computer-determined formant tracks obtained by the "analysis-by-synthesis" method. (a) Spectrogram of original speech. (b) Extracted formant tracks and square error measure. (After (Bell *et al.* [1961]))

input speech spectra; (2) generation of synthetic spectra; (3) control and adjustment of the synthetic spectra; (4) calculation of spectral difference according to a prescribed error criterion; and (5) storage and display of the parameters which yield minimum error. Provisions are made so that, if desired, the comparison and control functions can be performed by an human operator instead of by the automatic procedure.

In principle the programmed matching procedure is applicable both to vowel and consonant spectra, but the matching model for consonants is generally more complex. A typical result of the procedure is shown for the first three formants in Fig. 5.31. The (a) part of the figure shows a sound spectrogram of the utterance /həbɪb/ with sample intervals laid off along the top time axis. The (b) part of the figure shows the computerdetermined formant tracks for essentially the vowel portion of the second syllable (i.e., /ɪ/). The sample numbers on the abscissa of the (b) part correspond with those at the top of (a). The top diagram in part (b) is the square error for the spectral fit. The "analysis-by-synthesis" technique has also been implemented using a gradient-climbing calculation for matching the short-time spectrum (Olive [1971]). Other implementations have used sequential algorithms for fitting the spectrum (Fujisaki [1960]).

Another computer formant tracker uses a principle related to the pole-zero model of speech (Coker [1965]). The analyzing strategy is a combined peak-picking and spectral fitting approach. A filter bank, associated rectifiers and lowpass filters produce a short-time spectrum. The filter outputs are scanned by an electronic commutator, and the time waveform representing the spectral sections is led to an analog-to-digital converter. The output digital signal describing the successive spectra is read into the computer, and the short-time spectra are stored in the memory.

The automatic analyzing procedure, prescribed by a program, first locates the absolute maximum of each spectrum. A single formant resonance is then fitted to the peak. The single resonance is positioned at a frequency corresponding to the first moment of that spectral portion lying, say, from zero to 6 db down from the peak on both sides. The single formant resonance is then inverse filtered from the real

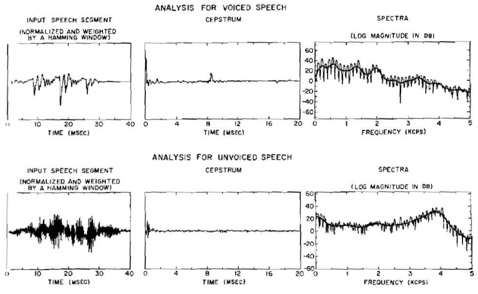


Figure 5.32: Spectrum and cepstrum analysis of voiced and unvoiced speech sounds. (After (Schafer and Rabiner [1970]))

speech spectrum by subtracting the log-amplitude spectral curves. The operation is repeated on the remainder until the required number of formants are located. Since the peak picking is always accomplished on the whole spectrum, the problem of formant segmentation is obviated! Proximate formants can also be resolved and accurate results can be obtained on running speech. The formant selections can be displayed directly on the spectral sections in a manner similar to that shown in Fig. 5.5. Again, the ability of the computer to store large amounts of data and to perform relatively complex operations at high speed permits a detailed fitting of the spectrum. The analysis is easily accomplished in real time, and the computer can essentially be used as the formant-tracking element of a complete formant-vocoder system (Coker and Cummiskey [1965]).

A still different method for formant analysis (Schafer and Rabiner [1970]) makes use of a special digital transform–the Chirp-Z transform (Rabiner *et al.* [1969]). The method also incorporates Fast Fourier Transform methods for spectral analysis (Cooley and Tukey [1965]). In its complete form, the method depends upon relations prescribed by a 3-pole model of voiced sounds and a single pole-zero model of voiceless sounds.

The point of departure is a short-time transform of the speech waveform for both voiced an voiceless sounds. The steps in the spectral analysis are depicted in Fig. 5.32.

The upper part of the figure shows the analysis of voiced speech. The waveform at the top left is a segment of voiced speech of approximately 40 msec duration, which has been multiplied by a Hamming window[7]. Over such a short time interval, the speech waveform looks like a segment of a periodic waveform. The detailed time variation of the waveform during a single period is primarily determined by the vocal

---

[7] The Hamming window is specified by the function

$$h(t) = \left\{0.54 + 0.46\cos\left(\frac{2\pi t}{\tau}\right)\right\} \quad \text{for} \quad -\frac{\tau}{2} \leq t \leq \frac{\tau}{2},$$

where τ is the window duration. This data window is attractive because the side lobes of its Fourier transform remain more than 40 db down at all frequencies (Blackman and Tukey [1959]).

tract response, while the fundamental period (pitch period) reflects the vocal-cord vibration rate.

The logarithm of the magnitude of the Fourier transform of this segment of speech is the rapidly-varying spectrum plotted at the top right of Fig. 5.32. This function can be thought of as consisting of an additive combination of a rapidly-varying periodic component, which is associated primarily with the vocal-cord excitation, and a slowlyvarying component primarily due to the vocal-tract transmission function. Therefore, the excitation and vocal-tract components are mixed and must be separated to facilitate estimation of formant values. The standard approach to the problem of separating a slowly-varying signal and a rapidly-varying signal is to employ linear filtering. Such an approach applied to the log magnitude of the short-time Fourier transform leads to the computation of the cepstrum (Bogert *et al.* [1963]).

The cepstrum is a Fourier transform of a Fourier transform. To compute the cepstrum the Fourier transform of the time waveform is computed. The logarithm is taken of the magnitude of this transform. Inverse Fourier transformation of this log-magnitude function produces the cepstrum. (See also Section 5.6.)

The cepstrum is plotted in the middle of the top row of Fig. 5.32. The rapidly-varying component of the log-magnitude spectrum contributes the peak in the cepstrum at about 8 msec (the value of the pitch period). The slowly-varying component corresponds to the low-time portion of the cepstrum. Therefore, the slowly-varying component can be extracted by first smoothly truncating the cepstrum values to zero above about 4 msec, and then computing the Fourier transform of the resulting truncated cepstrum. This yields the slowly-varying curve which is superimposed on the short-time spectrum, shown at the right of the top row in Fig. 5.32.

The formant frequencies correspond closely with the resonance peaks in the smoothed spectrum. Therefore, a good estimate of the formant frequencies is obtained by determining which peaks in the smoothed spectrum are vocal tract resonances. Constraints on formant frequencies and amplitudes, derived from a three-pole model of voiced sounds, are incorporated into an alogrithm which locates the first three formant peaks in the smoothed spectrum. The analysis of unvoiced speech segments is depicted in the bottom row of Fig. 5.32. In this case, the input speech resembles a segment of a random noise signal. As before, the logarithm of the magnitude of the Fourier transform of the segment of speech can be thought of as consistmg of a rapidly-varying component, due to the excitation, plus a slowly-varying component due to the spectral shaping of the vocal-tract transfer function. In this case, however, the rapidly-varying component is not periodic but is random. Again the low-time part of the cepstrum corresponds to the slowly-varying component of the transform, hut the high-time peak present in the cepstrum of voiced speech is absent for unvoiced speech. Thus, the cepstrum can also be used in deciding whether an input speech segment is voiced or unvoiced, and if voiced, the pitch period can be estimated from the location of the cepstral peak. Low-pass filtering of the logarithm of the transform, by truncation of the cepstrum and Fourier transformation, produces the smoothed

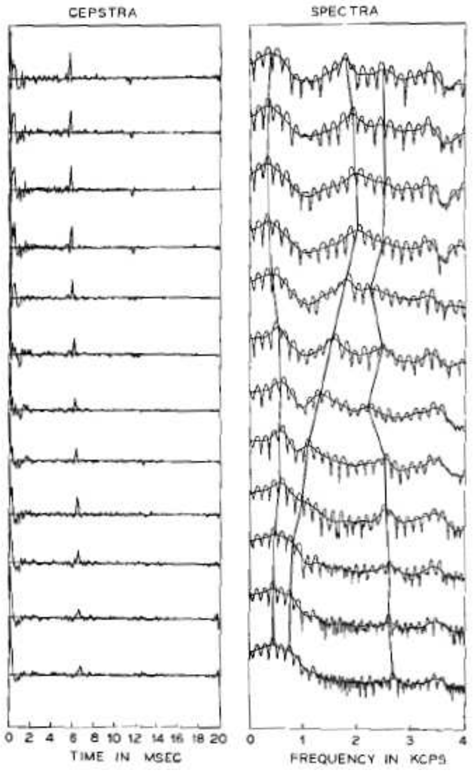


Figure 5.33: Cepstrum analysis of continuous speech. The left column shows cepstra of consecutive segments of speech separated by 20 ms. The right column shows the corresponding short-time spectra and the cepstrally-smoothed spectra

spectrum curve which is again superimposed on the short-time transform at the lower right of Fig. 5.32. In this case, an adequate specification of the spectrum shape can be achieved by estimating the locations of a single wide-bandwidth resonance and a single anti-resonance, i.e., a single pole and zero.

Continuous speech is analyzed by performing these operations on short segments of speech which are selected at equally-spaced time intervals, typically 10-20 msec apart. Fig. 5.33 illustrates this process for a section of speech which, as evidenced by the peaks in the cepstra, is voiced throughout. The short-time spectrum and smoothed spectrum corresponding to each cepstrum are plotted adjacent to the cepstrum. In going from top to bottom in Fig. 5.33, each set of curves corresponds to the analysis of segments of speech selected at 20 msec increments in lime. The formant peaks determined automatically by the program are connected by straight lines. Occasionally the formants come close together in frequency and pose a special problem in automatic extraction.

In the third and fourth spectra from the top, the second and third formants are so close together that there are no longer two distinct peaks. A similar situation occurs in the last four spectra where the first and second formants are not resolved. A

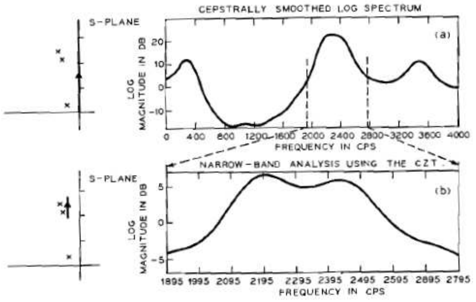


Figure 5.34: Enhancement of formant frequencies by the Chirp-z transform: (a) Cepstrally-smoothed spectrum in which F_2 and F_3 are not resolved. (b) Narrow-band analysis along a contour passing closer to the poles. (After (Schafer and Rabiner [1970]))

procedure for detecting such situations has been devised and a technique for enhancing the resolution of the formants has been developed. An example of the technique is shown in Fig. 5.34.

The curve shown in Fig. 5.34a is the smooth spectrum as evaluated along the $j\omega$-axis of the complex frequency s-plane. (The lowest three vocal tract eigen-frequencies corresponding to this spectrum are depicted by the x's in the s-plane at the left.) Because formants two and three (F2 and F3) are quite close together, only one broad peak is observed in the conventional Fourier spectrum. However, when the spectrum is evaluated on a contour which passes closer to the poles, two distinct peaks are in evidence, as shown in Fig. 5.34b. The Chirp z-transform alogrithm facilitates this additional spectral analysis by allowing a fast computation of the spectrum along an s-plane contour shown at the left of Fig. 5.34b.

Once the vocal excitation and formant functions are determined, they can be used to synthesize a waveform which resembles the original speech signal. (Systems for speech synthesis from formant data are discussed in Section 6.2.) Comparison of the formant-synthesized signal with the original speech signal is an effective means for evaluating the auomatic formant tracking. Fig. 5.35 shows a typical result of automatic analysis and synthesis of a voiced sentence. The upper curves show the pitch period and formant parameters as automatically estimated trom a natural utterance whose spectrogram is also shown in the figure. The bottom of the figure shows the spectrogram of speech synthesized from the automatically estimated pitch and formant parameters. Comparison of the spectograms of the original and synthetic speech indicates that the spectral properties are reasonably well preserved.

Another approach using computer processing is the analysis of real speech spectra in terms of a model of articulation (Heinz [1962], Heinz and Stevens [1964]). This approach differs from the preceding techniques essentially in the spectrum-generation and control strategy operations. The vocal tract poles and zeros are obtained from an articulatory or area function specification of the tract. These are obtained by

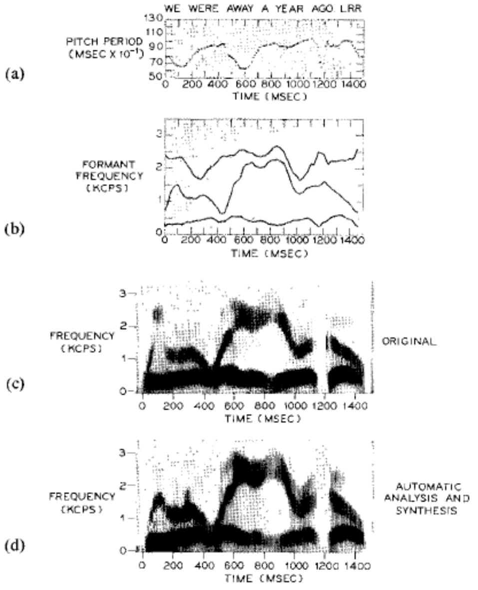
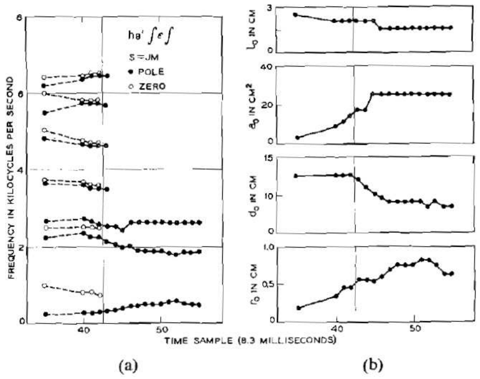


Figure 5.35: Automatic formant analysis and synthesis of speech. (a) and (b) Pitch period and formant frequencies analyzed from natural speech. (c) Spectrogram of the original speech. (d) Spectrogram of synthesis speech. (After (Schafer and Rabiner [1970]))

solving the Webster horn equation (see Chapter 4). A spectrum corresponding to the computed poles and zeros is generated and compared to the real speech spectrum. The error in fit is used to alter the synthetic spectrum by adjusting, on the articulatory level, the vocal tract area function. A modification of a three-parameter description of vocal configuration is used to specify the area function (Dunn [1950], Stevens and House [1955], Fant [1960]).

This formulation, provided the area function can be specified accurately enough, offers an important advantage over pole-zero models of the vocal system. The latter have as their input parameters the locations in the complex plane of the poles and zeros of the vocal transmission. The poles of the system are independent of source location and depend only on the configuration (see Chapter 4). They move in a continuous manner during the production of connected speech, even though the source may change in character and location. The zeros, however, depend upon source location as well as upon tract configuration. They may move, appear and disappear in a discontinuous fashion. This discontinuous behavior of the zeros–and the resulting large changes in the speech spectrum–makes pole-zero tracking difficult.

An articulatory description of the signal obviates these difficulties to a considerable extent. More realistic continuity constraints can be applied to the articulators. The location of the unvoiced source is generally implied by the configuration, and the vocal zero specification is an automatic by-product of the specification of configuration



(a) (b)

Figure 5.36: Pole-zero computer analysis of a speech sample using an articulatory model for the spectral fitting procedure. The (a) diagram shows the pole-zero positions calculated from the articulatory model. The (b) diagram shows the articulatory parameters which describe the vocal tract area function. (After (Heinz [1962]))

and excitation. In terms of articulatory parameters, the spectra of consonants and consonant-vowel transitions can be matched with little more difficulty than for vowels. A typical result of this articulatory fitting procedure is shown in Fig. 5.36.

I,'ig. 5.31 a and b.

The left diagram shows the temporal courses of the poles and zeros in the /ʃɛ/ portion of the bisyllabic utterance /həʃɛʃ/ (the time scale is the sample number multiplied by 8.3 msec). The vertical line, where the zero tracks disappear, represents the consonant-vowel boundary. (Only the first three formants are computed in the vowel part of the utterance.) The diagram to the right shows the corresponding temporal courses of the four articulatory parameters that were adjusted to make the spectral matches. They are:

r_0, the effective radius at the tongue constriction,
d_0, the location of the tongue constriction measured from the glottis,
a_0, the cross-sectional area of the mouth opening, and
l_0, the length of the lip tube (or mouth section).

Their trajectories are essentially continuous as the match proceeds across the consonant-vowel boundary. In going from the fricative /ʃ/ to the vowel /ɛ/, the mouth section becomes shorter and more open. The position of the constriction moves back toward the glottis, and the radius of the constriction becomes larger. The position of the unvoiced sound source during the fricative is taken 2.5 cm anterior to the constriction (i.e., $d_0 + 2.5$). The manner in which these relatively simple motions describe the more complicated pole-zero pattern is striking. Success of the method depends directly upon the accuracy with which the articulatory parameters describe the vocal-tract shape. Derivation of sophisticated articulatory models is an important area for research. (See Section 5.7.)

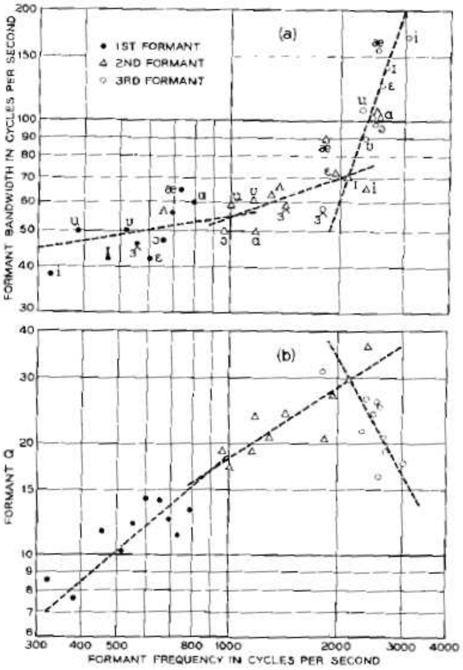


Figure 5.37: Measured formant bandwidths for adult males. (After (Dunn [1961]))

5.5.2 Measurement of Formant Bandwidth

The bandwidths of the formant resonances–or the real parts of the complex poles–are indicative of the losses associated with the vocal system. Not only are quantitative data on formant bandwidths valuable in corroborating vocal tract calculations (for example, those made in Chapter 4 for radiation, viscous, heat-conduction, cavity-wall and glottal losses), but a knowledge of the damping is important in the proper synthesis of speech.

A number of measurements have been made of vocal tract damping and formant bandwidth.[8] The measurements divide mainly between two techniques; either a measure of a resonance width in the frequency domain, or a measure of a damping constant (or decrement) on a suitably filtered version of the speech time waveform. In the former case the formant is considered as a simple resonance, and the half-power frequencies of the spectral envelope are determined. In the latter case the formant is considered a damped sinusoid, having amplitudes A_l and A_2 at times t_1 and t_2. The damping constant, σ, for the wave and its halfpower bandwidth, Δf, are related simply as

$$\sigma = \pi \Delta f = \frac{\ln A_2/A_1}{t_2 - t_1}.$$

---

[8] For a good summary and bibliography of earlier investigations, see (Dunn [1961]). Also, see (Fant [1958, 1959b,a]).

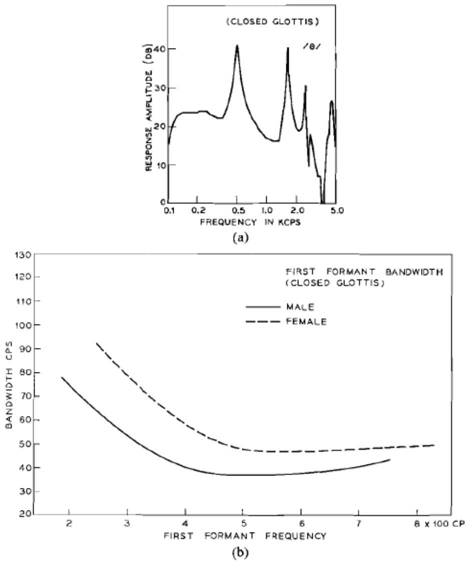


Figure 5.38: (a) Vocal-tract frequency response measured by sine-wave excitation of an external vibrator applied to the throat. The articulatory shape is for the neutral vowel and the glottis is closed. (After (Fujimura and Lindquist [1971])). (b) Variation in first-formant bandwidth as a function of formant frequency. Data for men and women are shown for the closed-glottis condition. (After (Fujimura and Lindquist [1971]))

The results of one of the more extensive formant bandwidth studies are summarized in Fig. 5.37 (Dunn [1961]). Part (a) of the figure shows the formant bandwidths measured by fitting a simple resonance curve to amplitude sections of vowels uttered in an /h–d/ syllable. The data are averages for 20 male voices producing each vowel. The second curve (b) represents the same data plotted in terms of $Q = f/\Delta f$. The upper graph shows that over the frequency ranges of the first and second formants, the nominal bandwidths are generally small–on the order of 40 to 70Hz. Above 2000Hz the bandwidth increases appreciably. The lower plot of formant-Q *vs* formant frequency shows that resonant Q's are largest in the frequency region around 2000Hz.

Formant bandwidths can also be effectively measured from a frequency response of the actual vocal-tract (Fujimura [1962]). A sine wave of volume velocity is introduced into the vocal-tract at the glottal end by means of a throat vibrator. The pressure output at the mouth is measured as the input source is changed in frequency. A typical vocal-tract frequency response is shown in Fig. 5.38a. The variation in first-formant handwidth, as a function of first-formant frequency, is shown in 5.33b.

These data are for a closed-glottis condition. The bandwidth is seen to increase as first formant frequency diminishes, owing primarily to the influence of cavity-wall loss. (See calculations of cavity-wall loss in Section 4.7.3)

The origins of the principal contributions to vocal-tract damping have already been indicated by the theory derived in Chapter 4. These are glottal loss and cavity-wall loss for the lower formants, and radiation, viscous and heat-conduction loss for the higher formants.

5.6 Analysis of Voice Pitch

Fundamental frequency analysis–or "pitch extraction"–is a problem nearly as old as speech analysis itself. It is one for which a complete solution remains to be found. The main difficulty is that voice pitch has yet to be adequately defined. Qualitatively, pitch is that subjective attribute that admits of rank ordering on a scale ranging from low to high. The voiced excitation of the vocal tract is only quasi-periodic. Not only does the exciting glottal waveform vary in period and amplitude, but it also varies in shape. Precisely what epochs on the speech waveform, or even on the glottal waveform, should be chosen for interval or period measurement is not clear. Furthermore, the relation between an interval, so measured, and the perceived pitch is not well established.

Most pitch-extracting methods take as their objective the indication of the epoch of each glottal puff and the measurement of the interval between adjacent pulses. Still, exactly how this relates to the pitch percept with all the random jitter and variation of the glottal wave is a question worthy of inquiry.

Most automatic or machine pitch extractors attempt either to describe the periodicity of the signal waveform (Grutzmacher and Lottermoser [1937], Jr. and Schott [1949], Dolansky [1955], Gill [1959]) or to measure the frequency of the fundamental component if it is present (Dudley [1939a]). Computer efforts at pitch extraction essentially do the same, but usually more elaborate constraints and decisions are applied (Inomata [1960], Gold [1962], Sugimoto and Hashimoto [1962]).

One particularly useful method for machine pitch extraction utilizes properties of the cepstrum to reveal signal periodicity (Noll [1967], Oppenheim *et al.* [1968]). As described in Section 5.5.1, the cepstrum is defined as the Fourier transform of the logarithm of the amplitude spectrum of a signal. Since it is a transform of a transform, and since the resulting independent variable is reciprocal frequency, or time, the terms "cepstrum" and "quefrency" were coined by its inventors (Bogert *et al.* [1963]) to designate the transform and its independent variable.

The log-taking operation has the desirable property of separating source and system characteristic (at least to the extent that they are spectrally multiplicative). If the output speech wave, $f(t)$, is the convolution of the vocal tract impulse response, $v(t)$, and the vocal excitation source, $s(t)$, the magnitudes of their Fourier transforms are related as

$$|F(\omega)| = |V(\omega)| \cdot |S(w)|,$$

where all the amplitude spectra are even functions. Taking the logarithm or both sides gives

$$\ln|F(\omega)| = \ln|V(\omega)| + \ln|S(\omega)|$$

Similarly, taking the Fourier transform[9] of both sides yields

$$\mathcal{F}\ln|F(\omega)| = \mathcal{F}\ln|V(\omega)| + \mathcal{F}\ln|S(\omega)|.$$

For voiced sounds, $|S(\omega)|$ is approximately a line spectrum with components spaced at the pitch frequency $1/T$. $\mathcal{F}\ln|S(\omega)|$ therefore exhibits a strong component at the "quefrency," T. $|V(\omega)|$, on the other hand, exhibits the relatively "slow" formant maxima. Consequently $\mathcal{F}\ln|V(\omega)|$ has its strongest component at a very low quefrency.

Because of the additive property of the transforms of the log amplitude spectra, the characteristics of the source and system are well separated in the cepstrum. The cepstrum is therefore also a valuable tool for formant analysis as well as pitch measurement (Schafer and Rabiner [1970]). (See Section 5.5.1.) Measurement of pitch and voiced-unvoiced excitation is accomplished by using a suitable strategy to detect the quefrency components associated with $\mathcal{F}\ln|S(\omega)|$. Because the method does not require the presence of the fundamental component, and because it is relatively insensitive to phase and amplitude factors (owing to the log-magnitude operations) it performs well in vocoder applications. In one test with a complete channel vocoder, it demonstrated superior performance in extracting the pitch and voiced-unvoiced control data (Noll [1967]). Because a large amount of processing is necessary, the method is most attractive for special purpose digital implementations where Fast Fourier Transform hardware can be used. An illustration of pitch determination by cepstrum computation has been shown previously in Fig. 5.33a and 5.35.

Perhaps a more basic measurement of voiced excitation is that of the glottal volume-velocity wave (Miller [1959], Fant [1959a], Mathews [1959], Holmes [1962]). Approximations to this function can be obtained by so-called inverse-filtering techniques. The idea is to pass the speech signal through a network whose transmission function is the reciprocal of that of the vocal tract for the particular sound. Zeros of the network are adjusted to nullify vocal tract poles, and the resulting output is an approximation to the input glottal volume current.

The inverse-filtering analysis presumes that the source and system relations for the speech-producing mechanism do not interact and can be uniquely separated and treated independently. This assumption is a treacherous one if the objective is an accurate estimate of the glottal volume velocity. In the real vocal tract they interact to a certain extent (particularly at the first-formant frequency). Another difficulty is that it is not always clear whether to ascribe certain properties (primarily, zeros) to the tract or to the source. The estimate obtained for the glottal wave obviously depends upon the vocal-tract model adopted for the inverse filter. The criterion of adjustment of the inverse filter also influences the answer. Under certain conditions, for example,

---

[9] Formally an inverse Fourier transform.

ripples on the inverse wave which may be thought to be formant oscillations might in fact be actual glottal variations.

One question often raised is “where in the pitch period does the excitation occur.” Presumably if such an epoch could be determined, the pulse excitation of a synthesizer could duplicate it and preserve natural irregularities in the pitch period. Because the glottal wave frequently changes shape, such a datum is difficult to describe. One claim is that this epoch commonly is at the close of the cords (Miller [1959]), while another (Holmes [1962]) is that it can occur at other points in the wave. To a first approximation, such an epoch probably coincides with the greatest change in the derivative of the glottal waveform. Often this point can occur just about anywhere in the period. For a triangular wave, for example, it would be at the apex. A perceptual study has been made of the effects of the glottal waveform on the quality of synthetic speech. The results support the notion that the significant vocal excitation occurs at the point of greatest slope change in the glottal wave (Rosenberg [1971b]). Natural speech was analyzed pitch-synchronously. The vocal-tract transmission and the glottal waveform were determined and separated by inverse filtering. Artificial glottal waveforms were substituted and the speech signal was regenerated. Listening tests showed that good quality speech can be obtained from an excitation function fixed in analytical form. The absence of temporal detail, period-to-period, does not degrade quality. A preferred glottal pulse shape has but a single slope discontinuity at closing. It is intrinsically asymmetric, so its spectral zeros never fall on or near the $j\omega$-axis for any combination of opening and closing times (Rosenberg [1971b]).

5.7 Articulatory Analysis of the Vocal Mechanism

The discussion of Chapter 4 showed that if the vocal tract configuration is known, the system response can be computed and the mode slructure specified. The cross-sectional area as a function of distance is sufficient to compute the lower eigenfrequencies of the tract. An accurate account of losses along the tract requires knowledge of the crosssectional shape or the circumference. [See Eq. 3.33).] Because the vocal mechanism is relatively inaccessible, the necessary dimensions are obviously difficult to obtain. Even at best, present methods of measurement yield incomplete descriptions of tract dimensions and dynamics.

X-ray techniques for motion and still pictures have provided most of the articulatory information available to date. The X-ray data generally are supplemented by other measures. Conventional moving pictures can be made of the external components of the vocal system. Palatograms, molds of the vocal cavities, and electromyographic recordings are also useful techniques for “filling in the picture.” Much of the effort in X-ray analysis is directed toward therapeutic goals, such as cleft palate repair and laryngeal treatment. Consequently, the results are often left in only a qualitative form. Several investigations, however, have aimed at measuring vocal dimensions and articulatory movements(Fant [1960], Chiba and Kajiyama [1941], Perkell [1965],

Figure 5.39: Sagittal plane X-ray of adult male vocal tract

Fujimura [1961], Houde [1967]).

One of the main problems in obtaining such data is keeping the radiation dose of the subject within safe limits. This usually means that only a very limited amount of data can be taken on a single individual. One ingenious solution to this problem utilizes a computer-controlled X-ray beam which, under program control, is made to irradiate and track only the physiological areas of interest (Fujimura [1961]).

Another problem is the detail of the X-ray photograph. This is particularly a problem in moving X-ray photography, even with the best image-intensifier tubes. Detail which looks deceptively good in the (visually-integrated) moving picture, disappears when one stops the film to study a single frame. Sound recordings are usually made simultaneously for analysis, but often are of poor quality because of the noise of the proximate movie camera.

The detail in still pictures is somewhat better but nevertheless lacking. An example of a typical medical X-ray is shown in Fig. 5.39. The tongue and lips of the subject were coated with a barium compound to make them more visible. The vocal tract position is appropriate to the production of a high-front vowel close to /i/.

The typical procedure for obtaining an area function from the X-ray picture can be illustrated. An axial line through the centers of gravity of the cross sectional areas is first located, as shown in Fig. 5.40a (Fant [1960]). The shape and area of the cross-sections at a number of locations are estimated, as shown in Fig. 5.40b. The shape estimates are deduced on the basis of all available data, including dental molds of the vocal and nasal cavities, conventional photographs and X-ray photographs from the front. These sections provide anchor points for an estimate of the whole area

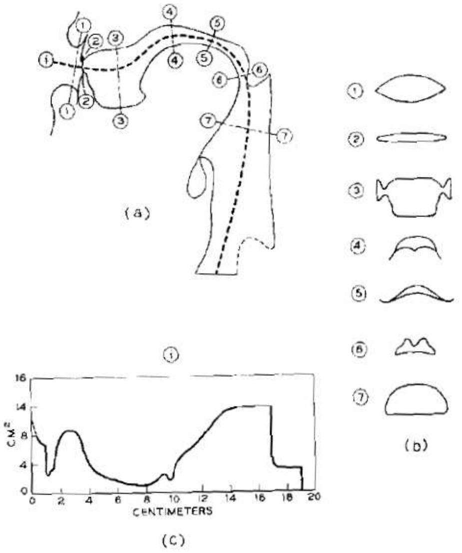


Figure 5.40: Method of estimating the vocal tract area function from X-ray data. (After (Fant [1960]))

curve. Intermediate values are established both from the sagittal plane X-ray tracing and from continuity considerations to give the complete area function, as shown in Fig. 5.40c. Typical results for several sounds produced by one man are shown in Fig. 5.41.

Even under best conditions, some of the vocal dimensions during natural speech are impossible to measure. For example, one often can only make crude estimates of the true shape and lateral dimensions of the pharynx cavity. In the same vein, the true dimensions of the constrictions for fricatives and affricates and the lateral pathways in /l/ are often very uncertain.

Similarly, the vocal source of excitation cannot be studied easily by direct methods. For sustained, open vowels, however, the vocal cord source can be examined by high-speed moving pictures. Measurements of subglottal pressure are also possible and give insight into vocal cord operation. Characteristics of the unvoiced sources, on the other hand, i.e., location, spectral properties and internal impedance, are best inferred from physiological configuration, air flow measurements and spectral analysis of the output sound.

Research interest in better methods for physiological measurements remains high. One active research area centers on the possibilities for relating electromyographic recordings of muscle potentials to the articulator movements observed in X-ray pictures. Several "exotic" schemes for vocal measurement have also been proposed, half humorously. They may, however, hold some promise. For example, a conducting dag loop might be painted around the circumference of the tract at a given position and electrical leads attached. The cross sectional area at that point could be measured

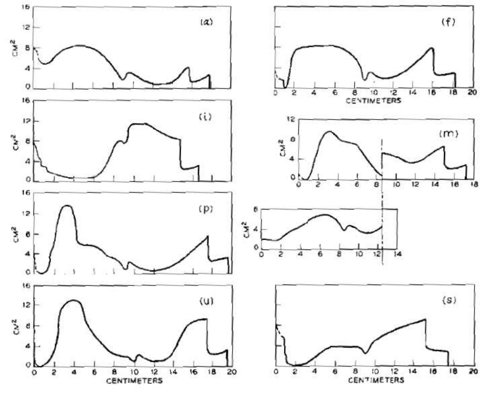


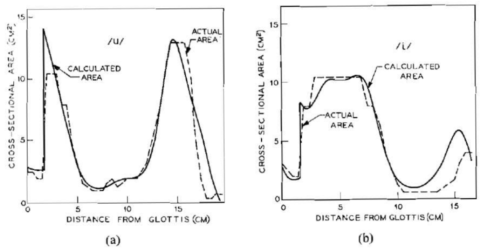
Figure 5.41: Typical vocal area functions deduced for several sounds produced by one man. (After (Fant [1960]))

by placing the subject in a magnetic field normal to the section and measuring the flux which links the dag loop. Other possibilities might be the attachment of miniature strain gauges at significant points, or the placement of inflatable annular cuffs or catheters at given positions in the tract. Still other possibilities include miniature ultrasonic transducers fixed to the articulators.

Acoustic measurements directly on the vocal-tract also promise useful estimation of the cross-sectional area function (Mermelstein [1967], Schroeder [1967], Gopinath and Sondhi [1970])). In one method the acoustic impedance of the tract is periodically sampled at the mouth (Gopinath and Sondhi [1970]). While the subject silently articulates into an impedance tube, pulses of sound pressure are produced periodically (typically at 100 $\sec^{-1}$) and the volume velocity response is measured. The pressure and volume velocity along the tract are assumed to obey Webster's horn equation [Eq. (**??**)], which is valid for frequencies below about 4000Hz. An asymptotic high-frequency behavior of the tract is assumed. No assumptions are made about the termination at the glottal end or about the length of the tract. Solution of an integral equation yields the integral of the cross-sectional area of an equivalent lossless, hard-walled pipe as a function of distance. Differentiation gives the area function. Typical results, compared to area functions from X-ray measurements, are shown in Fig. 5.42. The impedance tube calculations are made for hard-walled vocal-tracts having the shapes given by the X-ray data.

A question of considered importance is the influence of wall-yielding (as is present in the real vocal tract) upon the calculated area function. Present efforts aim to include wall vibration and wall loss into the area determination method. Further research is needed to test the method with real speakers and real speech, and to account for real vocal-tract conditions, including loss, yielding side walls and nasal coupling.

Vocal-tract models, electrical vocal-tract analogs and computational analyses have all been useful in inferring articulatory data and tract dynamics from acoustic mea-



(a)

(b)

Figure 5.42: Typical vocal-tract area functions (solid curves) determined from impedance mcusurements at the mouth. The actual area functions (dashed curves) are derived from X-ray data. (After (Gopinath and Sondhi [1970]))

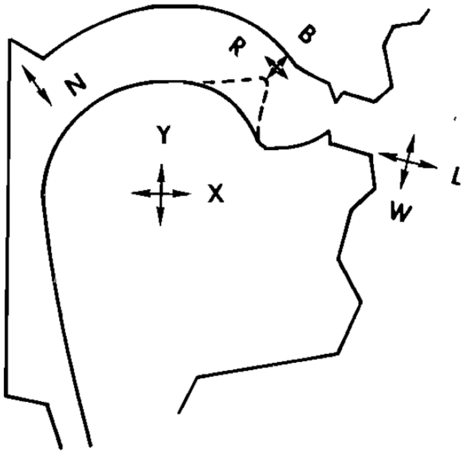


Figure 5.43: Seven-parameter articulatory model of the vocal tract. (After (Coker [1968]))

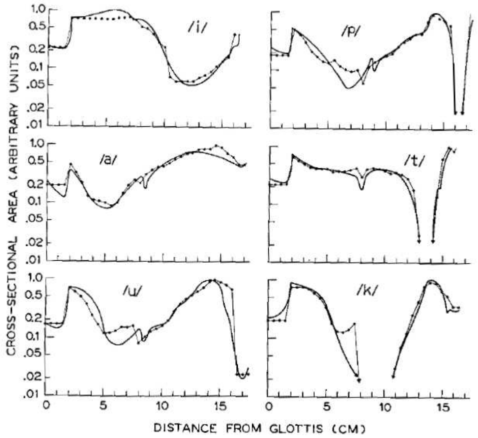


Figure 5.44: Comparison of vocal tract area functions generated by the artculatory model of Fig. 5.43 and human area data from X-rays. (After (Coker [1968]))

surements of speech sounds and from X-ray data. One articulatory model, which has an important application in synthesis (see Section 11.5), has also been useful in establishing physiological constraints and time constants associated with major articulators (Coker [1968])). The articulatory model describes the vocal area function in terms of seven parameters, shown in Fig. 5.43. The coordinates are: the position of the tongue body, X, Y; the lip protrusion, L; the lip rounding W; the place and degree of tongue tip constriction, R and B; and the degree of velar coupling, N. No nasal tract is incorporated in this version of the model, and velar coupling exerts its influence solely through the tract area function.

The area function described by the model can be used to synthesize connected speech, which in turn can be compared in spectral detail to real speech. Also, because of its correspondence to major vocal elements, the seven-parameter model can be used to duplicate articulatory motions observed from X-ray motion pictures. Further, its description of vocal-tract area can be compared with X-ray area data, as shown in Fig. 5.44. Such comparisons have been useful in analyzing priorities and time-constants for the motions of the articulators in real speech and in quantifying these effects for speech synthesis (Coker *et al.* [1971], Flanagan *et al.* [1970], Umeda [1970]).

5.8 Homework

Problem 5.1

A speech signal is sampled at a rate of 20,000 samples per second. A 12ms window is used for short-time spectral analysis.

a. How many speech samples are used in each segment?

b. If the window is a rectangular window, what analysis frame rate (in frames per second) will guarantee that no frequency-aliasing occurs? (Assume that the side-lobes have zero amplitude.)

c. If the window is a Hamming window, what analysis frame rate (in frames per second) will guarantee that no frequency-aliasing occurs? (Assume that the side-lobes have zero amplitude.)

d. What size Fast Fourier Transform is required to guarantee that no time-aliasing will occur?

e. Suppose that the window $w(n)$ is an ideal low-pass filter with a cutoff frequency of $f_c = 312.5$Hz. What size FFT should you use in order to construct a filter-bank with non-overlapping filters? Will time-aliasing occur?

Problem 5.2

Consider the following non-causal triangular window:

$$w_t[n] = \frac{1}{N} w_r[n] * w_r[-n] \tag{5.148}$$

where

$$w_r[n] = u[n] - u[n-N] \tag{5.149}$$

a. Sketch $w_t[n]$.

b. Notice the following property of the Fourier transform:

$$x[-n] \leftrightarrow X(-\omega)$$

Use the property above, the conjugate symmetry, and the convolution properties of the Fourier transform to show that

$$W_t(\omega) = \frac{1}{N} |W_r(\omega)|^2 \tag{5.150}$$

Sketch $W_t(\omega)$.

c. Now consider the following triangular window:

$$w_t[n] = \frac{N - |N-n|}{N}\left(u[n] - u[n-2N]\right) \tag{5.151}$$

Find $W_t(\omega)$. Hint: use the time-delay property.

Problem 5.3

Two of the most commonly used DSP windows—the Hanning window and Hamming window—can be written in the following form. In this equation, N must be odd, and B is the filter design parameter: $B = 0.5$ for a Hanning window, and $B = 0.46$ for a Hamming window.

$$w[n] = r[n]\left((1-B) + B\cos(2\pi n/N)\right)$$

where $r[n]$ is the zero-centered rectangular window

$$r[n] = \begin{cases} 1 & |n| \le (N-1)/2 \\ 0 & |n| > (N-1)/2 \end{cases}$$

a. What is the DTFT $R(\omega)$ of $r[n]$? At what frequency is the first null of $R(\omega)$? What is the amplitude of the first sidelobe of $R(\omega)$?

b. Express $W(\omega)$ as the sum of three scaled and frequency-shifted copies of $R(\omega)$. At what frequency is the first null of $W(\omega)$?

c. In terms of B, what is the amplitude of the first sidelobe of $W(\omega)$? Find the value of B which minimizes the amplitude of the first sidelobe, and say what that minimum amplitude turns out to be.

d. Sketch $H(\omega)$, the DTFT of the following digital filter. Label the amplitude, peak frequencies, and the frequencies of one or two zero crossings.

$$h[n] = \cos(\pi n/3)w[n]$$

e. Sketch $G(\omega)$, the DTFT of the following digital filter. Label the amplitude and cutoff frequencies.

$$g[n] = 0.25\text{sinc}(\pi n/4)\cos(\pi n/3)w[n]$$

Problem 5.4

Suppose that the autocorrelation coefficients of signal $x_1(n)$ are $R(0) = 500$ and $R(1) = 400$.

a. A first-order linear predictive model of the transfer function is given by $T_1(z) = G_1/(1 - a_1 z^{-1})$. Find G_1 and a_1.

b. Suppose that the spectrum of another signal, $x_2(n)$, can be modeled using the following transfer function:

$$T_2(z) = \frac{10}{1 - 0.9z^{-1}} \tag{5.152}$$

What is the Itakura-Saito distance between $T_1(z)$ and $T_2(z)$?

Problem 5.5

Write a program of the form **A = lpcana(X, P, N)** which performs LPC analysis of order **P** on the waveform **X** using frames of length **N** samples. The matrix **A** should contain one row for each frame; each row should contain the LPC filter coefficients for one frame.

Write a program **[F,BW] = formants(A,FS)** which finds the roots of the LPC polynomials stored in **A**, and calculates up to $p/2$ analog formant bandwidths BW_i and frequencies F_i per frame, such that

$$r_i = e^{-\frac{\pi BW_i + j2\pi F_i}{F_s}} \tag{5.153}$$

where r_i is one of the roots of $A(z)$.

Plot the formant frequencies as a function of time, and compare your plot to a spectrogram of the utterance. Which formants are tracked during voiced segments? What happens when there are less than $p/2$ trackable formants? What happens to the LPC-based formant estimates during unvoiced speech segments?

Problem 5.6

Write a program **[N0, B] = pitch(X, N, N0MIN, N0MAX)**. For each frame, set **B** and **N0** according to the following formulas:

$$B = \max r_x(m), \qquad N_{0,min} \le m \le N_{0,max} \tag{5.154}$$

$$N_0 = \arg\max r_x(m), \qquad N_{0,min} \le m \le N_{0,max} \tag{5.155}$$

$$\tag{5.156}$$

a. Try different values of $N_{0,min}$ and $N_{0,max}$. For each value you test, plot **B** and **N0** as a function of time, and compare them to the spectrogram. What values give the best pitch tracking? Is **B** always larger for voiced segments than unvoiced segments? What is the threshold value of **B** which best divides voiced and unvoiced segments?

b. Try pitch tracking using the autocorrelation of the LPC residual, rather than the signal autocorrelation. Do you find any improvement? Why or why not?

Problem 5.7

Consider the sequence

$$x[n] = \delta[n] - a\delta[n-1] \tag{5.157}$$

where $|a| < 1$. Suppose that we wish to approximate the complex cepstrum $\hat{x}[n]$ from samples of the logarithm of the Fourier transform:

$$\hat{x}_p[n] = \frac{1}{N}\sum_{k=0}^{N-1} \log\left(X(e^{j\frac{2\pi}{N}kn})\right) \tag{5.158}$$

Is it possible to choose N large enough such that $\hat{x}_p[n] = \hat{x}[n]$, without aliasing? If so, what is the minimum value of N? If not, what is the minimum value of N (give or take a few samples) such that

$$|\hat{x}_p[n] - \hat{x}[n]| < \left|\frac{\hat{x}[n]}{100}\right| \tag{5.159}$$

Note: You may find the following formula to be useful:

$$\log(1-x) = -\sum_{n=1}^{\infty}\frac{x^n}{n} \quad \text{if } |x| < 1 \tag{5.160}$$

Problem 5.8

Suppose that homomorphic analysis yields the following estimate of the vocal tract transfer function:

$$H(z) = \frac{G}{1-\sum_{k=1}^{2q} a_k z^{-k}} = G\prod_{k=1}^{q}\frac{1}{(1-b_k z^{-1})(1-b_k^* z^{-1})} \tag{5.161}$$

with pole locations $b_k = r_k e^{j\theta_k}$ and $b_k^* = r_k e^{-j\theta_k}$ which are located inside the unit circle. If the sampling rate is F_s, then the formant frequencies F_k and bandwidths B_k can be estimated by:

$$\hat{F}_k = \frac{F_s\theta_k}{2\pi} \tag{5.162}$$

$$\hat{B}_k = -\frac{F_s}{\pi}\log(r_k) \tag{5.163}$$

Suppose that we suspect that all of the bandwidth estimates $\hat{B}_k$ are too large. Show that the estimated formant bandwidths are reduced, without changing the estimated formant frequencies, if we replace $H(z)$ by the following transformed spectrum:

$$\tilde{H}(z) = H\left(\frac{z}{\alpha}\right) = G\prod_{k=1}^{q}\frac{1}{(1-b_k(z/\alpha)^{-1})(1-b_k^*(z/\alpha)^{-1})} \tag{5.164}$$

where α is real and greater than unity and $|\alpha b_k| < 1$.

Suppose $H(z)$ consists of a single complex pole pair of the form

$$H(z) = \frac{1}{(1-re^{j\theta}z^{-1})(1-re^{-j\theta}z^{-1})} \tag{5.165}$$

where r and θ are both real. Find expressions for the complex cepstra associated with $H(z)$ and $\tilde{H}(z)$ in this case. Find expressions for the real cepstra, and plot the real cepstra as functions of time.

Problem 5.9

Compute the following spectra for three different vowel segments segments, and plot the log-magnitude spectra (in dB) for frequencies between 0 and 4000Hz. You should turn in code, equations, or some combination of both which will make it clear how each spectrum was computed.

a. Power spectrum $S(\omega)$.

b. $S(\omega)$, smoothed using cepstral lifter $\hat{w}_1(n)$:

$$\hat{w}_1(n) = u(n-1) - u(n-L-1) \tag{5.166}$$

Choose L so that the window length is 1.5ms. What is the cutoff frequency of the magnitude lifter spectrum, $\tilde{W}_1(\omega) = |\hat{W}_1(\omega)|$?

c. $S(\omega)$, smoothed using cepstral lifter $\hat{w}_2(n)$:

$$\hat{w}_2(n) = \hat{w}_1(n)\left(1 + \frac{L}{2}\sin(\frac{n\pi}{L})\right) \tag{5.167}$$

d. LPC transfer function $H(\omega) = G/A(\omega)$.

e. $H(\omega)$, smoothed using cepstral lifter $\hat{w}_1(n)$.

f. $H(\omega)$, smoothed using cepstral lifter $\hat{w}_2(n)$.

g. Line spectra $1/P(\omega)$ and $1/Q(\omega)$, truncated at reasonable maximum and minimum values.

Problem 5.10

The three vowel signals you analyzed in problem 5.8 are different — but how different are they? Calculate the difference between the two vowels using the following spectral distortion metrics. Turn in code and/or equations showing how each distortion metric was computed.

a. L_2 spectral norm, calculated using log-FFT spectra.

b. Truncated cepstral distance $\hat{d}_c^2$ using a rectangular window.

c. Liftered cepstral distance $\hat{d}_2^2(L)$ using half of a Hamming window.

d. Likelihood-ratio distortions,

$$d_{LR}\left(\frac{1}{|A_1|^2}, \frac{1}{|A_2|^2}\right) \quad \text{and} \quad d_{LR}\left(\frac{1}{|A_2|^2}, \frac{1}{|A_1|^2}\right) \tag{5.168}$$

where the subscripts 1 and 2 represent the first and second vowel.

e. Itakura-Saito distortions,

$$d_{IS}\left(\frac{G_1^2}{|A_1|^2}, \frac{G_2^2}{|A_2|^2}\right) \quad \text{and} \quad d_{IS}\left(\frac{G_2^2}{|A_2|^2}, \frac{G_1^2}{|A_1|^2}\right) \tag{5.169}$$

Problem 5.11

Remember that a filter characteristic $H(\omega)$ is defined by its magnitude $|H(\omega)|$ and phase $\angle H(\omega)$. One particularly useful representation of the phase of $H(\omega)$ is the group delay $\tau_H(\omega)$, defined as

$$\tau_H(\omega) = -\frac{d\angle H(\omega)}{d\omega} \tag{5.170}$$

In general, a linear phase filter has a constant group delay—and the constant is equal to the amount by which the output is delayed with respect to the input. If $\tau(\omega)$ is not constant, components of the input $x[n]$ at different frequencies will be delayed by different amounts. If the differences are large, the result is a sort of reverberated sound.

a. In order to understand group delay, consider the filter

$$h_1[n] = \delta[n - D]$$

Suppose that

$$y_1[n] = h_1[n] * x[n]$$

Find a simple representation of $y_1[n]$ in terms of $x[n]$.

b. What is the DTFT of this filter, $H_1(\omega)$? What is the group delay? How is the group delay related to your answer to part a?

c. Consider $h_2[n]$, given by

$$h_2[n] = w[n]\left(\frac{\omega_c}{\pi}\right)\mathrm{sinc}\left(\omega_c(\mathrm{n}-\frac{\mathrm{N}-1}{2})\right) \tag{5.171}$$

where

$$w[n] = u[n] - u[n-N] \tag{5.172}$$

Sketch $h_2[n]$. By inspection (without doing any equation manipulation), find the group delay $\tau_2(\omega)$. Now do a little symbol manipulation: express the magnitude response, $|H_2(\omega)|$, in the form of the frequency domain convolution between two functions.

Problem 5.12

The process of voiced speech production (e.g., sung vowels) can be modeled as a linear filter,

$$y[n] = h[n] * x[n]$$

where $x[n]$ is a periodic excitation signal modeling the volume velocity coming through the singer's vocal folds,

$$x[n] = \sum_{r=-\infty}^{\infty} x_0[n+rN]$$

and $h[n]$ is an infinite-length impulse response modeling the frequency response of the mouth. Suppose it were possible to estimate the signal $x[n]$ in some way; then an approximation of $h[n]$ would be given by

$$\hat{h}[n] = \frac{1}{N}\sum_{k=0}^{N-1}\frac{Y(2\pi k/N)}{X(2\pi k/N)}e^{\frac{j2\pi kn}{N}}$$

where $Y(2\pi kn/N)$ and $X(2\pi kn/N)$ are the N-point DFT of any single period of $x[n]$ and $y[n]$, respectively, and N is the fundamental pitch period of $x[n]$. What is the relationship between $\hat{h}[n]$ and $h[n]$? Justify your answer.

Problem 5.13

Suppose that $x[n]$ is a cosine, given by

$$x[n] = \cos(\omega_0 n)$$

a. Suppose that the STFT $X_m(e^{j\psi})$ is computed using a rectangular window of length N. Find $X_m(e^{j\psi})$.

b. Suppose that the STFT is only computed once per M samples. Find $X_{fM}(e^{j\psi})$, the STFT of frame number f.

Problem 5.14

a. Suppose that the input to a room response is a periodic signal $x[n] = v[(n)_N]$, where the $()_N$ notation means "modulo N," i.e.

$$x[n] = \begin{cases} v[n+N] & -N \le n \le -1 \\ v[n] & 0 \le n \le N-1 \\ v[n-N] & N \le n \le 2N-1 \\ \vdots & \end{cases}$$

$x[n]$ is played through a speaker-room-microphone system with impulse response $h[n]$, so that $y[n] = h[n] * x[n]$ is the linear convolution of $h[n]$ and $x[n]$. Show that $y[n]$ is therefore the circular convolution of $v[n]$ and $h[n]$, repeated periodically with period N. Argue that therefore $Y(k) = X(k)H(k)$, where $Y(k)$ is the N-point DFT of $y[n]$.

b. Define the circular autocorrelation $r_v[n]$ and the estimated impulse response $q[n]$ as in the following equation

$$r_v[n] = \sum_{m=0}^{N-1} v[m]v[(m+n)_N], \quad q[n] = y[n] * v[-n] \tag{5.173}$$

Show that $q[n]$ is a time-aliased periodic repetition of $\hat{h}[n] = r_v[n] \circledast h[n]$, i.e.

$$q[n] = \sum_{k=-\infty}^{\infty} \hat{h}[n-kN], \quad \hat{h}[n] = r_v[n] \circledast h[n]$$

where $\circledast$ denotes circular convolution. Argue that therefore, if $r_v[n] = \delta[n]$, then $q[n]$ is a time-aliased periodic repetition of the true room response $h[n]$.

Chapter 6

Information and Communication

> "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages."
> SHANNON, *The Mathematical Theory of Communication* (Shannon and Weaver [1949])

The theory of communication addresses the relationship between transmitted and received *messages*. In everyday discourse, a *message* is usually understood to be a sequence of words, sounds, or symbols representing some intended meaning. Shannon proposed that the essential property of a message is not the meaning, but the sequence of symbols. According to his notation, the essential property of a message is that it is a sequence of symbols, selected from a set of possible symbol sequences according to some (usually unknown) probability distribution. Formally, we may write that a message X is selected from the set $\mathcal{X}^*$, where the $*$ superscript is called the "Kleene closure" operator, and it means that:

$$X = [x_1, x_2, \ldots, x_{|X|}], \tag{6.1}$$

$$x_t \in \mathcal{X} \tag{6.2}$$

$$0 \le |X| < \infty \tag{6.3}$$

The set $\mathcal{X}$ is called the "alphabet." The alphabet may be discrete or continuous, finite or infinite. For example, written English words are created by selecting letters from a 26-letter alphabet. On the other hand, an arbitrary acoustic waveform $x(t)$ can also be treated as a message, if we are willing to sample the waveform at discrete sampling times: then $x_1 = x(t_1)$, $x_2 = x(t_2)$, etc., and these "symbols" are acoustic pressure measurements drawn from the set of all real numbers ($\mathcal{X} = \mathcal{R}$).

In order to talk about communication, it is necessary to talk about noise. In order to talk about noise, it is necessary to talk about randomness. Most channels are characterized by random errors, such that the received signal $Y = [y_1, \ldots, y_{|Y|}]$ is related to the transmitted signal according to some probability distribution $P(Y|X)$. Clearly, the ideal channel is the one that introduces no errors, i.e., $P(X = Y|Y) = 1$ for all Y. Real-world communication channels are almost never so simple. Communication theory defines the *conditional entropy* of a channel, $H(X|Y)$, to be a measure of the degree to which the probability measure $P(X|Y)$ differs from error-free. Conditional entropy is always non-negative ($H(X|Y) \geq 0$), and conditional entropy is zero only if the channel never causes an error.

When a channel is noisy, it is sometimes possible to improve communication by simplifying the language. Everybody has experienced "mosh pit simplification:" when you are trying to communicate with somebody in an extremely noisy environment (e.g., a mosh pit), it helps to use only short, common, predictable words. In the extreme case, it is possible to make any channel error-free by limiting the language to just one possible utterance: no matter what the talker says, the listener always knows that the intended utterance was "hello," because protocol dictates that no other utterance is possible in the given context. A one-word language is characterized by the probability distribution $P(X = X_0) = 1$, where X_0 is the unique allowable message. The utility of a one-word language, unfortunately, is rather limited. Communication theory defines the *entropy* of a language, $H(X)$, to be a measure of the difference between $P(X)$ and the PDF of a one-word language.

The goal of effective communication, usually, is to adjust the language so that transmitted messages X are both interesting and intelligible. The balance between "interesting" and "intelligible" is measured by the *mutual information* between a transmitted message, X, and the corresponding received message, Y, defined as

$$I(X, Y) = H(X) - H(X|Y) \tag{6.4}$$

Mutual information can be defined in such a way that $I(X, Y) \geq 0$, and $I(X, Y) = 0$ only if knowing the received message (Y) provides one with no information about the content of the transmitted message (X).

The remainder of this chapter will describe entropy, conditional entropy, mutual information, and channel capacity (the maximum mutual information achievable by a particular channel). The first two sections will describe the entropy and other properties of messages, languages, and information sources. The third and fourth sections will introduce noise, and will more carefully define conditional entropy, mutual information, and channel capacity. The fifth section will more carefully discuss the relationship between sampling rate and mutual information across a continuous channel.

6.1 Discrete Sources

Printed books are an example of a discrete information source. Generally, a discrete information source will mean a system that selects choices from a finite set of elementary symbols $\mathcal{A} = \{a, b, c, \ldots\}$. The symbols are assumed to be ordered in time or in space, so that it makes sense to talk about the first symbol ($a(1) \in \mathcal{A}$), the second symbol ($a(2) \in \mathcal{A}$), the tth symbol ($a(t) \in \mathcal{A}$), and so on.

Suppose that the book is written in a European language, so that all of its characters may be written using the Latin-1 character set. Latin-1 is an 8-bit code: each character in the book (including letters, numbers, punctuation, and control characters) is encoded using exactly 8 bits. If a particular printing of the book is able to represent n characters per page, then it makes sense to say that this book generates information at a rate of $8n$ bits per page. A text file containing 100,000 characters always contains 800,000 bits.

Now suppose, instead, that the book contains characters from the international phonetic alphabet ([IPA]), as well as a few characters from other non-Latin scripts. The standard method for encoding non-Latin characters is unicode. Unicode is a mixed-length standard. Characters in the standard Latin alphabet, Arabic numerals, and a few other characters are coded using 8-bit code words, but characters from most other alphabets are coded using 16-bit symbols. A page with n characters contains somewhere between $8n$ and $16n$ bits of information, depending on the proportion of Latin characters on the page. In order to describe the average bit rate of the language (in bits/page), we need to know what percentage of the characters are in the Latin alphabet, and what percentage are symbols from IPA or other alphabets.

Both Latin-1 and unicode are inefficient encodings, because they fail to take into account the different frequencies of different letters in the alphabet. For example, recall the phoneme probabilities listed in Table 1.2. According to the table, the word "in" (ɪn) occurs with probability P(ɪ)P(n)= $0.0853 \times 0.0724 = .00618$, i.e., 0.6% of all two-phoneme sequences in the English language contain the word "in." At the other extreme, the word "joy" (dʒɔɪ) occurs with probability $0.0044 \times 0.0009 = 4 \times 10^{-7}$.

But this account is also simplistic; the word "joy" is less common than the word "in," but its frequency is not so low as 4×10^{-7}. A much better approximation can be obtained if we suppose that successive phonemes are not chosen independently but their probabilities depend on preceding phonemes. In the simplest model of this type, each phone depends only on the immediately preceding phone, and not on the ones before that. The probability of hearing phone j following phone i can thus be written as a conditional bigram probability $p(j|i)$:

$$p(j|i) = \frac{p(i,j)}{p(i)} = \frac{p(i,j)}{\sum_j p(i,j)} \tag{6.5}$$

For example, by counting occurrences in the XXX corpus, we find that the bigram probability P(ɔɪ—dʒ)= XXX. The probability of any particular two-phone sequence being the word “joy” is therefore P(dʒ)P(ɔɪ|dʒ)= $0.0044 \times XXXX = XXXX$. Perhaps there is a little more “joy” in the world than we first supposed.

This process can be extended to the estimation of trigram probabilities $p(i, j, k)$, 4-gram probabilities $p(i, j, k, m)$, 5-gram probabilities $p(i, j, k, m, n)$, and so on. In general, the probability of any particular T-phone sequence, $A = a_1, a_2, \ldots, a_T$ is given by

$$p(A) = p(a_1)p(a_2|a_1)\ldots p(a_T|a_1, a_2, a_3, \ldots, a_{T-1}), \tag{6.6}$$

where the last term is the T-gram conditional probability of a_T given its context.

To give a visual idea of how this series of processes approaches a natural human language, we can play a game invented by Shannon (Shannon and Weaver [1949]), and often called the “Shannon game.” The probabilities of Latin-1 character N-grams ranging in length from $N = 1$ to $N = 12$ were estimated from the Wall Street Journal corpus of American English. The following approximations to the English language were then generated at random, by choosing each character, a_t, using its N-gram conditional probability:

1. Zero-order approximation (symbols independent and equi-probable).
2. Unigram (1-gram) approximation (symbols chosen independently according to $p(a_t)$):
3. Bigram approximation (symbols chosen according to $p(a_t|a_{t-1})$):
4. Trigram approximation:
5. 6-gram approximation:
6. 12-gram approximation:

Chapter 7

The Ear and Hearing

Speech is a critical communication link for humans. Our auditory perceptual abilities determine the channel capacity necessary for the coding and transmission of our voice signal. It is pertinent therefore to inquire into the fundamental mechanism of hearing and to quantify the capabilities and limitations of human perception. This work began with the greeks, but was only fully developed with the invention of the telephone, and electronics Fletcher [1953].

As suggested earlier, speech information–originating from a speaker, traversing a transmission medium and arriving at a listener–might be considered at a number of stages of coding. On the transmitter side, the stages might include the acoustic wave, the muscular forces manipulating the vocal mechanism, or the physical shape and excitation of the tract. On the receiver side, the information might be considered in terms of the acoustic-mechanical motions of the hearing transducer, or in terms of the electrical pulses transmitted to the brain over the auditory nerve. Characteristics of one or more of these codings might have application in practicable transmission systems.

The previous chapter set forth fundamental relations between the acoustics and the physiology of the vocal mechanism. We will subsequently have occasion to apply the results to analysis-synthesis telephony. In the present chapter we wish to establish similar relations for the ear. Later we will utilize these in discussions of auditory discrimination and speech perception.

7.1 Mechanism of the Ear

The acousto-mechanical operation of the peripheral ear (pina, middle ear and cochlea) has been put on a rather firm base. This knowledge is, in part, due to the many experiments carried out by G. von Békésy, and for which he was awarded the Nobel Prize in 1961. However our present knowledge is the result of the last 20-30 years of detailed cochlear research, detailing the inner-ear's processes for converting mechanical motion into neural activity. Less is known about the transmission of neural information to the brain, and the ultimate mechanism of perception by the auditory cortex.

It is possible to quantify certain aspects of perception without such detail knowledge. Subjective (psychophical) behavior, in response to prescribed auditory stimuli, can of course be observed and measured, and such data are useful guideposts in the design of speech communication systems. In some instances the correlations between psychophysical behavior and the physiological operation of the peripheral ear can be placed in clear evidence. The present discussion aims to indicate current understanding of auditory physiology and psychoacoustic behavior, and to illustrate the extent to which the two can be brought into harmony.

The primary acoustic transducer of the human is shown schematically in Fig. 7.1. The acousto-mechanical components of the organ are conventionally divided according to three regions, namely, the outer ear, the middle ear, and the inner ear.

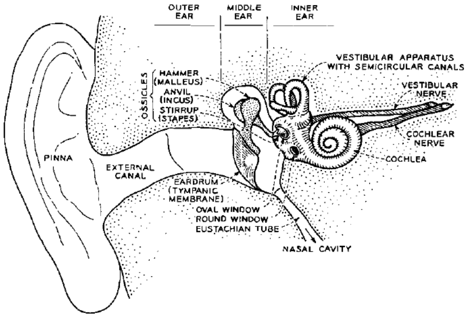


Figure 7.1: Schematic diagram of the human ear showing outer, middle and inner regions with the inner and middle ear structures shown enlarged.

7.1.1 The Outer Ear

As commonly understood, the term *ear* usually applies to the *pinna* (Fig. 7.1) which surrounds the *external canal*. Its main function in humans appears to act as a horn, increasing the collection area of the canal, and reducing the radiation reactance due to its increased area. Its directional characteristics at high audible frequencies are widely believed to facilitate localization of sound sources.

The human external canal (*meatus*) is about 2.7 cm in length and about 0.7 cm in diameter. Its volume is on the order of 1 cm^3, and its cross-section is oval-to-circular in shape with an area 0.3 to 0.5 cm^2 (Bekesy [1960]). The meatus is terminated by a thin membrane, the eardrum, or *tympanic membrane* (TM). The TM has the form of a stiff, inwardly-directed cone with an included angle of about $135\circ$ and a surface area of about 0.8 cm^2. To a rough approximation, the meatus is a uniform pipe–open at one end and terminated at the other in the *acoustic impedance* of the TM. The first mode of a quarter-wavelength pipe closed at one end, and open at the other,

work o this

falls at $f \approx c/4(2.7) \approx 3000$Hz. Such a resonance would be expected to aid the ear's sensitivity in this frequency range. Measurements do in fact show that it provides a sound pressure increase at the ear drum of between 5 and 10 db over the value at the canal entrance (Wiener and Ross [1946]). However, because the impedance of the TM is matched to the characteristic impedance of the canal, exactly how this resonance comes about is still uncertain **??**.

7.1.2 The Middle Ear

Behind the TM is the air-filled *middle-ear cavity* containg the three ossicular bones. Like the pinna and the TM, the function of the ossicles is mainly one of impedance transformation from the air medium of the outer ear to the liquid medium of the inner ear[1]. The malleus, or hammer, is fixed to and rests on the eardrum. It makes contact with the incus, or anvil, which in turn connects via a small joint to the stapes, or stirrup. The footplate of the stirrup seats in a port, the oval window, and is retained there by an annular ligament. The oval window is the entrance to the inner ear.

A sound wave impinging on the pinna is led down the external meatus and sets the TM into vibration. The vibration is transmitted via the three ossicular bones into the inner ear. The acousto-mechanical impedance of the inner ear is much greater than that of air, and for efficient transmission of sound energy an impedance transformation (a step up) is required. The ossicles provide such. First their lever action alone provides a force amplification of about 1.3 (Bekesy [1960]). That is, a force applied to the hammer appears at the stirrup footplate multiplied by 1.3. Second, the effective area of the eardrum is much greater than that of the stirrup, so that the ratio of pressure applied at the stirrup to that applied at the eardrum is essentially 1.3 times the ratio of the effective areas of drum and stirrup. Bekesy has measured this pressure transformation and finds it to be on the order of 15:1.

The middle ear structure serves another important purpose, namely, it provides protection against loud sounds which may damage the more delicate inner ear. The protective function is generally assumed to be served by two tympanic muscles– especially the tensor-tympani which connects the middle of the eardrum to the inner region of the head. Reflex contractions presumably attenuate the vibratory amplitude of the drum. Békésy points out, however, that voluntary contractions of the tensor and changes in the static pressure of the meatus only slightly reduce the vibrational amplitude of the drum. The contractions consequently can have only small effect in protecting against sound pressures that extend over a wide range of magnitudes. This fact can be established from measurements of the acoustic impedance at the drum.

In detailed studies on the mode of vibration of the ossicles, Békésy observed that at low and moderate sound intensities the stapes motion is principally a rotation

---

[1] This impedance transformation is important to the basic role of the middle ear; that is, the conversion of an external sound pressure into a fluid volume displacement in the inner ear (see Sec. 7.1.3).

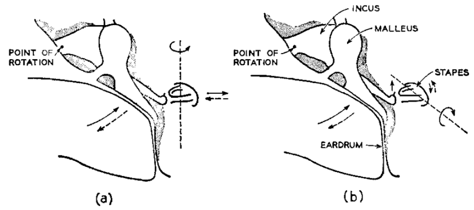


Figure 7.2: Vibration modes of the ossicles. (a) sound intensities below threshold of feeling (b) intensities above threshold of feeling. (After (Bekesy [1960]))

about an axis through the open "hoop" of the stirrup. The movement is illustrated in Fig. 7.2a. At sound intensities near and above the threshold of feeling, the motion of the stapes changes more to a rotation about an axis running longitudinally through the "arch" of the stapes, as shown in Fig. 4.2b. In the latter mode, the effective volume displacement is small because the upper-half of the rootplate advances by about as much as the lower half recedes.

Contraction of the middle ear muscles increases with sound intensity, so that the ossicles are prevented from bouncing out of contact and causing excessive distortion at the high levels. This control of distortion over the amplitude range from threshold-of-hearing to near thresholdof-feeling-while at the same time protecting the inner ear from harmful vibrational levels-apparently accounts for the elaborate middle-ear structure[2].

One of the important characteristics of the middle ear is its transmission as a function of frequency, that is, the volume displacement of the stapes footplate produced by a given sound pressure at the eardrum. A number of efforts have been made to measure or to deduce this characteristic (Bekesy [1960], Zwislocki [1957, 1959], Müller [1961, 1962]). The results are somewhat disparate, suggesting that not only is the characteristic a function of intensity in the living human, but that it may vary substantially from individual to individual.

If the fluid of the inner ear is considered incompressible and the walls of the cochlea rigid, then the volume displacement of the round window must be the same as that of the stapes footplate. At low frequencies the combined elasticity of the drum, ossicles and round window membrane controls the stirrup motion. That is, the system acts like a spring, with the stapes displacement proportional to, and in phase with, the eardrum pressure. Somewhere between about 1000 and 3000Hz the mass reactance of the system becomes important, and the motion passes from a stiffness-controlled vibration to a viscous-controlled one and finally to a mass-controlled motion. For

---

[2] One can appreciate the difficulties posed in duplicating this mechanical linkage with prosthetic devices. For example, one middle-ear prosthesis involves replacing damaged or diseased ossicles by a plastic strut joining the drum and the stapes footplate. The protection against distortion and high-amplitude vibration, normally provided by the middle ear, are difficult to include in such a construction.

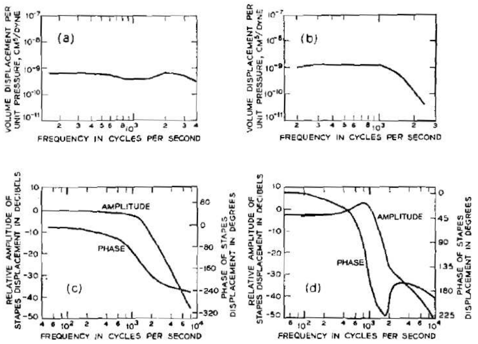


Figure 7.3: Data on middle ear transmission; effective stapes displacement for a constant sound pressure at the eardrum. (a) BÉKÉSY (1960) (one determination); (b) BÉKÉSY (1960) (another determination); (c) measured from an electrical analog circuit (after ZWISLOCKI, 1959); (d) measured from an electrical analog circuit (after (Müller [1961])).

a given sound pressure at the drum, the stirrup displacement begins to diminish in amplitude and lag in phase as frequency increases.

Békésy (Bekesy [1960]) has made a number of measurements of middle-ear transmission by directly observing the volume displacement of the round window. The transmission properties can also be deduced from a knowledge of the middle-ear topology, the input mechanical impedance to the inner ear, and the acoustic impedance at the eardrum. This approach has been used by Zwislocki (Zwislocki [1957, 1959]) and by Moller (Müller [1961]) to develop analog circuits of the middle ear. All these results agree in gross aspects but suggest that substantial variability can exist in the characteristic. By way of comparison, the transmission of the middle ear according to several determinations is shown in Fig. 7.3a-d.

For the data in Fig. 7.3b, Békésy obtains a critical "roll-off" frequency for middle-ear transmission of about 800 Hz. For the data in Fig. 7.3a, it is clearly higher, possibly around 3000Hz. Zwislocki's result in Fig. 7.3c places it somewhere in the vicinity of 1500Hz, and Moller's result in Fig. 7.3d is near 1000Hz. The common indication is that the middle-ear transmission has a low-pass characteristic. The effective cut-off frequency and the skirt slope are apparently subject to considerable variation.

7.1.3 The Inner Ear

As illustrated in Fig. 7.1, the inner ear is composed of the cochlea (normally coiled like a snail shell in a flat spiral of two and one-half turns), the vestibular apparatus and the auditory nerve terminations. It is in the cochlea that auditory mechanical-to-neural transduction takes place. The vestibular components (semi-circular canals, saccule and utricle) serve the sense of spatial orientation and apparently are not normally

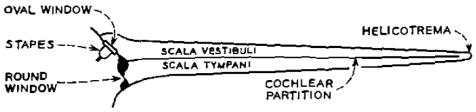


Figure 7.4: Simplified diagram of the cochlea uncoiled

used for detecting audio vibrations.

If the cochlea is uncoiled and stretched out, it appears schematically as in Fig. 7.4. The cochlear chamber is filled with a colorless liquid, perilymph, which has a viscosity about twice that of water and a specific gravity of about 1.03. The length of the canal in the spiral conch is about 35 mm. The cross-sectional area at the stirrup end is about 4 mm^2 and the area decreases to about 1 mm^2 at the tip.

The cochlear chamber is divided along almost its whole length by a partition. The half which receives the stapes is called the scala vestibuli; the other half is the scala tympani. The cochlear partition is itself a channel–the scala media–bounded partly by a bony shelf, a gelatinous membrane called the basilar membrane, and another membrane known as Reissner's membrane. The partition is filled with a different liquid, the endolymph. The basilar membrane and bony shelf both terminate a mm or two short of the ends of the scalas, permitting them to communicate at the helicotrema. The area of the connecting passage is about 0.3 to 0.4 mm^2(Békésy and Rosenblith). The basilar membrane is about 32 mm in length and tapers from a width of about 0.05 mm at the base (stirrup) to about 0.5 mm at the apex (Davis [1951]).

The inner ear is connected to the middle ear at the stapes footplate. The latter, supported by a ring-shaped ligament, seats into the oval window (about 3 mm^2 in area). In vibrating, the stapes acts as a piston and produces a volume displacement of the cochlear fluid. Because the cochlea is essentially rigid and its fluid incompressible, fluid displacements caused by inward motion of the stapes must be relieved. This function is accomplished at the round window which is covered by a compliant membrane (about 2 mm^2). Very slow vibrations of the stapes (say less than 20Hz) result in a to-and-fro flow of fluid between the scala vestibuli and scala tympani through the opening at the helicotrema. Higher frequency vibrations are transmitted through the yielding cochlear partition at a point which depends upon the frequency content of the stimulation.

A cross-section of the cochlea and its partition is shown in Fig. 7.5. The main functions and dynamical properties of the partition reside in the basilar membrane. It is upon the latter that the organ of Corti rests. Among several types of supporting cells, the organ of Corti contains some 30000 sensory cells (or hair cells), on which the endings of the auditory nerve (entering from the lower left in Fig. 7.5) terminate. The basilar membrane is stiffer and less massive at its narrow, basal end and more compliant and massive at the broad, apical end. Its resonant properties therefore vary continuously along its length. At low frequencies, Reissner's membrane normally

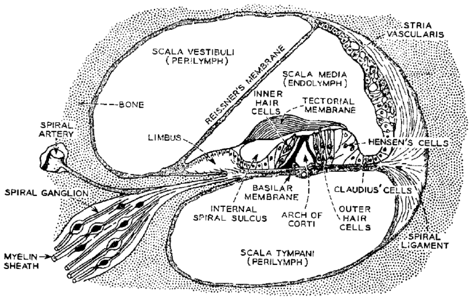


Figure 7.5: Schematic cross section of the cochlear canal. (Adapted from Davis (Davis [1957]))

moves cophasically with the basilar membrane.

Current knowledge of the acoustic-mechanical properties of the basilar membrane is due almost exclusively to the efforts of von Békésy. In physiological preparations, he vibrated the stapes footplate sinusoidally and measured the amplitude and phase of the membrane displacements along the length of the cochlea. The mechanical characteristics of the basilar membrane, as determined in these experiments, are shown in Fig. 7.6. Figs. 7.6a and b show the amplitude and phase of specific membrane points as functions of frequency. Fig. 7.6c shows the amplitude and phase as afunction of membrane place with frequency as the parameter.

The amplitude and phase response of a given membrane point is much like that of a relatively broad band-pass filter. The amplitude responses of successive points are roughly constant-Q in nature. Because of this constant percentage bandwidth property, the frequency resolution is best at the low-frequency (apical) end of the membrane, and the time resolution is best at the higher-frequency (basal) end[3].

All the amplitude responses of Fig. 7.6 are normalized to unity. Békésy's measurements suggest, however, that for constant amplitude of stapes displacement, the peak membrane response increases at about 5 db/octave for points resonant at frequencies up to about 1000Hz, and is approximately constant in peak displacement for points resonant at higher frequencies. Linear increments of distance along the basilar membrane correspond approximately to logarithmic increments of peak frequency, at least for frequencies less than about 1000Hz.

Excitation at the stapes is propagated down the membrane in the form of a travelling wave of displacement. Because of the taper of the distributed constants with distance, essentially no reflection takes place at the helicotrema, and no standing wave of displacement is created. The membrane is a dispersive transmission medium. The travelling wave loses more and more of its high frequency components as it progresses

---

[3] Recent measurements of basilar membrane vibration in animals, using the Mossbauer effect (Johnstone and Boyle [1967], Rhode [1971]), suggest that the mechanical response is sharper (higher in Q) than shown in Fig. 7.6. Also, the measurements suggest that the mechanical response is somewhat dependent upon sound intensity.

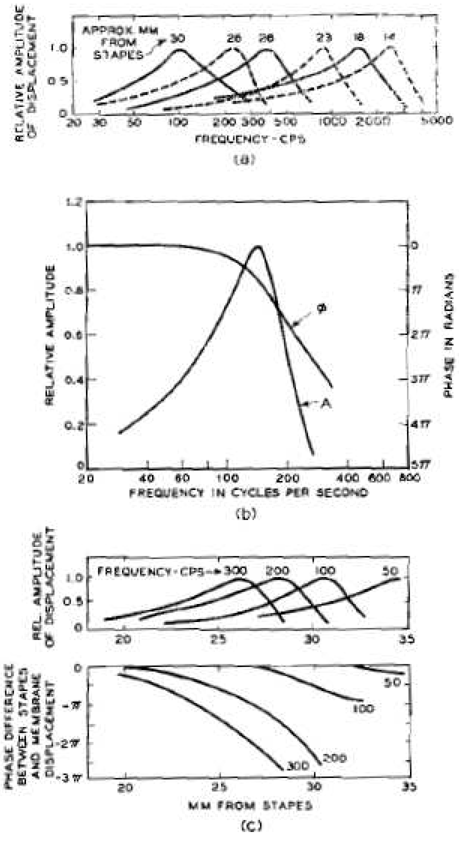


Figure 7.6: Amplitude and phase responses for basilar membrane displacement. The stapes is driven sinusoidally with constant amplitude of displacement. (After (Bekesy [1960]).) (a) Amplitude vs frequency responses for successive points along the membrane. (b) Amplitude and phase responses for the membrane place maximally responsive to 150 Hz. (c) Amplitude and phase of membrane displacement as a function of distance along the membrane. Frequency is the parameter

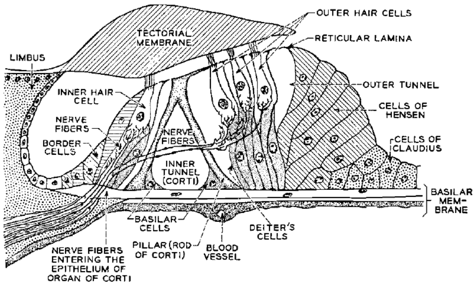


Figure 7.7: Cross section of the organ of Corti. (After (Davis [1951]))

toward the helicotrema, and its group delay increases.

7.1.4 Mechanical-to-Neural Transduction

Mechanical motion of the membrane is converted into neural activity in the organ of Corti. An enlarged view of this structure is shown in Fig. 7.7. The organ of Corti contains a collection of cells among which are the hair cells. The hairs emanating from these sensory cells protrude upward through the reticular lamina and contact a third membrane of the cochlear partition, the tectorial membrane. One set of cells lies in a single row, longitudinal to the basilar membrane and toward the axis of the cochlear spiral (left of the arch of Corti). They are termed the inner hair cells. Another set lies in three or four longitudinal rows, radially away from the center of the spiral. These are the outer hair cells. Estimates fix the number of the former at about 5000 and the latter at about 25000.

The tectorial and basilar membranes are anchored at their inner edges at spatially separate points. A deformation of the basilar membrane causes relative motion between the tectorial membrane and the reticular lamina and a resultant stress on the hairs passing between. By a process that presently is not understood, a bending of the hairs produces an electrical discharge in the cochlear portion of the VIIIth nerve[4]. The first-order fibers of this cochlear branch, or auditory nerve, enter from the lower left in Fig. 7.7 and run to the inner and outer hair cells.

Electrophysiological experiments suggest that the outer and inner hair cells of the organ of Corti differ in their sensitivities to mechanical stimulation(Bekesy [1953], Davis [1958]). The outer hair cells appear to be sensitive to bending only in a direction transverse to the long dimension of the membrane. Moreover, only outward bending of the hairs (away from the arch of Corti) produces an electrical potential in the scala media favorable for exciting the auditory nerve endings. This outward bending is produced on (unipolar) upward motions of the basilar membrane, that is, motions which drive it toward the tectorial membrane.

---

[4] The VIIIth nerve also serves the vestibular apparatus. See Fig. 7.1.

The inner hair cells, on the other hand–residing between the arch of Corti and the axis of the cochlear spiral–appear sensitive to bending in a direction parallel to the long dimension of the membrane (Bekesy [1953], Davis [1958]). In this case bending only toward the apex of the cochlea produces a scala media potential favorable for stimulating the nerve. So far as a given point on the membrane is concerned, the inner hair cells are essentially sensitive to the longitudinal gradient of displacement, that is, to the spatial derivative in the long dimension. Furthermore, the inner cells fire only on the polarity of the gradient which corresponds to bending toward the apex. The threshold for firing of the inner cells appears to be appreciably higher than that for the outer cells. Exactly how the pattern of mechanical displacement of the basilar membrane is reflected in the "transducer" potentials of the sensory cells and in the electrical pulses elicited in the auditory nerve has yet to be put on a firm basis.

The sensory cells of the ear connect to the brain via the bundle of nerve cells–or neurons–comprising the auditory nerve. The auditory nerve passes down the axis of the cochlear spiral–collecting more nerve fibers as it runs from apex to base–until it contains some 30000 neurons. Neurons presumably have only two states, namely, active or inactive. When excited by an electrical input above a particular threshold, they produce a standard electrical pulse of about a millisecond duration and are desensitized for a period of about one to three milliseconds thereafter. They consequently can be excited to maximum discharge rates on the order of 300 to 1000Hz.

The connections between the nerve cells and the hair cells in the organ of Corti are complex. Each inner hair cell is innervated by one or two nerve fibers, and each fiber connects with one or two hair cells. Innervation of the outer cells is more compound. Most nerve fibers make connections with a number of outer cells, and each outer cell usually receives connections from several nerve fibers (Davis [1957]). The exact functional significance of this complex multiple distribution of the nerve supply is not presently known. One study has suggested that it contributes to the great intensity range of the ear (van Bergeijk [1961]).

The fibers of the auditory nerve twist like strands of rope about a central core. The nerve itself is short and enters the lower brain stem (medulla oblongata) after a run of about 5mm (Davis [1957]). The incoming fibers divide, and the branches run respectively to the dorsal and to the vertral portions of the cochlear nucleus. Here the first synapses (junctions connecting one nerve cell to another) of the auditory system reside. The fibers of the auditory nerve, and the cells of the cochlear nucleus to which they join, essentially preserve the orderly arrangement of the corresponding sensory cells on the basilar membrane. The same general tendency toward orderly arrangement, with respect to membrane place of origin, seems to be maintained throughout the auditory system.

Relatively little is known about the mechanism by which the basilar membrane displacements are converted into neural activity. Still less is known about how information is coded in nerve pulses and assimilated into an auditory percept by the brain. Qualitatively, however, several deductions seem to be generally accepted. First, the hairs of the sensory cells, in experiencing a lateral shear owing to relative motion of basilar membrane and tectorial membrane (see Fig. 7.7), generate local electrical potentials which represent the local basilar membrane displacement. More precisely, the shearing forces on the sensory hairs "modulate" (as would a variable resistor) a current passing between the scala media and the base of the hair cell (Davis [1965]).

Second, this facsimile alternating potential, acting at the base of the hair cell, modulates the liberation of a chemical mediator about some quiescent rate. The mediator, in sufficient quantity, stimulates the dendritic endings of the first-order nerve fibers and causes the fibers to fire. Because of its quiescent bias rate, the hypothesized chemical mediator is secreted more on one phase of the sensory potential than on the other; that is, a rectifying function is implied in triggering the nerve fiber.

Lastly, the chemical stimulation of the nerve endings produces an all-or-none electrical firing, which is propagated axonally to subsequent higher-order fibers in the central nervous system.

There are two basic electrical phenomena in the cochlea: the resting (dc) polarization of the parts, and the ac output of the organ of Corti (which, as stated, appears to reflect the displacement of the cochlear partition). Current thinking holds that the ac output, or the cochlear microphonic[5], is a potential produced by the sensory or receptor cells and is derived from the pre-existing resting polarization of the receptor cells by a change in the ohmic resistance of a mechanically-sensitive portion of the cell membrane. "This change in resistance is presumably brought about by a deformation, however slight, of a critical bit of the polarized surface" (Davis [1965]).

Energy considerations argue for an active (power-amplifying) type of transduction, as in a carbon microphone. The biasing current (power supply) for the transducer is produced by the biological battery which is the resting polarization of the hair cell. The mechanical energy of the basilar membrane is not transduced into electrical current, rather it controls or modulates the flow of current across the interface (cell membrane) which separates the negative polarization inside the hair cell from the positive, endo-cochlear potential of the endolymph inside the cochlear partition.

A map of the cochlear resting potentials is shown in Fig. 7.8. The scala tympani is taken as the zero reference potential, and regions of similar potential are often found within the organ of Corti. Other areas, presumably intracellular, are strongly negative. The endolymphatic space (scala media) is strongly positive. (Refer to Fig. 7.5 for more details on the organ of Corti.)

If a microelectrode penetrates upward through the basilar membrane, recording simultaneously the de potentials (which serve to locate the electrode tip) and the cochlear microphonic response to a 500Hz tone, the result is shown in Fig. 7.9. The conclusion is that the electrical interface at which the phase reversal of the cochlear

---

[5] This potential is typically observed by an electrode placed at the round window or inserted into a scala.

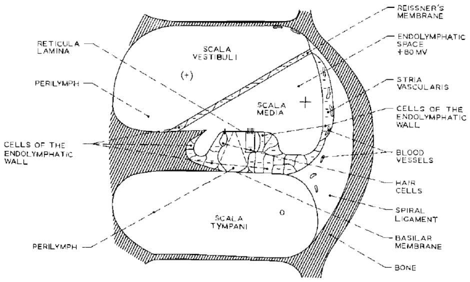


Figure 7.8: Distribution of resting potentials in the cochlea. Scala tympani is taken as the zero reference. The tectorial membrane is not shown. The interiors of all cells are strongly negative. (After (Tasaki *et al.* [1954]))

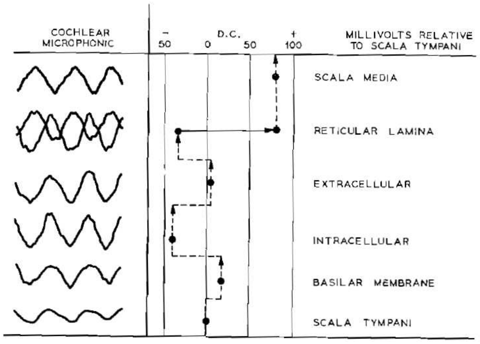


Figure 7.9: Cochlear microphonic and dc potentials recorded by a microelectrode penetrating the organ of Corti from the scala tympani side. The cochlear microphonic is in response to a 500Hz tone. (After (Davis [1965]))

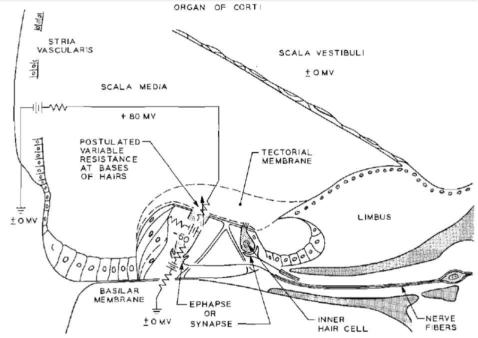


Figure 7.10: A "resistance microphone" theory of cochlear transduction. (After DAVIS, 1965)

microphonic occurs is the hair-bearing surface of the hair cell (although one cannot differentiate between the surface and base location of the hair cell).

Two biological batteries therefore act in series: the internal (negative) polarization of the hair cells and the (positive) dc polarization of the endocochlear voltage (which is probably generated by the stria vascularis). This action leads to the conception of the equivalent circuit for cochlear excitation shown in Fig. 7.10.

The cochlear microphonic, as already mentioned, is viewed as the fluctuating voltage drop across the cell membrane due to alternating increase or decrease in its ohmic resistance. It appears to be a facsimile representation of the local displacement of the basilar membrane (TEAS et al.). The dynamic range of the microphonic is relatively large. Its amplitude appears linearly related to input sound pressure over ranges greater than 40 dB (Teas *et al.* [1962]).

Although the functional link between the cochlear microphonic (or the facsimile membrane displacement) and the all-or-none electrical activity in the fibers of the auditory nerve remains obscure, it is nevertheless clear that a local deformation of the membrane (of sufficient amplitude), and a consequent bending of the sensory hairs in the area, causes the sensory cells to generate a scala media potential favorable for triggering the neurons in that region. The greater the displacement magnitude, the greater the number of neurons activated. A periodic displacement of sufficiently low frequency elicits neural firing synchronous with the stimulus. The periodicity of tones of frequencies less than about 1000Hz may therefore be represented by the periodicity of the neural volleys. This mechanism may be one method of coding for the subjective attribute of pitch. The fact that the neurons leading away from a given region of the frequency-selective basilar membrane maintain their identity in the auditory nerve offers a further possibility for the coding of pitch, namely, in terms of membrane place of maximum stimulation.

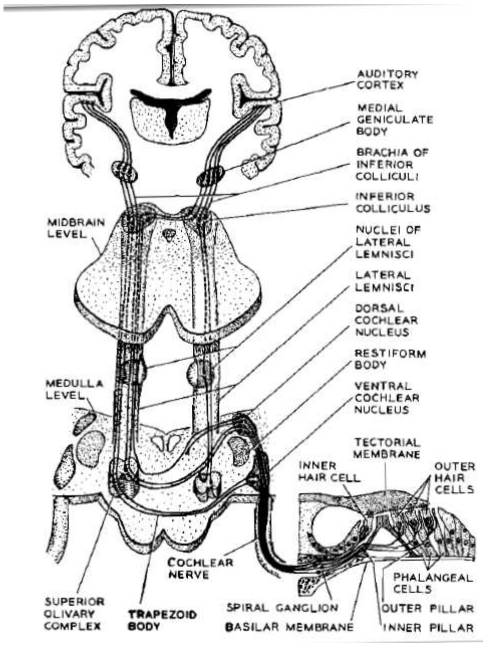


Figure 7.11: Schematic diagram of the ascending auditory pathways. (Adapted from (Netter [1962]))

7.1.5 Neural Pathways in the Auditory System

A schematic representation of the ascending neural pathways associated with one ear are shown in Fig. 7.11. Beginning in the organ of Corti, the roughly 30000 individual neurons innervate singly or multiply about the same number of sensory (hair) cells. (In general, the inner hair cells are served by only one or two neurons, the outer cells by several.) The dendritic arbors of the first-order neurons bear on the sensory cells. The cell bodies of the first-order neurons are located in the spiral ganglion, and their axons pass via the cochlear nerve (about 5 mm) to the dorsal and ventral cochlear nuclei in the medulla. Here the first synapses of the auditory pathway are located. From these nuclei, some second-order neurons pass to the superior olive on the same side, some decussate to the opposite side. Some pass upward to the medial geniculate body, with or without intermediate synapses with other neurons located in the lateral lemnisci and the inferior colliculi. The latter nuclei are located in the midbrain, and a second, smaller pathway of decussation runs between them. Thus, stimuli received at the two ears may interact both at the medulla and midbrain levels. The last stage in the pathway is the auditory cortex. The exact neuro-electrical representation of sound stimuli at these various levels is not well understood, and considerable research effort is presently aimed at studying these processes.

The first-order fibers of the auditory nerve connect to different places along the cochlear partition. Starting at the point (apex) of the cochlea, they are progressively collected in the internal auditory meatus until, at the base, the whole nerve trunk is

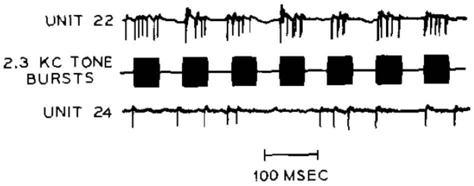


Figure 7.12: Electrical firings from two auditory nerve fibers. The characteristic frequency of unit 22 is 2.3 kHz and that for unit 24 is 6.6 kHz, The stimulus is 50 msec bursts of a 2.3 kHz tone. (After (Kiang and Peake [1960]))

formed. Because the basilar membrane is a mechanical frequency analyzer, it is not surprising that individual fibers exhibit frequency specificity. Owing to the way in which the fibers are collected and the trunk formed, those fibers which have greatest sensitivity to high frequencies lie on the outer part of the whole nerve, while those more sensitive to low frequencies tend to be found toward the core. This "tonotopic" organization of the auditory system (that is, its place-frequency preserving aspect) seems to be maintained at least to some degree all the way to the cortical level (Tunturi [1955]).

The electrical response of individual fibers is a standard pulse. Characteristically, the pulse exhibits a duration on the order of a millisecond. The activity is statistical in two senses. First, the firing patterns of an individual fiber are not identical in successive repetitions of a given stimulus. Second, individual fibers exhibit spontaneous firing (electrical output) of a random nature. The latter appears to be much the same for all first-order fibers.

Comprehensive investigation of first-order electrical behavior in cats has been carried out by Kiang et al. (Kiang and Peake [1960]). Since the structure of the cochlea and auditory nerve follows the same general plan in all mammals, data from these studies should give insight into the human auditory system.

Typical microelectrode recordings from single primary fibers are illustrated in Fig. 7.12. In this instance, the signal comprises 50 msec tone bursts of a 2.3 kHz frequency. The upper recording is from a fiber that is maximally sensitive to this frequency, while the lower is from a fiber maximally sensitive to 6.6 kHz. The electrical output of the former is highly correlated with the stimulus, while the electrical output of the latter is not. The nerve response potential is recorded with respect to a convenient reference potential, in this case the head holder for the preparation. A positive voltage is indicated by a downward deflection.

By choosing a suitable criterion of response, the frequency characteristic (tuning curve) of an individual first-order fiber can be measured. Results of such measurements are illustrated for several fibers in Fig. 7.13. The frequency for which the threshold is lowest (the minimum of each curve) is called the characteristic frequency (CF) of the fiber (unit). These minima appear to match well the shape of the audiogram determined from behavioral measurements. An interesting aspect of these data

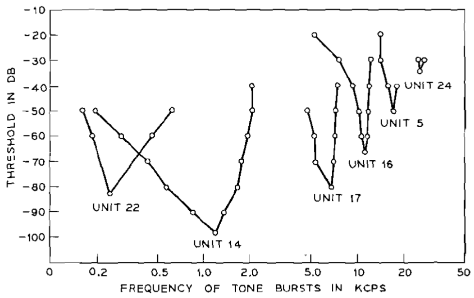


Figure 7.13: Frequency sensitivities for six different fibers in the auditory nerve of cat. (After (Kiang and Peake [1960]))

is that while over the low-frequency range the shapes of the tuning curves appear to be nearly constant percentage bandwidth in character (constant Q) and display a bandwidth which correlates reasonably well with Békésy's mechanical responses, the tuning curves of high-frequency units are much sharper and display Qincreasing with frequency. (Békésy's observations on human basilar membrane were, of course, limited to the low end of the spectrum–2400Hz and down.)

Kiang et al. have also observed the electrical response of primary units to punctuate signals, namely, broadband clicks. Individual responses to 10 successive rarefaction clicks of 100–sec duration are plotted in Fig. 7.14. The figure shows the electrical response recorded at the round window (RW, primarily the cochlear microphonic) and the response of the individual fiber. The time origin is taken as the time the pulse is delivered to the earphone. The characteristic frequency of the unit is 540Hz. The pattern of firing differs in successive stimulations, but the responses show a consistent periodic aspect. Multiple firings in response to a single click are apparent.

A convenient way to analyze this periodic feature is, in successive presentations of the signal, to measure the number of times the fiber fires at a prescribed time after the signal onset. This number plotted against the time of firing forms the post-stimulus time (PST) histogram. Some quantization of the time scale is implied and this quantization (or "bin width") is made sufficiently small to resolve the periodicities of interest. (For click signals, a bin width of 0.063 msec was customarily used.) A digital computer is a valuable tool for calculating and displaying the histogram. One minute of data from the conditions in Fig. 7.14 produces the histogram of Fig. 4.15. (Since the clicks are delivered at a rate of 10 Hz, this histogram is the result of 600 signal presentations.) The times of firings show a periodic structure, or "preferred" times for firing. In the midfrequency range for the animal, the histogram may exhibit as many as five or six distinct peaks or preferred times. At the upper end of the animal's frequency range, the tendency is for the histogram to display a single major peak.

The preferred times for firing appear to be intimately linked to the characteristic frequency of the unit, and the interval between peaks in the PST histogram is


RW
UNIT

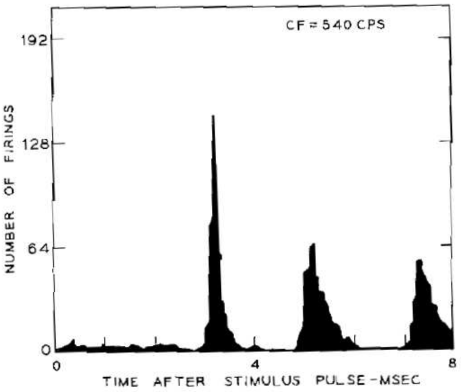


Figure 7.15: Post stimulus time (PST) histogram for the nerve fiber shown in Fig. 7.14. $CF = 540$Hz. Stimulus pulses 10 Hertz. (After (Kiang and Peake [1960]))

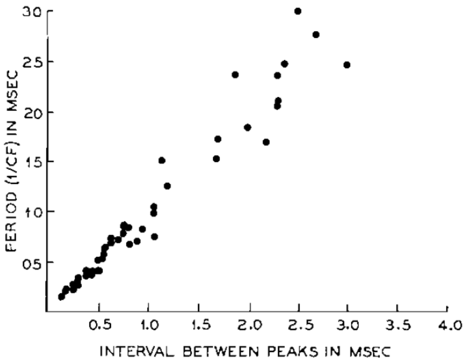


Figure 7.16: Characteristic period (l/CF) for 56 different auditory nerve fibers plotted against the interpeak interval measured from PST histograms. (After KIANG et at.)

approximately equal to l/CF. Higher frequency units consequently show smaller intervals between the PST histogram peaks. The interval between peaks in the histogram and l/CF are related as shown in Fig. 7.16. Data for 56 different units are plotted. The multiple responses of single primary units to single clicks almost certainly reflect the mechanical response of the cochlea. (See the derivations of Section 7.2 for the impulse response of the basilar membrane.)

Microelectrode studies of the electrical activity of single neurons at other levels in the auditory pathway have been, and are being, carried out. Varying experimental techniques and methods of anesthesia have sometimes led to disagreements among the results, but as research progresses the neural schema is becoming increasingly better understood.

According to at least one investigation on cat, the rate of single unit firings is monotonically related to stimulus intensity at all neural stages from the periphery up to the medial geniculate body (KATSUKI). This is exemplified for sinusoidal tones

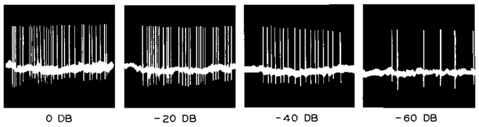


Figure 7.17: Responses of a single auditory neuron in the trapezoidal body of cat. The stimulus was tone bursts of 9000Hz produced at the indicated relative intensities. (After KATSUKI)

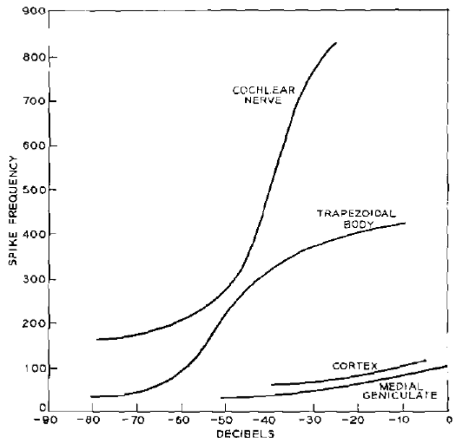


Figure 7.18: Relation between sound intensity and firing (spike) frequency for single neurons at four different neural stages in the auditory tract of cat. Characteristic frequencies of the single units: Nerve: 830Hz; Trapezoid: 9000Hz; Cortex: 3500Hz; Geniculate: 6000 Hz.(After KATSUKI)

in Figs. 7.17 and 4.18. Fig. 7.17 shows the spikes (firings) of a single neuron in the trapezoidal body of cat in response to tone bursts of 9000Hz, delivered at four different levels. The spike duration is on the order of the conventional 1 msec, and the firings are more numerous for the more intense sounds.

Fig. 7.18 shows a monotone relation between firing rate and intensity for different neural stages. The firing rate for the first-order single neuron (the top curve for the cochlear nerve) has a maximum value close to its best (characteristic) frequency, namely 830 Hz. This suggests that for the sinusoidal stimulation, the first-order neuron fires at most once per period. The rates at the higher neural stages appear substantially less than their characteristic frequencies.

Microelectrode recordings from single first-order neurons often show appreciable spontaneous activity. At higher neural stages and in the cortex, spontaneous activity apparently is not as pronounced (KATSUKI).

The cochlear nucleus complex of cat has been another particular area of study (Rose *et al.* [1959]). Strong evidence for a distinct tonotopical organization is found in the major subdivision of the cochlear nucleus. Typical of this finding is the sagittal section

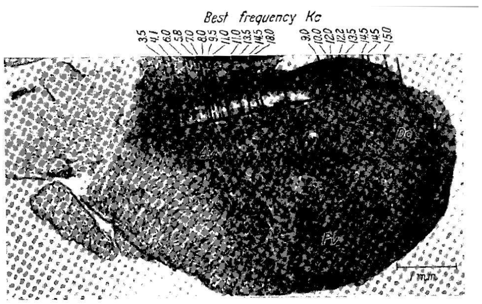


Figure 7.19: Sagittal section through theleft cochlear complex in cat. The electrode followed the track visible just above the ruled line. Frequencies of best response of neurons along thetrack are indicated. (After (Rose *et al.* [1959]))

through the left cochlear complex shown in Fig. 7.19. The frequency scale indicates the best (most sensitive) frequencies of the neurons located along the ruled axis.

Some tonotopical organization appears to exist at the cortical level, although its degree and extent seems to be controversial (for example, (Katsuki [1960], Tunturi [1955])).

The relations between threshold sound amplitude and tone frequency (that is, the tuning curves) for single units at the cochlear nucleus level have been found to vary in shape (Rose *et al.* [1959]). Some appear broad, others narrow. All, however, resemble roughly the mechanical resonance characteristic of the basilar membrane. That is, the tuning curve (or threshold amplitude) rises more steeply on the high-frequency side than on the low-frequency side. Typical narrow and broad tuning curves obtained from single units in the cochlear nucleus are shown in Fig. 7.20a and b, respectively. For tones up to about 60 db above the threshold, the frequency range of response for both narrow and broad units does not extend over more than about 0.3 of an octave above the best frequency. The frequency range below the best frequency can range from about 0.4 to 3.8 octaves for the narrow units, to almost the whole lower frequency range for the broad units. Single units at this level display adaptive and inhibitory behavior which is strongly intensity dependent.

The mechanism of neural transmission across a synapse also remains to be firmly established. A temporal delay–typically on the order of 1 msec–is usually incurred at the junction. Response latencies at the level of the cochlear nucleus have minimum times on the order of 2 to 3 msec, but latencies as great as 6 to 8 msec have been measured. At the cortical level, latencies as great as 20 to 30 msec and as small as 6 to 8 msec are possible.

7.2 Computational Models for Ear Function

It has been emphasized in the preceding discussion that the complete mechanism of auditory perception is far from being adequately understood. Even so, present

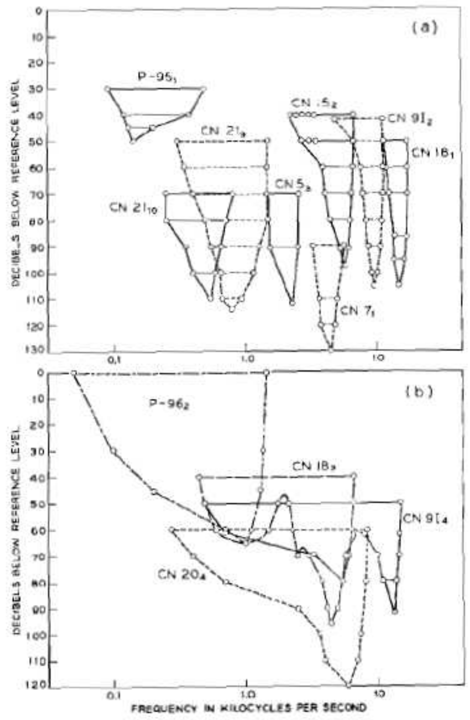


Figure 7.20: Intensity us frequency" threshold" responses for single neurons in the cochlear nucleus of cat. The different curves represent the responses of different neurons. (a) Units with narrow response areas; (b) units with broad response areas. (After (Rose *et al.* [1959]))

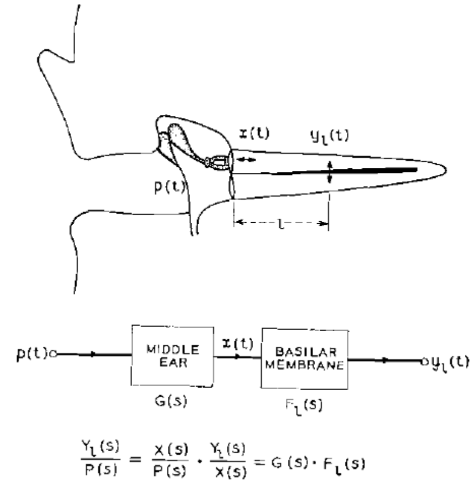


Figure 7.21: Schematic diagram of the peripheral ear. The quantities to be related analytically are the eardrum pressure, $p(t)$: the stapes displacement, $x(t)$; and the basilar membrane displacement at distance l from the stapes, $y_l(t)$

knowledge of ear physiology, nerve electrophysiology, and subjective behavior make it possible to relate certain auditory functions among these disparate realms. Such correlations are facilitated if behavior can be quantified and analytically specified. As a step in this direction, a computational model has been derived to describe basilar membrane displacement in response to an arbitrary sound pressure at the eardrum (Flanagan [1962a]).

The physiological functions embraced by the model are shown in the upper diagram of Fig. 7.21. In this simplified schematic of the peripheral ear, the cochlea is shown uncoiled. $p(t)$ is the sound pressure at the eardrum, $x(t)$ is the equivalent linear displacement of the stapes footplate, and $y_l(t)$ is the linear displacement of the basilar membrane at a distance l from the stapes. The desired objective is an analytical approximation to the relations among these quantities. It is convenient to obtain it in two steps. The first step is to approximate the middle-ear transmission, that is, the relation between $x(t)$ and $p(t)$. The second is to approximate the transmission from the stapes to the specified point l on the membrane. The approximating functions are indicated as the Laplace transforms $G(s)$ and $F_l(s)$, respectively, in the lower part of Fig. 7.21.

The functions $G(s)$ and $F_l(s)$ must be fitted to available physiological data. If the ear is assumed to be mechanically passive and linear over the frequency and amplitude ranges of interest, rational functions of frequency with stable normal modes (left half-plane poles) can be used to approximate the physiological data. Besides computational convenience, the rational functions have the advantage that they can be realized in terms of lumped-constant electrical circuits, if desired. Because the model is an input-output or "terminal" analog, the response of one point does not require explicit computation of the activity at other points. One therefore has the freedom to calculate the displacement $y_l(t)$ for as many, or for as few, values of l as are desired.

7.2.1 Basilar Membrane Model

The physiological data upon which the form of $F_l(s)$ is based are those of BÉKÉSY, shown in Fig. 7.6[6]. If the curves of Fig. 7.6 are normalized with respect to the frequency of the maximum response, one finds that they are approximately constant percentage bandwidth responses. One also finds that the phase data suggest a component which is approximately a simple delay, and whose value is inversely proportional to the frequency of peak response. That is, low frequency points on the membrane (nearer the apex) exhibit more delay than high frequency (basal) points. A more detailed discussion of these relations and the functional fitting of the data has been given previously (Flanagan [1962a]). In this earlier work, the fit afforded by three different forms of $F_l(s)$ was considered. For purpose of the present discussion, only the results for the first, a fifth-degree function, will be used.

The physiological data can, of course, be approximated as closely as desired by selecting an appropriately complex model. The present model is chosen to be a realistic compromise between computational tractability and adequacy in specifying the physiological data. One function which provides a reasonable fit to Békésy's results is

$$F_l(s) = c_1 \beta_l^4 \left(\frac{2000\pi\beta_l}{\beta_l + 2000\pi} \right)^{0.8} \left(\frac{s + \epsilon_l}{s + \beta_l} \right) \left[\frac{1}{(s + \alpha_l)^2 + \beta_l^2} \right]^2 e^{\frac{-3\pi s}{4\beta_l}} \tag{7.1}$$

---

[6] More recent data on basilar membrane vibration, determined in animal experiments using the Mossbauer effect (Johnstone and Boyle [1967], Rhode [1971]), may also serve as this point of departure.

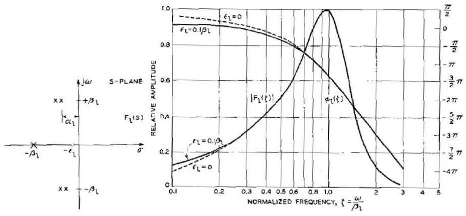


Figure 7.22: (a) Pole-zero diagram for the approximating function $F_l(s)$ (After FLANAGAN, 1962a). (b) Amplitude and phase response of the basilar membrane model $F_l(s)$. Frequency is normalized in terms of the characteristic frequency β_l

where

$s = \alpha + j\beta$	is the complex frequency,
$\beta_l = 2\alpha_l$	is the radian frequency to which the point l-distance from the stapes responds maximally,
c_1	is a real constant that gives the proper absolute value of displacement,
$e^{\frac{-3\pi s}{4\beta_l}}$	is a delay factor of $3\pi/4\beta_l$ seconds which brings the phase delay of the model into line with the phase measured on the human ear. This factor is primarily transit delay from stapes to point l on the membrane,
$\left(\frac{2000\pi\beta_l}{\beta_l+2000\pi}\right)^{0.8}\beta_l^4$	is an amplitude factor which matches the variations in peak response with resonant frequency β_l as measured physiologically by (Bekesy [1943]).
ϵ_l/β_l	= 0.1 to 0.0 depending upon the desired fit to the response at low frequencies.

The membrane response at any point is therefore approximated in terms of the poles and zeros of the rational function part of $F_l(s)$. As indicated previously in Fig. 7.6, the resonant properties of the membrane are approximately constant-Q (constant percentage bandwidth) in character. The real and imaginary parts of the critical frequencies can therefore be related by a constant factor, namely, $\beta_l = 2\alpha_l$. To within a multiplicative constant, then, the imaginary part of the pole frequency, β_l, completely describes the model and the characteristics of the membrane at a place l-distance from the stapes. The pole-zero diagram for the model is shown in Fig. 7.22a.

The real-frequency response of the model is evidenced by letting $s = j\omega$. If frequency is normalized in terms of $\zeta = \omega/\beta_l$, then relative phase and amplitude responses of $F_l(j\zeta)$ are as shown in Fig. 7.22b. Because of the previously mentioned

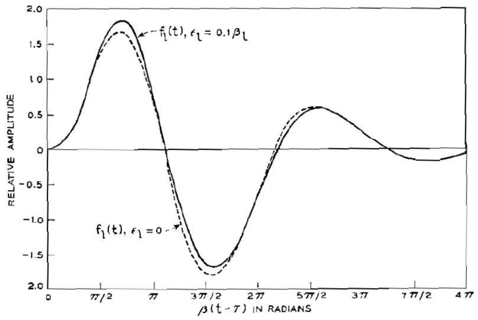


Figure 7.23: Response of the basilar membrane model to an impulse of stapes displacement

relations, $F_l(\zeta)$ has (except for the multiplicative constant) the same form for all values of l.

The inverse Laplace transform of (7.1) is the displacement response of the membrane to an impulse of displacement by the stapes. The details of the inverse transformation are numerically lengthy, but if the mathematics is followed through it is found to be

$$
\begin{aligned}
f_l(t) = c_1 \left(\frac{2000\pi}{\beta_l + 2000\pi} \right)^{0.8} & \beta_l^{1+r} \left\{ [0.033 + 0.360\beta_l(t-T)] \right. \qquad (7.2) \\
& \times e^{-\frac{\beta_l(t-T)}{2}} \sin \beta_l(t-T) + [0.575 - 0.320\beta_l(t-T)] \\
& e^{-\frac{\beta_l(t-T)}{2}} \cos \beta_l(t-T) - 0.575 e^{-\beta_l(t-T)} \left. \right\} = 0 \\
& \text{for } t \geq T \quad \text{and} \quad \epsilon_l/\beta_l = 0.1,
\end{aligned}
$$

where the delay $T = 3\pi/4\beta_l$, as previously stated. A plot of the response (7.2) is shown in Fig. 7.23.

Note, too, from the form of (7.1) that the complex displacement response can be determined as a function of the place frequency β_l for a given stimulating frequency $s = j\omega_n$ The radian frequency β_l can, in turn, be related directly to the distance l (in mm) from the stapes by

$$
(35 - l) = 7.5 \log \beta_a / 2\pi(20)
$$

(see (Flanagan [1962a])). Therefore (7.1) can be used to compute $F(s,l)|_{s=j\omega_n} = A(l)e^{j\phi(l)}$ to give spatial responses of amplitude and phase similar to those shown in Fig. 7.6c.

7.2.2 Middle Ear Transmission

To account for middle ear transmission, an analytical specification is necessary of the stapes displacement produced by a given sound pressure at the eardrum (see

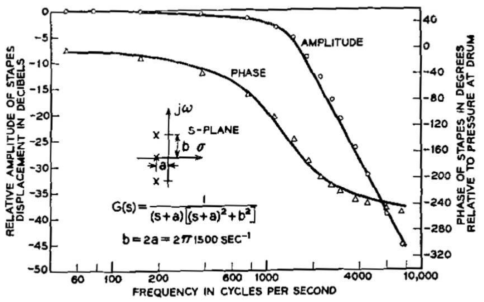


Figure 7.24: Functional approximation of middle ear transmission. The solid curves are from an electrical analog by ZWISLOCKl (see Fig. 7.3c). The plotted points are amplitude and phase values of the approximating function $G(s)$. (Flanagan [1962a])

Fig. 7.21). Quantitative physioacoustical data on the operation of the human middle ear are sparse. The data which are available are due largely to Békésy and, later, to Zwislocki and to Moller. These results have been shown in Fig. 7.3. The data suggest appreciable variability and uncertainty, particularly in connection with the critical (roll-off) frequency and damping of the characteristic. All agree, however, that the middle ear transmission is a low-pass function. Békésy's results were obtained from physiological measurements. Zwislocki's and Moller's data are from electrical analogs based upon impedance measurements at the eardrum, a knowledge of the topology of the middle ear circuit, and a knowledge of some of the circuit constants. In gross respects the data are in agreement[7].

If ZWISLOCKI'S results in Fig. 7.3 are used, they can be approximated reasonably well by a function of third degree. Such an approximating function is of the form

$$G(s) = \frac{c_0}{(s+a)\left[(s+a)^2+b^2\right]} \tag{7.3}$$

where c_0 is a positive real constant. [When combined with $F_l(s)$, the multiplying constants are chosen to yield proper absolute membrane displacement. For convenience, one might consider $c_0 = a(a^2+b^2)$ so that the low-frequency transmission of $G(s)$ is unity.] When the pole frequencies of $G(s)$ are related according to

$$b = 2a = 2\pi(l500) \text{ rad/sec}, \tag{7.4}$$

the fit to Zwislocki's data is shown by the plotted points in Fig. 7.24. The inverse transform of (7.3) is the displacement response of the stapes to an impulse of pressure at the eardrum. It is easily obtained and will be useful in the subsequent discussion. Let

$$G(S) = G_1(s)G_2(s),$$

---

[7] Recent measurements on middle-ear transmission in cat (Guinan and Peake [1967], Peake *et al.* [1962]) also correspond favorably with these data.

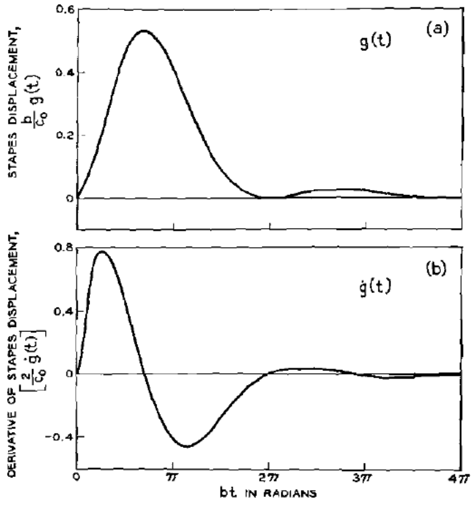


Figure 7.25: Displacement and velocity responses of the stapes to an impulse of pressure at the eardrum

where

$$G_1(s) = \frac{c_0}{s+a}; \qquad G_2(s) = \frac{1}{(s+a)^2 + b^2} \tag{7.5}$$

The inverses of the parts are

$$g_1(t) = c_0 e^{-at}; \qquad g_2(t) = \frac{e^{-at}}{b} \sin bt. \tag{7.6}$$

The inverse of $G(s)$ is then the convolution of $g_1(t)$ and $g_2(t)$

$$g(t) = \int_0^t g_1(\tau) g_2(t-\tau) d\tau,$$

or

$$g(t) = c_0 \frac{e^{-at}}{b}(1 - \cos bt) = \frac{c_0 e^{-bt/2}}{b}(1 - \cos bt). \tag{7.7}$$

Also for future use, note that the time derivative of the stapes displacement is

$$\dot{g}(t) = \frac{c_0 e^{-bt/2}}{2}(2 \sin bt + \cos bt - l). \tag{7.8}$$

Plots of $g(t)$ and $\dot{g}(t)$ are shown in Fig. 7.25. For this middle ear function, the response is seen to be heavily damped. Other data, for example Moller's in Fig. 4.3, suggest somewhat less damping and the possibility of adequate approximation by a still simpler, second-degree function. For such a transmission, the stapes impulse response would be somewhat more oscillatory[8].

---

[8] The modelling technique does not of course depend critically upon the particular set of data being modeled. When more complete physiological measurements are forthcoming, the rational function can be altered to fit the new data.

7.2.3 Combined Response of Middle Ear and Basilar Membrane

The combined response of the models for the middle ear and basilar membrane is

$$\begin{aligned} H_l(s) &= G(S)F_l(s) \\ h_l(t) &= g(t) * f_l(t). \end{aligned} \tag{7.9}$$

For the $F_l(s)$ model described here, the combined time response is easiest obtained by inverse transforming $H_l(s)$. [For other $F_l(s)$ models, the combined response may be more conveniently computed from timedomain convolution.]

The details of the inverse transform of $H_l(s)$ are numerically involved and only the result is of interest here. When the inverse transform is calculated, the result has the form

$$\begin{aligned} h_l(\tau) = \; & Ae^{-b\tau/2} + Be^{-b\tau/2}(\cos b\tau - \tfrac{1}{2}\sin b\tau) + C(e^{-b\tau/2}\sin b\tau) \\ & + De^{-\eta b\tau} + E(e^{-\eta b\tau/2}\sin \eta b\tau) + F(\eta b\tau e^{-\eta b\tau/2}\sin b\tau) \\ & + G(e^{-\eta b\tau/2}\cos \eta b\tau) + H(\eta b\tau e^{-\eta b\tau/2}\cos \eta b\tau); \quad \text{for } \tau \geq 0, \end{aligned} \tag{7.10}$$

where $\tau = (t - T)$; $T = 3\pi/4\beta_l$; $\eta = \beta_l/b$; $\beta_l = 2\alpha_l$; $b = 2a$; $\epsilon_l = 0$; and the A, B, C, D, E, F, G, H are all real numbers which are functions of β_l and b (see (Flanagan [1962a]), for explicit description).

The form of the impulse response is thus seen to depend upon the parameter $\eta = \beta_l/b$. Values of $\eta < 1.0$ refer to (apical) membrane points whose frequency of maximal response is less than the critical frequency of the middle ear. For these points, the middle-ear transmission is essentially constant with frequency, and the membrane displacement is very nearly that indicated by $f_l(t)$ in Eq. (7.2). On the other hand, values of $\eta > 1.0$ refer to (basal) points which respond maximally at frequencies greater than the critical frequency of the middle ear. For these points, the middle-ear transmission is highly dependent upon frequency and would be expected to influence strongly the membrane displacement. To illustrate this point, Eq. (7.10) has been evaluated for η =0.1, 0.8, and 3.0. The result is shown in Fig. 7.26.

For an impulse of pressure delivered to the eardrum, the three solid curves represent the membrane displacements at points which respond maximally to frequencies of 150, 1200, and 4500Hz, respectively. Each of the plots also includes a dashed curve. In Figs. 7.26a and 7.26b, the dashed curve is the membrane displacement computed by assuming the middle-ear transmission to be constant, or flat, and with zero phase. This is simply the response $[\mathcal{L}^{-1}F_l(s)]$. In Fig. 7.26c the dashed curve is the time derivative of the stapes displacement, $g(t)$, taken from 7.25. Fig. 7.25c therefore suggests that the form of the membrane displacement in the basal region is very similar to the derivative of the stapes displacement.

The individual frequency-domain responses for $G(s)$ and $F_l(s)$ have been shown in Figs. 7.22 and 7.24, respectively. The combined response in the frequency domain is simply the sum of the individual curves for amplitude (in db) and phase (in radians).

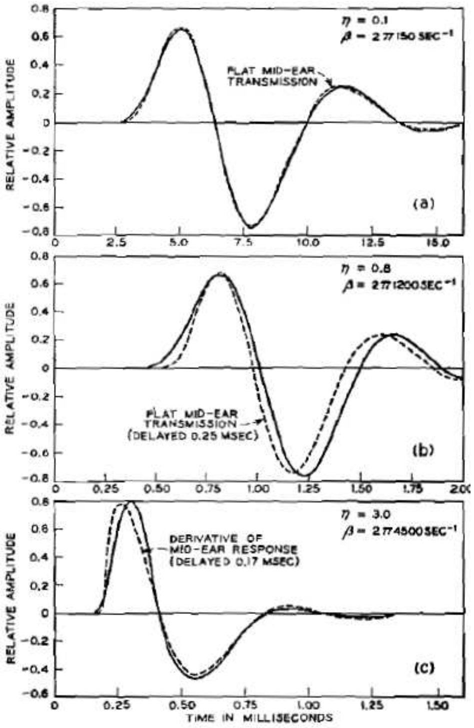


Figure 7.26: Displacement responses for apical, middle and basal points on the membrane to an impulse of pressure at the eardrum. The responses are computed from the inverse transform of $[G(s)F_l(s)]$

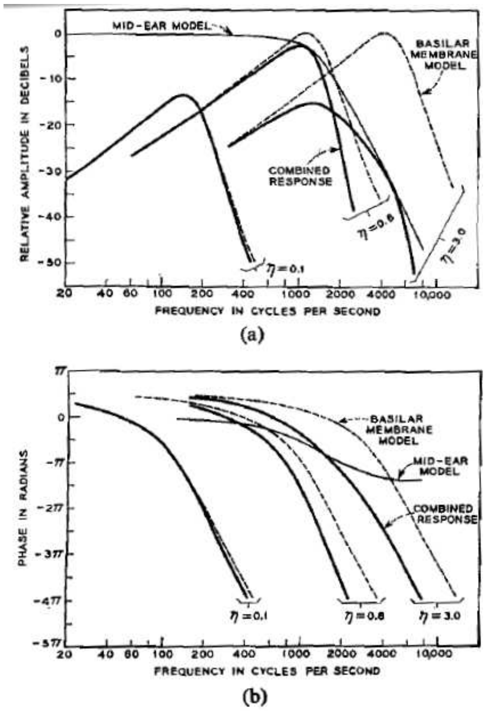


Figure 7.27: (a) Amplitude*vs* frequency responses for the combined model. (b) Phase*vs* frequency responses for the combined model

The combined amplitude and phase responses for the model $G(s)F_l(s)$ are shown in Figs. 7.27a and 7.27b, respectively.

As already indicated by the impulse responses, the response of apical (low-frequency) points on the membrane is given essentially by $F_l(s)$, while for basal (high-frequency) points the response is considerably influenced by the middle-ear transmission $G(s)$. Concerning the latter point, two things may be noted about the frequency response of the membrane model [i.e., $F_l(\omega)$]. First, the low-frequency skirt of the amplitude curve rises at about 6 db/octave. And second, the phase of the membrane model [i.e. $\angle F_l(\omega)$] approaches $+\pi/2$ radians at frequencies below the peak amplitude response[9]. In other words, at frequencies appreciably less than its peak response frequency, the membrane function $F_l(\omega)$ behaves crudely as a differentiator. Because the middle-ear transmission begins to diminish in amplitude at frequencies above about 1500Hz, the membrane displacement in the basal region is roughly the time derivative of the stapes displacement. The waveform of the impulse response along the basal part of the membrane is therefore approximately constant in shape. Along the apical part, however, the impulse response oscillates more slowly (in time) as the apex is approached. This has already been illustrated in Fig. 7.26.

One further point may be noted from Fig. 7.27. Because the amplitude response of the middle-ear declines appreciably at high frequencies, the amplitude response of a basal point is highly asymmetrical. (Note the combined response for $\eta = 3.0$.) The result is that a given basal point–while responding with greater amplitude than any other membrane point at its characteristic frequency–responds with greatest amplitude (but not greater than some other point) at some lower frequency.

7.2.4 An Electrical Circuit for Simulating Basilar Membrane Displacement

On the basis of the relations developed in the previous sections [Eqs. (7.1) and (7.3)], it is possible to construct electrical circuits whose transmission properties are identical to those of the functions $G(s)$ and $F_l(s)$. This is easiest done by representing the critical frequencies in terms of simple cascaded resonant circuits, and supplying the additional phase delay by means of an electrical delay line. Such a simulation for the condition $\epsilon_l = 0$ is shown in Fig. 7.28.

The voltage at an individual output tap represents the membrane displacement at a specified distance from the stapes. The electrical voltages analogous to the sound pressure at the eardrum and to the stapes displacement are also indicated. The buffer

---

[9] This phase behavior is contrary to the physiological phase measurements shown in Fig. 7.6b. Nevertheless, calculations of minimum phase responses for the basilar membrane indicated that the low-frequency phase behavior must approach $n/2$ radians lead (Flanagan *et al.* [1962a]). This earlier analytical prediction (and hence justification for the choice 1=0) has been confirmed by recent measurements. These measurements, using the Mossbauer effect, in fact reveal a leading phase at low frequencies (Johnstone and Boyle [1967], Rhode [1971]).

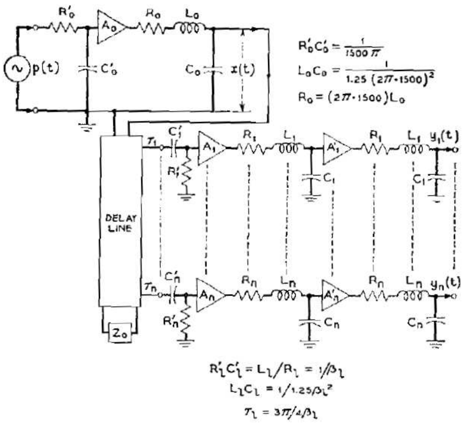


Figure 7.28: Electrical network representation of the ear model

amplifiers labelled A have fixed gains which take account of the proper multiplicative amplitude constants.

The circuit elements are selected according to the constraints stated for $G(s)$ and $F_l(s)$. The constraints are represented by the equations shown in Fig. 7.28 and, together with choice of impedance levels, completely specify the circuit. For each membrane point the relative gains of the amplifiers are set to satisfy the amplitude relations implied in Fig. 7.27a. The gains also take account of the constant multiplying factors in the rational function models. Some representative impulse responses of the analog circuit of Fig. 7.28 are shown in Fig. 7.29a. One notices the degradation in time resolution as the response is viewed at points more apical ward. That is, the frequency resolution of the membrane increases as the apex is approached.

The electrical circuit can also be used in a simple manner to provide an approximation to the spatial derivative of displacement. This function, like the displacement, may be important in the conversion of mechanical-to-neural activity. As mentioned earlier, it has been noted that the inner hair cells in the organ of Corti appear sensitive to longitudinal bending of the membrane, Whereas the outer cells are sensitive to transverse bending (Bekesy [1953]). The former may therefore be more sensitive to the spatial gradient or derivative of membrane displacement, while the latter may be primarily sensitive to displacement.

The differences between the deflection of adjacent, uniformly-spaced points can be taken as an approximation to the spatial derivative. Fig. 7.29b shows the first spatial difference obtained from the analog circuit by taking

$$\frac{\partial y}{\partial x} = \frac{y(t, x + \Delta x) - y(t, x)}{\Delta x},$$

where

$$\Delta x = 0.3\text{mm}$$

The similarity to the displacement is considerable.

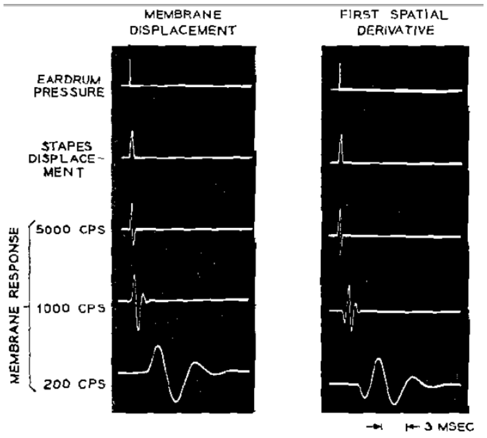


Figure 7.29: (a) Impulse responses measured on the network of Fig. 7.28. (b) First difference approximations to th espatial derivative measured from the network of Fig. 7.28

7.2.5 Computer Simulation of Membrane Motion

If it is desired to simulate the membrane motion at a large number of points and to perform complex operations upon the displacement responses, it is convenient to have a digital representation of the model suitable for calculations in a digital computer. One such digital simulation represents the membrane motion at 40 points (Flanagan [1962b]).

As might be done in realizing the analog electrical circuit, the digital representation of the model can be constructed from sampled-data equivalents of the individual complex pole-pairs and the individual real poles and zeros. The sampled-data equivalents approximate the continuous functions over the frequency range of interest. The computer operations used to simulate the necessary poles and zeros are shown in Fig. 7.30. All of the square boxes labelled D are delays equal to the time between successive digital samples. The input sampling frequency, $1/D$, in the present simulation is 20 KHz, and the input data are quantized to 11 bits. All of the triangular boxes are "amplifiers" which multiply their input samples by the gain factors shown next to the boxes.

Each of the digital operations enclosed by dashed lines is treated as a component block in the program. The block shown in Fig. 7.30a is labelled CP for conjugate-pole. It has the transfer function

$$\frac{Y_a(s)}{X_a(s)} = \left[e^{-2\theta}e^{-2sD} - 2e^{-\theta}\cos\Phi e^{-sD} + 1\right]^{-1} \tag{7.11}$$

which has poles at

$$e^{-(\theta+sD)} = \cos\Phi \pm j\sin\Phi$$

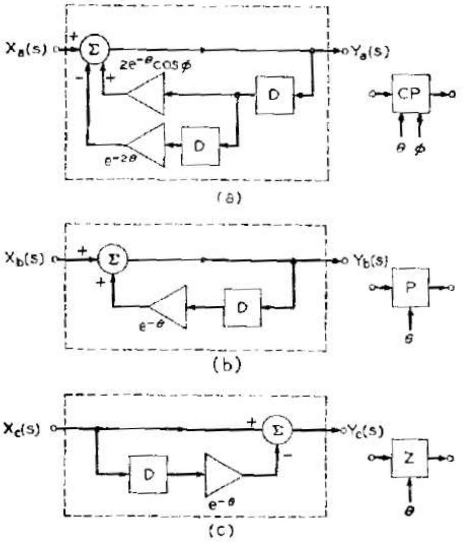


Figure 7.30: Sampled-data equivalents for the complex conjugate poles, real-axis pole, and real-axis zero

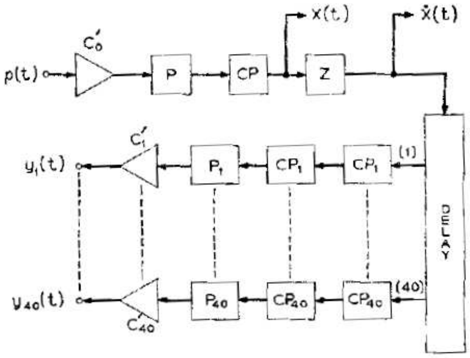


Figure 7.31: Functional block diagram for a digital computer simulation of basilar membrane displacement

or

$$s = \frac{1}{D}\left[-\theta \pm j(\Phi + 2n\pi)\right], \quad n = 0, 1, 2, \ldots$$

so that

$$\theta_l = \alpha_l D \quad \text{and} \quad \Phi_l = \beta_l D,$$

where α_l and β_l are the real and imaginary parts of the pole-pair to be simulated. The pole constellation of the sampled-data function repeats at $\pm j2n\pi/D$ (or at $j2n\pi/5 \times l0^{-5}$ for the 20kHz sampling frequency).

Single real-axis poles are approximated as shown by the P block in Fig. 7.30b. The transfer function is

$$\frac{Y_b(x)}{X_b(s)} = \left[1 - e^{-(\theta + sD)}\right] \tag{7.12}$$

and has poles at

$$s = \frac{1}{D}\left(-\theta \pm j2n\pi\right), n = 0, 1, 2, \ldots$$

The single zero is simulated by the Z block in Fig. 7.30c. Its transfer function is the reciprocal of the P block and is

$$\frac{Y_c(s)}{X_c(s)} = l - e^{-(\theta + sD)} \tag{7.13}$$

with zeros at

$$s = \frac{1}{D}(-\theta j2n\pi), n = 0, 1, 2, \ldots$$

In the present simulation the zero is placed at the origin, so that $\theta = 0$ (i.e., $\epsilon_l = 0$).

The computer operations diagrammed by these blocks were used to simulate the model $G(s)F_l(s)$ for 40 points along the basilar membrane. The points represent 0.5 mm increments in distance along the membrane, and they span the frequency range 75 to 4600Hz. The blocks are put together in the computer program as shown in Fig. 7.31[10] The amplifier boxes c_0' and c_1' in Fig. 7.31 take into account not only the model amplitude constants c_0 and c_1 and the $(2000\pi\beta_l/(\beta_l + 2000\pi))^{0.8}$ factor, but also the amplitude responses of the digital component blocks. For example, it is convenient to make the zero-frequency gain of the CP boxes unity, so each c_1' amplifier effectively includes a $[e^{-2\theta} - 2e^{-\theta} \times \cos\Phi + 1]^2$ term. The overall effect of the c_0' and c_1' gain adjustments is to yield the amplitudes specified by $G(s)F_l(s)$. The delay to each membrane point, $3\pi/4\beta_l$, is simulated in terms of integral numbers of sample intervals. In the present simulation it is consequently represented to the nearest 50 usee.

An illustrative impulse response from the simulation, plotted automatically by the computer, is shown in Fig. 7.32. The displacement response of the membrane at 40

---

[10] In the present case the simulation was facilitated by casting the operations in the format of a special compiler program (Jr. and Lochbaum [1962b], Vyssotsky [1961]).

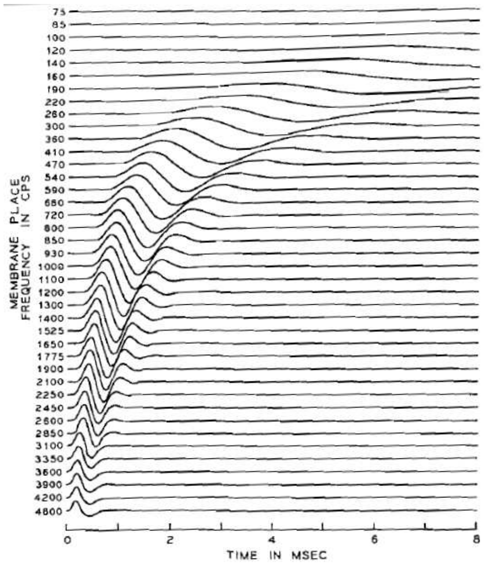


Figure 7.32: Digital computer simulation of the impulse responses for 40 ponits along the basilar membrane. The input signal is a single rarefaction pulse, 100μsec in duration, delivered to th eeardrum at time $t = 0$. (After (Flanagan [1962b]))

points is shown as a function of time. The characteristic frequencies of the membrane points are marked along the y-axis, starting with 4600Hz at the lower (basal) end and going to 75Hz at the upper (apical) end. Time is represented along the x-axis. The input pressure signal $p(t)$ is a single positive pulse 100 μsec in duration and delivered at $t = 0$. The responses show that the basal points respond with short latency and preserve a relatively broad-band version of the input pulse. The apical points display increasingly greater latency and progressive elimination of high-frequency content from the signal.

These same attributes of the membrane are put in evidence by a periodic pulse signal, which will be of interest in the subsequent discussion. Fig. 7.33 shows the reponse to an input signal composed of alternate positive and negative pulses of 100μsec duration, produced at a fundamental frequency of 100Hz and initiated at $t = 0$. The time between alternate pulses is therefore 5 msec. At the apical (low-frequency) end of the membrane, the frequency resolution is best, and the displacement builds up to the fundamental sinusoid. At the basal (highfrequency) end, the membrane resolves the individual pulses in time. The responses also reflect the transit delay along the membrane.

The utility of the computation model depends equally upon its mathematical tractability and its adequacy in approximating membrane characteristics. Given both, the model can find direct application in relating subjective and physiological auditory behavior. More specifically, it can be useful in relating psychoacoustic responses to patterns of membrane displacement and in establishing an explanatory framework for

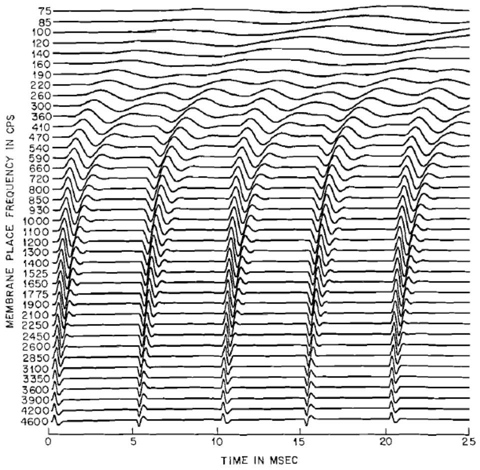


Figure 7.33: Digital computer output for 40 simulated points along the basilar membrane. Each trace is the displacement response of a given membrane place to alternate positive and negative pressure pulses. The pulses have 100μsec duration and are produced at a rate of 200 Hz. The input signal is applied at the eardrum and is initiated at time zero. The simulated membrane points are spaced by 0.5mm. Their characteristic frequencies are indicated along the ordinate. (After (Flanagan [1962b]))

the neural representation of auditory information.

7.2.6 Transmission Line Analogs of the Cochlea

The preceding discussion has concerned an "input-output" formulation of the properties of the middle ear and basilar membrane. This approach, for computational and applicational convenience, treats the mechanism in terms of its terminal characteristics. A number of derivations have been made, however, in which the distributed nature of the inner ear is taken into account, and the detailed functioning of the mechanism is examined (Peterson and Bogert [1950], Bogert [1951], Ranke [1942], Zwislocki [1948], Oetinger and Hauser [1961])). At least two of these treatments have yielded transmission line analogs for the inner ear.

The simplifying assumptions made in formulating the several treatments are somewhat similar. By way of illustration, they will be indicated for one formulation (Peterson and Bogert [1950]). The cochlea is idealized as shown in Fig. 7.34. The oval window is located at O and the round window at R. The distance along the cochlea is reckoned from the base and denoted as x. The cross-sectional areas of the scalas vestibuli and tympani are assumed to be identical functions of distance, $S_0(x)$. The width of the basilar membrane is taken as $b(x)$, and the per-unit-area distributed mass, resistance and stiffness of the basilar membrane (or, more precisely, of the cochlear duct separating the scalas) are respectively $m(x)$, $r(x)$ and $k(x)$. The mechanical constants used are deduced from the physiological measurements of Békésy.

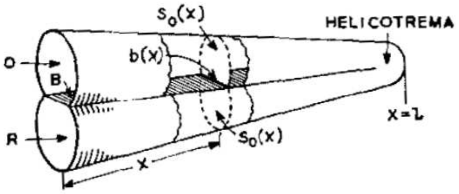


Figure 7.34: Idealized schematic of the cochlea. (After PETERSON and BOGERT)

The following simplifying assumptions are made. All amplitudes are small enough that nonlinear effects are excluded. The stapes produces only plane compressional waves in the scalas. Linear relations exists between the pressure difference across the membrane at any point and the membrane displacement, velocity and acceleration at that point. The vertical component of particle velocity in the perilymph fluid is small and is neglected. A given differential element of the membrane exerts no mutual mechanical coupling on its adjacent elements.

The relations necessary to describe the system are the equations for a plane compressional wave propagating in the scalas and the equation of motion for a given membrane element. For a plane wave in the scalas, the sound pressure, p, and particle velocity, u, are linked by the equation of motion

$$\rho\frac{\partial u}{\partial t} = -\frac{\partial p}{\partial x}, \tag{7.14}$$

where p is the average density of the perilymph fluid. If the membrane displacements are small, the equations of continuity (mass conservation) for the two scalas are

$$\frac{\partial(u_v S)}{\partial x} = -\frac{S}{\rho c^2}\frac{\partial p_v}{\partial t} - vb$$

$$\frac{\partial(u_t S)}{\partial x} = -\frac{S}{\rho c^2}\frac{\partial p_t}{\partial t} + vb \tag{7.15}$$

where v is the membrane velocity and the subscripts v and t denote vestibuli and tympani, respectively. These relations state that the rate of mass accumulation for an elemental volume in the scala is equal to the temporal derivative of the fluid density.

The equation of motion for the membrane is

$$(p_v - p_t) = m\frac{dv}{dt} + rv + k\int vdt, \tag{7.16}$$

where the pressure difference between the scalas $(p_v - p_t)$ is the forcing function for a membrane element.

Eqs. (7.14) to (7.16) can be solved simultaneously for the pressures and velocities involved. A typical solution for the instantaneous pressure difference produced

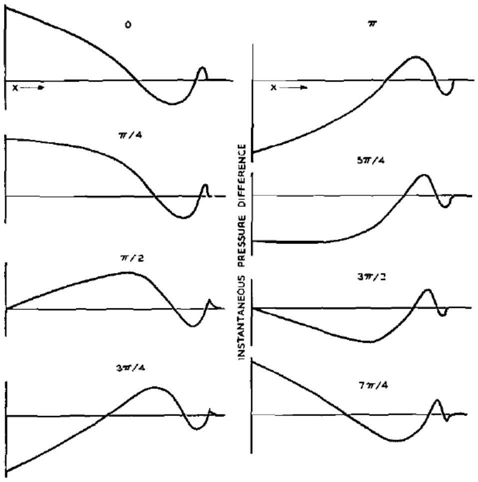


Figure 7.35: Instantaneous pressure difference across the cochlear partition at successive phases in one period of a 1000Hz excitation. (After (Peterson and Bogert [1950]))

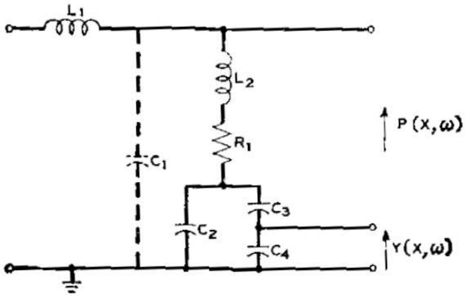


Figure 7.36: Electrical network section for representing an incremental length of the cochlea. (After (Bogert [1951]))

across the membrane by an excitation of 1000Hz is shown in Fig. 7.35. The pressure difference is shown at $\frac{1}{8}$msec intervals (every $\pi/4$ radians of phase) for one cycle. The traveling wave nature of the excitation is apparent, with the speed of propagation along the membrane being greater at the basal end and becoming slower as the apex (helicotrema) is approached.

From the pressure and velocity solutions, an equivalent four-pole network can be deduced for an incremental length of the cochlea. Voltage can be taken analogous to sound pressure and current analogous to volume velocity. Such a network section is shown in Fig. 7.36 (BOGERT). Here L_1 represents the mass of the fluid in an incremental length of the scalas; C_1 the compressibility of the fluid; and L_2, R_2, C_2, C_3, and C_4 represent the mechanical constants of the membrane. The voltage $P(x, \omega)$ represents the pressure difference across the membrane as a function of distance and

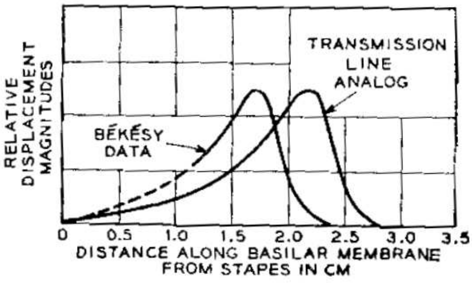


Figure 7.37: Comparison of the displacement response of the transmission line analog of the cochlea to physiological data for the ear. (After BOGERT)

frequency, and the voltage $Y(x, \omega)$ represents the membrane displacement.

A set of 175 such sections has been used to produce a transmission line analog of the cochlea (Bogert [1951]). The displacement responses exhibited by the line compare well in shape with those measured by Békésy on real cochleas. An illustrative response is shown in Fig. 7.37. Some differences are found in the positions of peak response and in the lowest frequencies which exhibit resonance phenomena. Probable origins of the differences are the uncertainties connected with the spatial variation of the measured mechanical constants of the membrane and the neglect of mutual coupling among membrane elements. Despite the uncertainties in the distributed parameters, the transmission line analog provides a graphic demonstration of the traveling-wave nature of the basilar membrane motion.

7.3 Illustrative Relations between Subjective and Physiological Behavior

The ear models discussed above describe only the mechanical functions of the peripheral ear. Any comprehensive hypothesis about auditory perception must provide for the transduction of mechanical displacement into neural activity. As indicated earlier, the details of this process are not well understood. The assumptions that presently can be made are of a gross and simplified nature. Three such assumptions are useful, however, in attempting to relate physiological and subjective behavior. Although oversimplifications, they do not seem to violate known physiological facts. The first is that sufficient local deformation of the basilar membrane elicits neural activity in the terminations of the auditory nerve. A single neuron is presumably a binary (fired or unfired) device. The number of neurons activated depends in a monotonic fashion upon the amplitude of membrane displacement[11]. Such neural ac-

---

[11] Psychological and physiological evidence suggests that the intensity of the neural activity is a power-law function of the mechanical displacement. A single neuron is also refractory for a given period after firing. A limit exists, therefore, upon the rate at which it can fire.

tivity may exist in the form of volleys triggered synchronously with the stimulus, or in the form of a signalling of place localization of displacement. Implicit is the notion that the displacement–or perhaps spatial derivatives of displacement-must exceed a certain threshold before nerve firings take place. Second, neural firings occur on only one "polarity" of the membrane displacement, or of its spatial derivative. In other words, some process like half-wave rectification operates on the mechanical response. Third, the membrane point displacing with the greatest amplitude originates the predominant neutral activity. This activity may operate to suppress or inhibit activity arising from neighboring points. These assumptions, along with the results from the models, have in a number of instances been helpful in interpreting auditory subjective behavior. Without going into any case in depth, several applications can be outlined.

7.3.1 Pitch Perception

Pitch is that subjective attribute which admits of a rank ordering on a scale ranging from low to high. As such, it correlates strongly with objective measures of frequency. One important facet of auditory perception is the ability to ascribe a pitch to sounds which exhibit periodic characteristics.

Consider first the pitch of pure (sinusoidal) tones. For such stimuli the basilar membrane displacements are, of course, sinusoidal. The frequency responses given previously in Fig. 7.27a indicate the relative amplitudes of displacement versus frequency for different membrane points. At any given frequency, one point on the membrane responds with greater amplitude than all others. In accordance with the previous assumptions, the most numerous neural volleys are elicited at this maximum point. For frequencies sufficiently low (less than about 1000Hz), the volleys are triggered once per cycle and at some fixed epoch on the displacement waveform. Subsequent processing by higher centers presumably appreciates the periodicity of the stimulus-locked volleys.

For frequencies greater than about 1000 to 2000Hz, electro-physiological evidence suggests that synchrony of neural firings is not maintained (Galambos [1958]). In such cases, pitch apparently is perceived through a signalling of the place of greatest membrane displacement. The poorer frequency resolution of points lying in the basal part of the basilar membrane probably also contributes to the psychoacoustic fact that pitch discrimination is less acute at higher frequencies.

Suppose the periodic sound stimulus is not a simple sinusoidal tone but is more complex, say repeated sharp pulses. What pitch is heard? For purpose of illustration, imagine the stimulus to be the alternately positive and negative impulses used to illustrate the digital simulation in Fig. 7.33. Such a pulse train has a spectrum which is odd-harmonic. If the pulses occur slowly enough, the membrane displacement at all points will resolve each pulse in time. That is, the membrane will have time to execute a complete, damped impulse response at all places for each pulse, whether positive or negative. Such a situation is depicted by the analog membrane responses shown

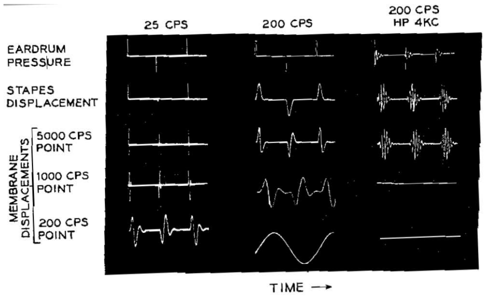


Figure 7.38: Membrane displacement responses for filtered and unfiltered periodic pulses. The stimulus pulses are alternately positive and negative. The membrane displacements are simulated by the electrical networks shown in Fig. 7.28. To display the waveforms more effectively, the traces are adjusted for equal peak-to-peak amplitudes. Relative amplitudes are therefore not preserved

in the left column of Fig. 7.38. The fundamental frequency of excitation is 25Hz (50 pps). The waveforms were measured from analog networks such as illustrated in 7.28.

For this low pulse rate condition, one might imagine that neural firings synchronous with each pulse–regardless of polarity–would be triggered at all points along the membrane. The perceived pitch might then be expected to be equal to the pulse rate. Measurements show this to be the case (Flanagan and Guttman [1960]). Furthermore, the model indicates that a pulse signal of this low rate causes the greatest displacements near the middle portion of the membrane, that is, in the vicinity of the place maximally responsive to about 1500Hz.

If, on the other hand, the fundamental frequency of excitation is made sufficiently high, say 200Hz or greater, the fundamental component will be resolved (in frequency) at the most apically-responding point. This situation is illustrated for a 200Hz fundamental by the traces in the second column of Fig. 7.38. The 200Hz place on the membrane displaces with a nearly pure sinusoidal motion, while the more basal points continue to resolve each pulse in time. At the apical end, therefore, neural volleys might be expected to be triggered synchronously at the fundamental frequency, while toward the basal end the displacements favor firings at the pulse rate, that is, twice per fundamental period. Psychoacoustic measurements indicate that the apical, fundamental-correlated displacements are subjectively more significant than the basal, pulse-rate displacements. The fundamental-rate volleys generally predominate in the percept, and the pitch is heard as 200 sec−1. At some frequency, then, the pitch assignment switches from pulse rate to fundamental.

The pulse pattern illustrating the computer simulation in Fig. 7.33 is the same positive-negative pulse alternation under discussion, but it is produced at a fundamental frequency of 100 Hz. This frequency is immediately in the transition range between the fundamental and pulserate pitch modes. One notices in Fig. 7.33 that the ear is beginning to resolve the fundamental component in relatively low amplitude at the apical end of the membrane, while the pulse rate is evident in the basal displacements. One might suppose for this condition that the pulse rate and fundamental cues are strongly competing, and that the pitch percept is ambiguous. Subjective measurements bear this out.

Another effect becomes pronounced in and near the pitch-transition region corresponding to the conditions of Fig. 7.33. A fine structure in the perception of pulse pitch becomes more evident. The membrane region where displacement amplitude is greatest is in the place-frequency range 600 to 1500Hz. In this region the displacement response to a pulse has a period which is an appreciable fraction of the pulse repetition period. That is, the half-period time of the pulse response is a significant percentage of the pulse period. Assume as before that neural firings occur only on positive deflections of the membrane. The intervals between firings on fibers originating from a given place in this region should, therefore, be alternately lengthened and shortened. The change in interval (from strict periodicity) is by an amount equal to the half-period of the pulse response at that place. One might expect, therefore, a bimodality in the pitch percept. If f_d is the place-frequency of dominant membrane motion and r the signal pulse rate, the perceived pitch f_p should correspond to

$$f_p = \left[\frac{1}{r} \pm \frac{1}{2f_d}\right]^{-1}.$$

This bimodality in the pitch percept is in fact found (Rosenberg [1965], Ritsma [1967]).

If the 200Hz stimulus in the middle column of Fig. 7.38 is high-pass filtered at a sufficiently high frequency, only the basal displacements remain effective in producing the pitch percept. For example, the membrane displacements for a high-pass filtering at 4000 Hz are shown in the third column of Fig. 7.38. If the present arguments continue to hold, such a filtering should change the percept from the fundamental mode back to the pulse-rate mode. The reason, of course, is that the time resolution of the basal end separates each pulse, whether positive or negative. This hypothesis is in fact sustained in psychoacoustic measurements (Guttman and Flanagan [1964]).

A somewhat more subtle effect is obtained if the high-pass filtering is made at a fairly small harmonic number, for example, at the second harmonic, so as to remove only the fundamental component. Under certain of these conditions, the membrane may generate displacements which favor a difference-frequency response. For a stimulus with odd and even components, the pitch percept can be the fundamental, even though the fundamental is not present in the stimulus.

7.3.2 Binaural Lateralization

Another aspect of perception is binaural lateralization. This is the subjective ability to locate a sound image at a particular point inside the head when listening over

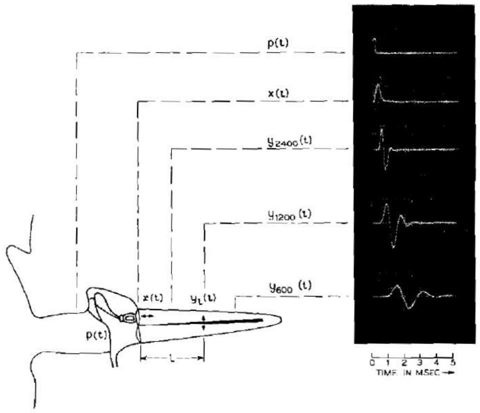


Figure 7.39: Basilar membrane responses at the 2400, 1200 and 600Hz points to a pressure-rarefaction pulse of 100μsec duration. The responses are measured on the electrical analog circuit of Fig. 7.28. Relative amplitudes are preserved

earphones. If identical clicks (impulses of sound pressure) are produced simultaneously at the two ears, a normal listener hears the sound image to be located exactly in the center of his head. If the click at one ear is produced a little earlier or with slightly greater intensity than the other, the sound image shifts toward that ear. The shift continues with increasing interaural time or intensity difference until the image moves completely to one side and eventually breaks apart. One then begins to hear individual clicks located at the ears.

Naively we suppose the subjective position of the image to be determined by some sort of computation of coincidence between neural volleys. The volleys originate at the periphery and travel to higher enters via synaptic pathways. The volley initiated earliest progresses to a point in the neural net where a coincidence occurs with the later volley. A subjective image appropriately off-center is produced. To the extent that intensity differences can shift the image position, intensity must be coded–at least partially–in terms of volley timing. As has been the case in pitch perception, there are several research areas in binaural phenomena where the computational model described in Section 7.2 has been helpful in quantifying physiological response and relating it to subjective behavior. One such area concerns the effects of phase and masking upon the binaural lateralization of clicks.

If a pulse of pressure rarefaction is produced at the eardrum, the drum is initially drawn outward. The stapes is also initially drawn outward, and the membrane is initially drawn upward. The stapes and membrane displacements (as described by the model) in response to a rarefaction pulse of 100μsec duration are shown by the waveforms at the right of Fig. 7.39. The pulse responses of three different membrane points are shown, namely, the points maximally responsive to 2400Hz, 1200Hz, and 600Hz, respectively. The stapes displacement is a slightly integrated version of the input. The membrane responses reflect the vibratory behavior of the particular points as well as the travelingwave transit delay to the points.

According to the model, broadband pulses produce the greatest displacements near the middle of the membrane, roughly in the region maximally responsive to about 1500Hz. The magnitude of displacement is less at places either more toward the base or more toward the apex. It has been hypothesized that the most significant neural activity is generated at the membrane point displacing with the greatest amplitude. Further, electro-physiological data suggest that neural firings occur at some threshold only on unipolar motions of the basilar membrane. (For the outer hair cells, these are motions which drive the basilar membrane toward the tectorial membrane.) The oscillatory behavior of the pulse response suggests, too, that multiple or secondary neural firings might be elicited by single stimulus pulses.

If pulses are supplied to both ears, a centered sound image is heard if the significant neural activity is elicited simultaneously. Suppose that the input pulses are identical rarefaction pulses. The maximum displacements occur near the middle of the membrane. For simplicity imagine that the neural firings are triggered somewhere near the positive crests of the displacement waves. For this cophasic condition, a centered image is heard if the input pulses are produced simultaneously, or if the interaural time is zero. Suppose now that the pulse to one of the ears is reversed in phase to a pressure condensation. The membrane responses for this ear also change sign and are the negatives of those shown in Fig. 7.39. Their first positive crests now occur later by about one-half cycle of the displacement at each point. At the middle of the membrane this half-cycle amounts to about 300 to 400μsec. To produce a centered image for the antiphasic condition, then, one would expect that the condensation pulse would have to be advanced in time by this amount.

The membrane point which displaces with the greatest coherent amplitude can be manipulated by adding masking noise of appropriate frequency content. That is, the place which normally responds with greatest amplitude can be obscured by noise, and the significant displacement caused to occur at a less sensitive place. For example, suppose that the basal end of the membrane in one ear is masked by high-pass noise, and the apical end of the membrane in the other ear is masked by low-pass noise. If the listener is required to adjust stimulus pulses to produce a centered image, the fusion must be made from apical-end information in one ear and basal-end in the other. The resulting interaural time would then reflect both the oscillatory characteristics of the specific membrane points and the traveling-wave delay between them.

Experiments show these time dependencies to be manifest in subjective behavior (Flanagan *et al.* [1962a]). The test procedure to measure them is shown in Fig. 7.40. Identical pulse generators produce 100μsec pulses at a rate of 10 per second. Pulse amplitude is set to produce a 40 db sensation level. The subject, seated in a sound-treated room, listens to the pulses over condenser earphones. (Condenser phones are used because of the importance of good acoustic reproduction of the pulses.) He has a switch available to reverse the polarity of the pulses delivered to the right ear so that it can be made a condensation instead of the normal rarefaction.

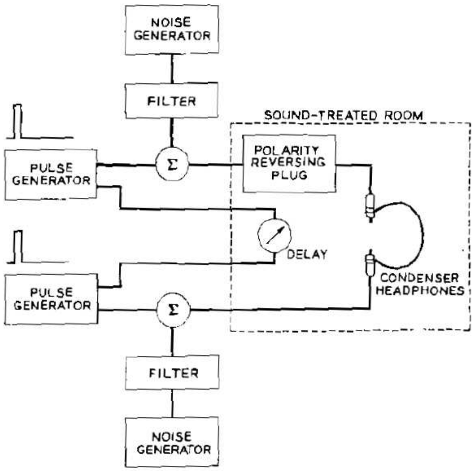


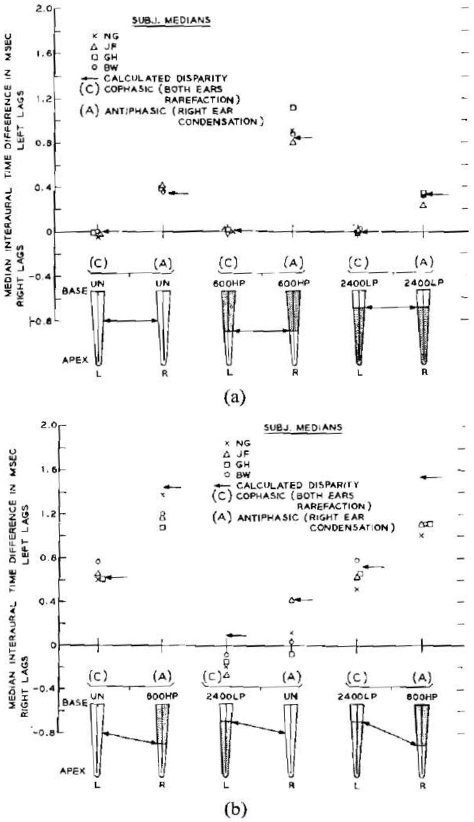
Figure 7.40: Experimental arrangement for measuring the interaural times that produce centered sound images. (After (Flanagan *et al.* [1962a])

The subject also has a delay control which varies the relative times of occurrence of the two pulses over a range of 5 msec. Two uncorrelated noise generators supply masking noise via variable filters. (A separate experiment was conducted to determine the filtered noise levels necessary to mask prescribed spectral portions of the pulse stimuli.)

For a given masking and pulse polarity condition, the subject is required to adjust the delay to produce a centered sound image in his head. Multiple images are frequently found, with the more subtle, secondary images apparently being elicited on secondary bounces of the membrane.

Fig. 7.41 shows the results for principal-image fusions under a variety of masking conditions. Fig. 7.41a gives results for unmasked and symmetrically-masked conditions, and Fig. 7.41b gives the results for asymmetrical masking. The data are for four subjects, and each point is the median of approximately 15 principal-image responses. Each two sets of points is bracketed along the abscissa. The set labelled C is the cophasic response and that labelled A is the antiphasic. The cophasic conditions are rarefaction pulses in both ears. The antiphasic conditions are rarefaction in the left ear and condensation in the right ear.

Each bracket corresponds to the masking conditions represented by the schematic cochleas drawn below the brackets. The labelling at the top of each cochlea gives the masking condition for that ear. For example, the UN means unmasked. The dark shading on the cochleas indicates the membrane regions obscured by masking noise. The double arrow between each pair of cochleas indicates approximately the points of maximum, unmasked displacement. For example, in the first case of Fig. 7.41a, which is the unmasked case, the maximum displacements occur near the middles of the two membranes.



(a)

(b)

Figure 7.41: Experimentally measured interaural times for lateralizing cophasic and antiphasic clicks. Several conditions of masking are shown. (a) Unmasked and symmetrically masked conditions. (b) Asymmetrically masked conditions. The arrows indicate the interaural times predicted from the basilar membrane model

The single arrows in the vicinity of the plotted responses are estimates or the interaural times calculated from the basilar membrane model. The estimates are made by assuming the neural firings to be produced at the positive crest of the displacement at the most significant place. The arrows therefore represent the time differences between the first positive crests at the places indicated in the cochlear diagrams. As such, they include the transit time to the particular place, plus the initial quarter-cycle duratiori of the pulse response.

The actual threshold for neural firing is of course not known, and is very likely to be dependent upon place. In the symmetrically-masked conditions, an actual knowledge of the threshold is not of much consequence since the threshold epoch, whether it is at the crest or down from the crest, should be about the same in the two ears. For these cases, therefore, it is the half-cycle time of the displacement wave that is important. Fig. 7.41a shows that the measured responses do, in fact, agree relatively well with this simple estimate of the interaural time. All of the principal cophasic fusions are made for essentially zero time, and the antiphasic lateralizations reflect the half-cycle disparity of the appropriate places, with the condensation pulse always leading.

The agreement is not as good for the asymmetrically-masked cases shown in Fig. 7.41b. Signal loudnesses are different in the two ears, and the neural thresholds probably vary with place. The times of the initial positive crests would not be expected to give very realistic estimates of the interaural times. It becomes much more important to have a knowledge of the actual threshold levels and the relative amplitudes of the displacements. Even so, it is interesting to note to what extent the simple positive-crest estimates follow the data.

In the first condition, the left ear is unmasked and the right ear has masking noise high-pass filtered at 600Hz (600 HP). The cophasic interaural time is predicted to be on the order of 600 usee, and the measurements do give essentially this figure. The antiphasic condition is expected to be on the order or 1450 μsec, but the measured median response is a little less, about 1200 μsec.

The next case has the left ear masked with noise low-pass filtered at 2400Hz (2400 LP), and the right ear is unmasked. The cophasic condition is expected to yield an interaural time of slightly less than 100 μsec, with the left ear lagging, but the experimental measurements actually give a right ear lag of about 150 μsec. The relatively wide spread of the subject medians in the asymmetrical cases, and in particular for the cases involving 2400 LP, show that these lateralizations are considerably more difficult and more variable than the symmetrical cases. The antiphasic response for this same condition is estimated to give an interaural time on the order of 400 μsec, but again the responses are variable with the median falling at about 100 μsec. One subject's median actually falls on the right-lag side of the axis.

The final condition has 2400 LP in the left ear and 600 HP in the right ear. The cophasic fusion is expected to be in the neighborhood of 700 μsec, and the measured response is found about here. The antiphasic condition should yield an interaural

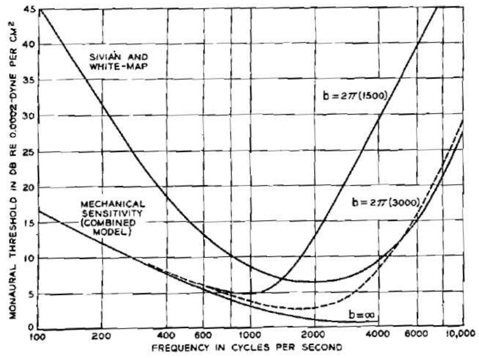


Figure 7.42: Relation between the mechanical sensitivity of the ear and the monaural minimum audible pressure threshold for pure tones

time on the order of 1550 μsec, but the measurements produce a median slightly greater than 1100 μsec.

Clearly, the simple assumption of neural firing at the positive crests (or some other fixed epoch) of the displacement is not adequate to specify all of the interaural times. The real thresholds for firing are likely to vary considerably with place. In fact, by taking data such as these, plus the displacement waves from the model, the possibility exists for working backwards to deduce information about neural threshold epochs. More broadly than this, however, the present results suggest strong ties between subjective response and the detailed motion of the basilar membrane.

7.3.3 Threshold Sensitivity

The combined response curves in Fig. 7.27a indicate that the ear is mechanically more sensitive to certain frequencies than to others. A similar frequency dependence exists subjectively. To what extent are the variations in the threshold of audibility accounted for simply by the mechanical sensitivity of the ear?

The envelope of the peak displacement responses in Fig. 7.27a can be compared with the subjectively determined minimum audible pressure for pure (sine) tones. Fig. 7.42 shows this comparison. The agreement is generally poor, although the gross trends are similar. One curve in the figure is based on the 1500 Hz critical frequency for the middle ear. The earlier discussion has pointed up the uncertainty and variability of this figure. If a critical frequency of 3000Hz is taken for the middle car, the fit to the threshold curve at high frequencies is more respectable[12]. The match at low frequencies, however, is not improved, but this is of less concern for a different reason.

At the low frequencies, the disparity between the mechanical and subjective sensi-

---

[12] Note, too, that the membrane velocity response $\dot{y}(t)$ provides a better fit to the tone threshold than does the displacement, $y(t)$. $\dot{y}(t)$ includes an additional +6 db/oct. component.

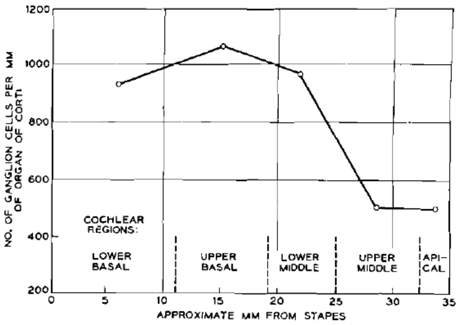


Figure 7.43: Average number of ganglion cells per mm length of organ of Corti. (After GUILD et at.)

tivity may be partially a neural effect. According to the earlier assumptions, the number of neurons activated bears some monotonic relation to amplitude of membrane displacement. Perception of loudness is thought to involve possibly temporal and spatial integrations of neural activity. If a constant integrated neural activity were equivalent to constant loudness, the difference between mechanical and subjective sensitivities might be owing to a sparser neural density in the apical (low-frequency) end of the cochlea. There is physiological evidence to this effect.

In histological studies, Guild*et al.* counted the number of ganglion cells per mm length of the organ of Corti (Guild *et al.* [1931]). Their results for normal ears are summarized in Fig. 7.43. These data show a slight decrease in the number of cells at the basal end, and a substantial decrease in the density as the apex is approached. The innervation over the middle of the membrane is roughly constant.

One can pose similar questions about threshold sensitivity to short pulses or clicks of sound. For pulses of sufficiently low repetition rate, the maximal displacement of the membrane–as stated before–is near the middle. According to the model, the place of maximum displacement remains near the middle for pulse rates in excess of several hundred per second. In its central region, the resonance properties of the membrane permit temporal resolution of individual exciting pulses for rates upwards of 1000Hz. If the predominant displacement takes place in one vicinity for a large range of pulse rates, polarity patterns and pulse durations, how might the subjective threshold vary with these factors, and how might it be correlated with the membrane motion? At least one examination of this question has been made (Flanagan [1961]). The results can be briefly indicated.

Binaural thresholds of audibility for a variety of periodic pulse trains with various polarity patterns, pulse rates and durations are shown in Fig. 7.44. The data show that the thresholds are relatively independent of polarity pattern. For pulse rates less than 100Hz, the thresholds are relatively independent of rate, and are dependent only upon duration. Above 100Hz, the thresholds diminish with increasing pulse rate. The amplitude of membrane displacement would be expected to be a function of pulse

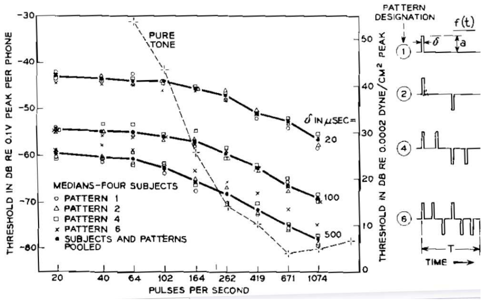


Figure 7.44: Binaural thresholds of audibility for periodic pulses. (After FLANAGAN, 1961a)

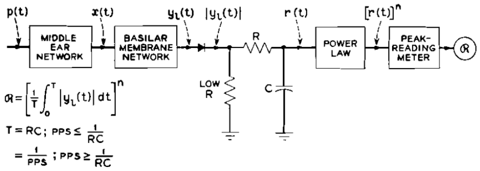


Figure 7.45: Model of the threshold of audibility for the pulse data shown in Fig. 7.44

duration and to produce a lower threshold for the longer pulses, which is the case. For rates greater than 100Hz, however, some other nonmechanical effect apparently is of importance. The way in which audible pulse amplitude diminishes suggests a temporal integration with a time constant of the order of 10 msec.

Using the earlier assumptions about conversion of mechanical to neural activity, one might ask "what processing of the membrane displacement at the point of greatest amplitude would reflect the constant loudness percept at threshold." One answer is suggested by the operations illustrated in Fig. 7.45. The first two blocks represent middle ear transmission [as specified in Eq. (7.3)] and basilar membrane displacement [vicinity of the 1000Hz point, as specified in Eq. (7.1)]. The diode represents the half-wave rectification associated with neural firings on unipolar motions of the membrane. The RC integrator has a 10 msec time constant, as suggested by the threshold data. The powerlaw element (exponent=0.6) represents the power-law relation found in loudness estimation[13]. A meter indicates the peak value of the output of the power-law device. When the stimulus conditions represented by points on the threshold curves in Fig. 7.44 are applied to this circuit, the output meter reads the same value, namely, threshold.

One can also notice how this simple process might be expected to operate for sine wave inputs. Because the integration time is 10 msec, frequencies greater than about 100Hz produce meter readings proportional to the average value of the half-

---

[13] The power-law device is not necessary for simple threshold indications of "audible-inaudible." It is necessary, however, to represent the growth of loudness with sound level, and to provide indications of subjective loudness above threshold.

wave rectified sinusoid. In other words, the meter reading is proportional to the amplitude of the sine wave into the rectifier. Two alterations in the network circuitry are then necessary. First, the basilar membrane network appropriate to the point maximally responsive to the sine frequency must be used. This may be selected from an ensemble of networks. And second, to take account of the sparser apical innervation, the signal from the rectifier must be attenuated for the low-frequency networks in accordance with the difference between the mechanical and subjective sensitivity curves in Fig. 7.42. The power-law device is still included to simulate the growth of loudness with sound level.

7.3.4 Auditory Processing of Complex Signals

The preceding discussions suggest that the extent to which subjective behavior can be correlated with (and even predicted by) the physiological operation of the ear is substantial. Recent electrophysiological data link neural activity closely with the detailed mechanical motion of the basilar membrane. Subjective measurements, such as described in the foregoing sections, lend further support to the link. Psychological and physiological experimentation continue to serve jointly in expanding knowledge about the processes involved in converting the mechanical motions of the inner ear into intelligence-preserving neural activity.

The physiological-psychoacoustic correlations which have been put forward here have involved only the simplest of signals–generally, signals that are temporally punctuate or spectrally discrete, or both. Furthermore, the correlations have considered only gross and salient features of these signals, such as periodicity or time of occurrence. The primary aim has been to outline the peripheral mechanism of the ear and to connect it with several phenomena in perception. Little has been said about classical psychoacoustics or about speech perception. As the stimuli are made increasingly complex–in the ultimate, speech signals–it seems clear that more elaborate processing is called into play in perception. Much of the additional processing probably occurs centrally in the nervous system. For such perception, the correlations that presently can be made between the physiological and perceptual domains are relatively rudimentary. As research goes forward, however, these links will be strengthened.

The literature on hearing contains a large corpus of data on subjective response to speech and speech-like stimuli. There are, for example, determinations of the ear's ability to discriminate features such as vowel pitch, loudness, formant frequency, spectral irregularity and the like. Such data are particularly important in establishing criteria for the design of speech transmission systems and in estimating the channel capacity necessary to transmit speech data. Instead of appearing in this chapter, comments on these researches have been reserved for a later, more applied discussion where they have more direct application to transmission systems.

7.4 Homework

Problem 7.1

Consider the acoustic pressure signal

$$p(t) = A\cos(2\pi 499t) + A\cos(2\pi 501t) + A\cos(2\pi 999t) + A\cos(2\pi 1001t) \tag{7.17}$$

a. Suppose that A is adjusted so that the intensity level of the signal is $I_p = 40$dB SPL. What is A?

b. Assume that a 500Hz tone with an amplitude of A is exactly as loud as a 1000Hz tone with an amplitude of A. In terms of the variable A, what is the loudness in sones of the signal $p(t)$ given in equation 7.17?

Problem 7.2

a. What is the loudness level, in phons, of a 40dB SPL tone at 1000Hz? What is the loudness, in sones, of the same tone?

b. What is the loudness level, in phons, of a 40dB SPL tone at 500Hz? What is the loudness, in sones, of the same tone?

c. What is the loudness, in sones, of a stimulus consisting of the 500Hz and 1000Hz tones (each at level 40dB SPL) played at the same time?

d. What is the loudness, in sones, of a stimulus consisting of two tones played at the same time: one at 1000Hz/40dB SPL, and one at 1050Hz/40dB SPL?

Problem 7.3

a. Weber's law states that many perceived quantities Ω (e.g. pitch, loudness, light intensity) are related to a real-world stimulus S by the relationship

$$d\Omega = \frac{dS}{S}$$

Prove that Weber's law describes the relationship between the loudness level of a tone at 1000Hz (in phons) and its acoustic intensity (in Watts/ square meter).

b. Stevens proposed that some perceived quantities, instead, follow the general law

$$\Omega = C(S - S_T)^E$$

where S_T is a threshold stimulus quantity, and C and E are constants that depend on the problem. If the perceived quantity is loudness (in sones), and the acoustic quantity is intensity (in Watts/square meter), what are the variables C, S_T, and E?

Problem 7.4

The auditory resolution of the purple-spotted stonewall toad is considerably less than that of human beings. The equivalent rectangular bandwidth of a purple-spotted stonewall toad is given by

$$\text{ERB}(x) = \frac{f_c(x)}{2} + 100\text{Hz}$$

Construct a set of triangular pseudo-critical-band-filters centered at the frequencies $f_c(x) = f_c(x-1) + \text{ERB}(x-1)$, of the form

$$H(x, 2\pi f) = \max\left(0, \frac{\text{ERB}_x - |f - f_c(x)|}{\text{ERB}(x)}\right)$$

Draw the first four filters $H(x, 2\pi f)$ for $1 \leq x \leq 4$. Assume that $\Omega_c(0) = 0$.

Problem 7.5

In this problem, you will devise an experimental protocol for measuring the shape of auditory filters. The derivation will be similar to that described in the notes. Recall that intensity is related to the power spectrum of an acoustic pressure signal by

$$I_f = z_0 r_f(0) = \frac{z_0}{\pi} \int_0^\infty R_f(\Omega) d\Omega \tag{7.18}$$

a. Suppose that

$$f(t) = A\sin(\Omega_c t), \quad g(t) = h(t) * f[n]$$

where $H(\omega)$ is a filter of unknown shape. Find the power spectrum $R_g(\Omega)$ and intensity I_g of $g(t)$. Write I_g in terms of I_f, the intensity of $f(t)$.

b. Suppose that $v(t)$ is a band-stop noise signal. Specifically, suppose that

$$P_v(\omega) = \begin{cases} 0 & \Omega_c - \alpha < |\Omega| < \Omega_c + \alpha \\ N_0 & \text{otherwise} \end{cases}$$

If $w(t) = h(t) * v(t)$, find the power spectrum $R_w(\Omega)$. Show that the intensity I_w is proportional to

$$C(\alpha) = \int_0^{\Omega_c-\alpha} |H(\Omega)|^2 d\Omega + \int_{\Omega_c+\alpha}^{\infty} |H(\Omega)|^2 d\Omega$$

c. Suppose that

$$x(t) = f(t) + v(t), \quad y(t) = h(t) * x(t)$$

where $f(t)$ and $v(t)$ are as given in previous parts. Assume that $f(t)$ and $v(t)$ are uncorrelated, so that their spectra add in quadrature. What is the intensity I_y? Write I_y in terms of I_f, the intensity of $f(t)$.

d. Norm Needscash has signed up for a psychology experiment. In this experiment, Norm will listen to two noise bursts, and try to figure out which of the two noise bursts contained an embedded sine wave. Suppose that Norm can reliably hear the sine wave if and only if the signal+noise loudness level exceeds the noise-only loudness level by some minimum amount θ:

$$10\log_{10} I_y \geq 10\log_{10} I_w + \theta$$

The experimenter plans to keep the noise variance N_0 fixed, and vary the signal power I_f until Norm can reliably hear the sine wave. The experimenter will call this threshold power P, that is, Norm can hear the sine wave if and only if

$$I_f \geq P$$

Write P as a function of θ, N_0, $H(\Omega_c)$, and $C(\alpha)$.

e. The experiment described above is repeated for many different combinations of noise-gap width α and center frequency Ω_c. For each combination of α and Ω_c, Norm has to listen to many different signal levels, in order to determine the signal level P at which he can first hear the sine wave. Fortunately, the experimenter provides lots of coffee, and of course Norm is paid for his time.

Based on all of these data, the experimenter estimates $dP/d\alpha$. Show that, given $dP/d\alpha$, it is possible to estimate the shape of the filter $|H(\Omega)|^2$. Specifically, show that

$$\frac{|H(j(\Omega_c+\alpha))|^2 + |H(j(\Omega_c-\alpha))|^2}{|H(\Omega_c)|^2} \propto \frac{dP}{d\alpha}$$

Problem 7.6

Consider an acoustic pressure signal, measured at the oval window, with a missing fundamental frequency:

$$p[n] = \sum_{k=2}^{N_0/2} 2P_k \cos(kn\omega_0 + \theta_k)$$

a. Suppose that the cochlear filters $H_i(\omega)$, $i = 1, \ldots, N_0/2$, are conveniently defined as

$$H_i(\omega) = \begin{cases} 1 & \left(i - \frac{1}{2}\right)\omega_0 \le |\omega| < \left(i + \frac{1}{2}\right)\omega_0 \\ 0 & \text{else} \end{cases}$$

Remember that the basilar membrane velocity at position x_i is given by

$$U_i(\omega) = P(\omega)H_i(\omega)$$

Find $u_i[n]$.

b. Suppose that the brainstem correlogram computation is applied to the basilar membrane velocities directly, without half-wave rectification, i.e.

$$r_i[n] = \frac{1}{N} \sum_{\tau=n}^{N-1} u_i[\tau]u_i[\tau - n]$$

Find $r_i[n]$. Assume that N is large enough that there are no significant cross-terms, and that the autocorrelation of $\cos(kn\omega_0)$ is roughly $0.5w_T[n]\cos(kn\omega_0)$.

c. Find the lowest value of the delay n, $n > 0$, such that $r_i[n]$ has a peak in every band except $i = 1$.

Problem 7.7

a. This lab will use three test signals. All test signals should be downsampled to about 6.4kHz; higher sampling rates don't make much sense, because inner hair cells only phase lock up to about 2kHz.

The first test signal should be a sine wave at some frequency between about 100Hz and 600Hz. The second test signal should be a periodic signal with the same fundamental frequency as the sine wave, but with the first harmonic component missing (like the signal in problem 9.1; include at least three higher harmonics). The third test signal is any audio clip of interest to you, downsampled to about 6.4kHz.

Construct the three test signals. Compute and plot the long-term power spectra of all three signals, using the "periodogram" spectral estimator. Plot all three power spectral estimates in subplots of the same figure; label the axes in Hertz. The periodogram spectral estimator is created by averaging the short-time power spectral estimates in a series of M consecutive frames, with frame length N and frame skip parameter S:

$$\hat{R}_x(\omega) = \frac{1}{M}\sum_{m=1}^{M}\hat{R}_x(\omega,m) \tag{7.19}$$

$$\hat{R}_x(\omega,m) = \frac{1}{N}|X(\omega,m)|^2 \tag{7.20}$$

$$X(\omega,m) = \sum_{n=0}^{N-1} x[mS+n]e^{-j\omega n} \tag{7.21}$$

b. Write a function `F=cochlea(f)` that computes a model of inner hair cell voltages in response to an acoustic signal `f`. The function should do the following things:

- Bandpass filter `x`. You may use either uniformly spaced, non-overlapping filters (with bandwidths of about 100Hz) or ERB-spaced filters (with bandwidths equal to about one ERB). Uniformly spaced filters may be designed by creating a prototype 50Hz FIR lowpass filter in matlab, and then modulating your prototype filter up to each of the different center frequencies. ERB-spaced filters may be created, for example, by defining the impulse responses to be gammatone functions, and by adjusting the center frequency and bandwidths of the gammatone functions according to the ERB scale.
- Half-wave rectify each bandpass-filtered signal.

Each row of the matrix `F` should contain the rectified band-pass filter output from one hair cell channel.

Use `cochlea` to estimate hair cell voltages in response to the sine wave, the missing-fundamental wave, and the audio clip. Create a plot with three subplots. From each of these different `F` matrices, pick out one channel with relatively high energy, and plot about 100ms from that channel, as a function of time. For the third plot (audio clip), be sure to pick out a 100ms segment with high energy. Label the axes in milliseconds or seconds, *not samples*. Provide the following information in the title of each plot: the type of signal, and the band center frequency.

c. Write a function `[CG, CG0]=correlogram(F)` that computes the correlogram, `CG`, given estimates of the inner hair cell voltages. The correlogram `CG` should be a rank-3 array of the form `CG(t,b,m)`, in which the dimensions are autocorrelation lag (`t`), band number (`b`), and frame number (`m`). The energy matrix `CG0` should be a matrix of the form `CG0(b,m)`, in which the dimensions are band number (`b`) and frame number (`m`). The function `correlogram` should do the following things:

- Window each of the rectified signals using `enframe`. Use non-overlapping frames of about 33.3ms.
- Compute the autocorrelation of each frame of each bandpass filtered signal. Set `CG0` equal to the zero-lag autocorrelation $CG(0)$. Set the correlogram `CG` equal to the half of the autocorrelation function for which $\tau > 0$; discard the other half of the autocorrelation function.

Compute the correlogram of the sine wave, and of the missing-fundamental signal. Choose any representative frame from the sine wave correlogram, and compute the whitened correlogram of that frame, i.e. if NCOR is the number of autocorrelation lags, you could write:

```
WCG = CG(:,:,m)./repmat(CG0(:,m),[1 NCOR]);
```

Create an image plot of this whitened correlogram, e.g., if `TAU` contains the autocorrelation lags in milliseconds, and `CF` contains the center frequencies in Hertz, you could type

```
imagesc(TAU,CF,64*WCG); axes xy
```

Do the same with the missing-fundamental wave, and hand in both plots. The two plots should have some similarities, and some differences.

d. Create a movie displaying the correlogram of your audio clip as a function of time.

First, normalize the entire correlogram according to the maximum energy—compute the normalized correlogram NCG, as

$$NCG(t,b,m) = \frac{CG(t,b,m)}{\max_t \max_m CG(t,b,m)} \tag{7.22}$$

Note this is NOT the same as computing a whitened correlogram—the entire correlogram should increase and decrease from frame to frame. You can probably compute this using commands such as

```
MAXENERGY = max(CG0);
NCG = CG./repmat(MAXENERGY,[NCOR 1 NFRAMES]);
```

Then create a movie using code something like this (but instead of “strange animal noises,” be sure to write the title of your own audio clip):

```
mov=avifile('correlogram.avi','fps',30);
for m=1:size(CG,3),
```

```
    image(TAU,CF,64*NCG(:,:,m));
    xlabel('Autocorrelation lag (seconds)');
    ylabel('Center frequency (Hz)');
    title('Correlogram of strange animal noises');
    frame=getframe(gca);
    mov=addframe(mov,frame);
  end
  mov=close(mov);
```

Play back the AVI file using windows media player or the matlab `aviread` function.

When you're satisfied with the movie quality, clip out 4-9 frames, showing the structure of the correlogram as a function of time. Plot them in a figure, and attach this figure to your assignment.

Problem 7.8

Model the head as a sphere, 10cm in radius. Plot the inter-aural time delay, in milliseconds, as a function of the angle of incidence, for $0 \leq \theta \leq \pi/2$.

Problem 7.9

An approximate formula relating frequency to position on the basilar membrane is given by

$$x(f_c) \approx 35\text{mm} - 11\text{mm} \ln\left(1 + \frac{46 f_c}{f_c + 14700\text{Hz}}\right) \tag{7.23}$$

This formula is actually not based directly on physiological measurements. Instead, it is based on the physiological rule of thumb which says that one equivalent rectangular bandwidth is about 1mm on the basilar membrane—in other words, vibration of a position x on the basilar membrane necessarily includes vibration of positions between $x - 0.5\text{mm}$ and $x + 0.5\text{mm}$. The form of Eq. 7.23 is too detailed to have been derived from physiological experiments; it is, rather, derived from psycho-acoustic measurements of ERB, of the type that we will consider in lecture 11. In the mean time, this problem will consider some of the implications of Eq. 7.23.

a. Approximate Eq. 7.23 as a log-linear form, i.e.,

$$\text{ERB} = 11 \ln\left(1 + \frac{f}{F_r}\right) \tag{7.24}$$

for some reference frequency F_r. What is F_r?

b. Divide Eq. 7.24 into two frequency regions, a “linear” region and a “logarithmic” region, using the following approximation:

$$\ln(1+z) \approx z \quad \text{for } z < e-1 \tag{7.25}$$

Sketch $x(f_c)$ as a function of f_c over the audible frequency band (about 0 to 20kHz). Label the linear and the logarithmic regions.

c. Use the linear approximation from part (b) to estimate the equivalent rectangular bandwidth (ERB) of a cochlear filter at relatively low frequencies. Using your approximation from part (b), you should find that ERB is independent of f_c for low frequencies.

d. At very high center frequencies ($f_c \gg F_r$), how many millimeters on the basilar membrane are traversed by one octave in frequency? That is, if you double the center frequency, how many millimeters do you move?

e. Invert your answer to part (d) in order to estimate the fraction of an octave that is spanned (in the ERB sense) by a cochlear filter at high frequencies. How many semitones is one ERB?

f. Invert your answer to part (a) or part (b) (whichever you like) in order to create a formula $f_c(x)$ that specifies the center frequency f_c corresponding to any position x on the basilar membrane.

Problem 7.10

Compute the number of critical bands along the length of the cochlea. Assume that there are $N_c = 2.75W(X)$ CBs, where W(X) is the width of the cochlea, and is given by $W(X) = A + B * exp((X-L)/C)$, L is the length of the BM, C is the characteristic length, and A and B are constants that depend on species.

Chapter 9

Human Speech Recognition

As a general topic, auditory perception can be divided a number of ways. From the standpoint of communication, one separation might be between classical auditory psychophysics, on the one hand, and the recognition of acoustic signals presented within a linguistic framework, on the other. The former relates principally to the abilities and limitations of the hearing organ as a mechano-neural transducer of all acoustic signals. The latter bears mainly upon the identification and classifications of auditory patterns which are significant within the communicative experience of the listener.

Classical auditory psychophysics strives to discover the "resolving power" of the hearing mechanism. Discrimination is usually examined along fundamental dimensions of the stimulus–usually along only one dimension at a time. The measurements are generally conducted under conditions which are most favorable for making the relevant discriminations, that is, differential discriminations or close comparisons. Differential thresholds for dimensions such as intensity and frequency fall into this classification. Intuitively one feels that large neural storage and complex central processing probably are not brought into play in such detections. The measures more likely reflect the capacity of the transducer and the peripheral neural net to preserve details about a given stimulus dimension. The discussion in Chapter 7, for example, touched upon these properties of the peripheral system. The apparent relations between physiological and psychoacoustic response were analyzed for several stimulus cases. The acoustic signals were of the "classical" type in that they were either temporally punctate or spectrally simple, or both.

Speech, on the other hand, is a multidimensional signal that elicits a linguistic association. For it to be an effective communication code, some sort of absolute perceptual categorization must be made of its content. That is, the signal must be broken down into a finite number of discrete message elements. The "size" of these perceptual elements, and the manner in which they are processed to yield the percept, are questions of considerable debate and not little speculation. Our present knowledge brings us nowhere near a good understanding of the process. Theorizing about speech perception–cloaked in all of its linguistic and over-learned functions–abounds with pitfalls. An even larger problem, perhaps, is reconciling physiological, psychophysical and linguistic factors. As in other difficult situations, it is tempting to push back to some still-higher center the final decision-making process that is the real seat of perception.

Although a complete theory of speech perception remains in the future, a good deal can be said about auditory discrimination. Some of the "classical" measurements relate strongly to signal dimensions important to speech–even though the measurements are made outside of linguistic or contextual frames. In addition, a respectable amount of information has been accumulated on the acoustic cues associated with synthetic approximants to simple speech elements–for example, syllables and phonemes.

From the practical point of view, articulation tests and intelligibility measures based upon absolute recognition of sentences, words, syllables, and isolated phonemes can be used to good effect in evaluating transmission facilities. For a given processing of the voice signal, these tests often help to identify factors upon which perception depends (although they serve poorly, if at all, in supplying a description of the perception process itself). Under certain conditions, the so-called articulation index can be used to compute intelligibility scores from physical measurements on the transmission system. Still ancillary to intelligibility testing, some data are available on the influences of linguistic, contextual and grammatical constraints. Contrariwise, measures of the prosodic and quality features of speech are not well established.

The present chapter proposes to circumscribe some of these problems. In particular, the aim is to indicate current level of understanding in the perception of speech and speech-related sounds.

9.1 Differential vs, Absolute Discrimination

Classical psychophysics generally deals with discriminations made in close comparison. Speech perception, on the other hand, seems more likely an absolute classification of an acoustic signal. Can the former provide useful information about the latter, or vice versa?

Man is highly sensitive to differences in the frequency or intensity of sounds presented for comparison. Under certain conditions the threshold for detecting a difference in the frequencies of two successively presented pure tones may be as small as one part in 1000 (Rosenblith and Stevens [1953]). The threshold for detecting a difference in intensity may be less than one db (Riesz [1928]). On the basis of comparative judgments, it has been estimated that the normal listener can distinguish about 350000 different tones (Stevens and Davis [1938]).

Contrasting with this acute differential sensitivity, a listener is relatively inept at identifying and labelling sounds presented in isolation. When equally-loud pure tones are presented individually for absolute judgment of frequency, listeners are able to accomplish perfect identification among only five different tones (Pollack [1952]). This identification corresponds to an information transfer of about 2.3 bits per stimulus presentation. If, however, the sound stimulus is made multidimensional, for example by quantizing it in frequency, loudness, duration, etc., the correct identifications increase, and the information transferred may be as high as five to seven bits per stimulus presentation (Pollack and Ficks [1954]). This rate is equivalent to correct identification from an ensemble of from 32 to 128 stimuli.

It is clear that absolute and differential discriminations yield substantially different estimates of man's informational capacity. The former suggest that few subdivisions along a given dimension can be identified, whereas the latter indicate that a much larger number may be discriminated. The differential measure, however, reflects the ability to discriminate under the most favorable circumstances for detecting a difference (namely, a close comparison usually along a single dimension). In a sense, it represents an upper bound on the resolving ability of the perceptual mechanism.

So far as absolute judgments are concerned, the differential estimate of discrimination is an optimistic one. The probability is extremely small that stimulus quantizations as small as a difference limen[1] could ever be detected absolutely. Even so, differential measures quantify perception in a "clinical" way, and they place a rough ceiling (albeit over-optimistic) on the ability to detect changes in a signal. In any speech processing system, fidelity criteria based upon differential discriminations would be expected to be very conservative. Lacking more directly applicable measures, however, they can often be useful in estimating the performance and channel capacity requirements of a transmission system (Flanagan [1956b]).

9.2 Differential Discriminations Along Signal Dimensions Related to Speech

The results of Chapters 4 and 7 suggest that significant dimensions for the speech signal might be defined in either the acoustic or articulatory domains. Both domains have perceptual correlates. The analyses in Chapter 4, for example, attempted to separate the properties of vocal transmission and excitation. Perceptually-important acoustic dimensions of the system function are those of the mode pattern–that is, the complex frequencies of the poles and zeros of the transmission. Alternatively, the same information is specified by the bandwidths and frequencies of the maxima and minima of the amplitude spectrum and by the values of the phase spectrum. In a similar manner, relevant dimensions of the excitation source for voiced sounds are intensity, fundamental frequency and perhaps spectral zero pattern (or equivalently, glottal wave asymmetry and duty factor). For the unvoiced source, intensity and duration are significant dimensions.

---

[1] The terms difference limen (DL) and just-noticeable difference (JND) are synonomous with differential threshold or just-discriminable change.

Auditory sensitivity to some of these factors–divorced of any linguistic or contextual frame–has been measured in psychoacoustic experiments. For example, data are available on just-discriminable changes in formant frequency, fundamental frequency, over-all intensity and formant bandwidth. Without going into the details of any specific experiment, the nature of the results can be summarized.

9.2.1 Limens for Vowel Formant Frequencies

Using synthetic vowel sounds generated by a terminal-analog synthesizer (see Section 11.4, Chapter 11), just-discriminable changes in the frequencies of the first and second formants have been measured (Flanagan [1955b]). The synthesizer was operated as it would be used in a formant-vocoder system. Although the difference limens (DL's) depend to an important extent upon the proximity of the formants, they are found to be on the order of three to five percent of the formant frequency[2].

9.2.2 Limens for Formant Amplitude

The results of Chapter 4 and 7 show that the relative amplitude of a given formant in the speech signal is a function of several factors, among them formant frequency, vocal damping, transmission zeros and excitation characteristics. One measure of the differential discriminability of formant amplitude has been made with a parallel-connected, terminal-analog synthesizer (Flanagan [1957a]). The intensity limen for the second formant of a near-neutral vowel (/æ/) is found to be about 3 db.

A related measurement of the limen for over-all intensity of a synthetic vowel gives a value of about 1.5 db (Flanagan [1955]). Because the first formant is usually the most intense formant in vowel sounds, the over-all figure might be taken as a rough estimate of the first-formant intensity limen.

Another experiment determined the intensity limens for single harmonic components of synthetic vowels (Flanagan [1965]). Values found for intensity changes at the first and second formant frequencies support well the values just mentioned. Intensity limens for harmonic components located in spectral "valleys" can be quite large, as much as $+13$db to $-\infty$ db, i.e., complete absence.

9.2.3 Limens for Formant Bandwidth

Apparently, no direct measures of the discriminability of changes in formant bandwidth, or damping, have been made on synthetic vowels. However, some related

---

[2] This experiment considered changes in the frequency of only one formant at a time. In real speech–and in formant-coding of speech–the formants usually move simultaneously. A relevant and practically-useful extention of the experiment might be the determination of "DL solids" in F1-F2-F3 space. Proximity effects of the formants should, in general, give these "solids" ellipsoidal shapes. Similar comments about discrimination of simultaneous changes in signal dimensions apply in several of the following experiments.

measurements, and their extrapolations, suggest what might be expected.

Stevens (Stevens [1952]) measured the descriminability of changes in the tuning and damping of a single electrical resonator. The resonator was excited by periodic pulses at a fundamental frequency of 125 Hz. The output signal was therefore representative of a one-formant vowel. In general, changes on the order of 20 to 40% in formant bandwidth were just-discriminable.

Also, the results of Chapter 4 show that the amplitude of a formant peak is inversely related to formant damping. The 1.5 db figure found for the amplitude limen of the first formant corresponds to a bandwidth change of about 20%. Similarly, the 3 db figure for the second formant corresponds to a bandwidth change of about 40%[3].

9.2.4 Limens for Fundamental Frequency

Following an experimental procedure similar to that used with the formant measurements, a difference limen has been measured for the fundamental excitation frequency of synthetic vowel sounds (Flanagan and Saslow [1958]). For synthetic vowels appropriate to a man, the fundamental-frequency limen is found to be about 0.3 to 0.5 per cent of the fundamental-frequency. From this and the previously-mentioned measurements, a hierarchy in frequency acuity emerges. The formant-frequency limen is an order of magnitude more acute than the formantbandwidth limen, and the fundamental-frequency limen is an order of magnitude more acute than the formant-frequency limen.

9.2.5 Limens for Excitation Intensity

For a given glottal wave shape and vocal transmission, the over-all intensity of a voiced sound is directly proportional to the amplitude of the glottal pulse. As mentioned previously, a measure of the limen for over-all vowel intensity gives a value of about 1.5 db.

Similarly, the over-all intensity of an unvoiced sound is directly related to the amplitude of the unvoiced source. Fricative consonants are relatively broadband, noise-excited, continuant sounds. The discriminability of changes in their over-all amplitude might be expected to be somewhat similar to that of wide-band noise. Intensity limens have been measured for the latter (Miller [1947]). They are found to be of the order of 0.4 db for sensation levels above 30 db. The minimum perceptible intensity change is therefore about 5%. Only a few fricative consonants have relatively flat spectra, but the figure might be used as an order-of-magnitude estimate. Experience with speech synthesis confirms that it is a conservative figure.

---

[3] Another multidimensional DL of interest might be that for simultaneous changes in formant bandwidth and frequency. In other words, one might determine DL "areas" in the complex-frequency plane for vocal tract poles.

9.2.6 Limens for Glottal Zeros

The differential discriminability of changes in the spectral zero pattern of the vocal cord source (see Section 11.6, Chapter 11), or in the detailed spectrum of the glottal wave, have, to the author's knowledge, been observed only informally (Flanagan [1961]). The glottal source may contribute significant factors to speech quality and to speaker recognition. Therefore, liminal measures of parameters such as the duty factor and asymmetry of the glottal wave could be valuable in establishing bounds on their importance to speech naturalness.

It is clear that if complex source zeros lie far enough away from the $j\omega$-axis of the frequency plane they have negligible effect on signal quality. One experiment in which only gross features of the source spectrum and waveform were preserved suggests that many temporal and spectral details are unimportant to quality (Rosenberg [1965]) (see Section 11.6.1).

9.2.7 Discriminability of Maxima and Minima in a Noise Spectrum

The vocal tract transmission for fricative consonants, like other sounds, is characterized by certain poles and zeros. Broadband noise excitation is filtered by this transmission. Some of the poles and zeros (and their related spectral maxima and minima) are perceptually significant. Others are not. One measurement considered the differential discriminability of a single peak or valley in an otherwise flat noise spectrum (Malme [1959]). A single pole and zero filtering of a broadband noise was used to produce the spectral variations shown in the insert of Fig. 9.1. The equivalent complex frequencies (half-power bandwidths vs. center frequencies) of the irregularities which were just-detectable from the flat spectrum are also plotted in Fig. 9.1. The db numbers next to the points are the just-perceptible peak heights and notch depths, respectively. These data indicate that, at least in a flat-noise surround, spectral peaks with Q's (i.e., ratios of center frequency to bandwidth) less than about 5, and spectral notches with Q's less than about 8 are not differentially perceptible.

The results suggest, therefore, that many of the small spectral irregularities seen in a fricative consonant such as /f/ are not perceptually significant. In the same vein, certain spectral peaks such as in /s/ or /ʃ/ are of course significantly different from a flat spectrum. Synthesis of fricative consonants has been demonstrated by representing the spectrum in terms of two poles and one zero (Heinz and Stevens [1961]). Appropriate Q's for the poles are found in the range of about 5 to 13. For the zero, Q's of the order of 2 to 4 appear appropriate. The suggestion is, therefore, that to the extent the results in Fig. 9.1 can be applied, the poles are more significant perceptually than the zero, the latter apparently having importance only in contributing to the gross spectral shape. This appears to be the case, for the zero has been found to be relatively noncritical of frequency position and often can be

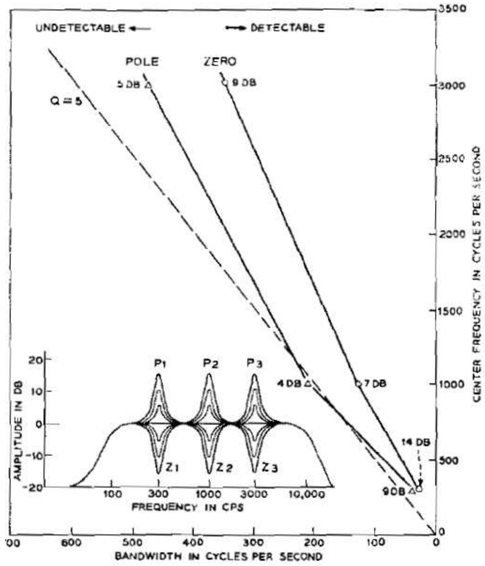


Figure 9.1: Detectability of irregularities in a broadband noise spectrum. (After (Malme [1959]))

placed automatically about an octave below the first pole (Heinz and Stevens [1961]).

A similar discrimination measurement has been made for a noise spectrum with exactly periodic maxima–that is, for a comb filtering of the noise (Atal and Schroeder [1956]). The objective was to investigate the perceptual effects of irregularities in the frequency response of rooms. The comb-filtered noise was differentially compared with white noise of equal power, and a limen was obtained for the minimum detectable periodic irregularity. The minimum detectable ratio of maximum-to-minimum spectral amplitude was found to be about one db. This figure is in close agreement with the intensity limen measured for white noise (see Section 9.2.5).

The results of this same experiment provide information on the weighting function used by the ear in performing its short-time spectral analysis. The weighting function deduced from the measurements is approximately exponential in form, with an initial slope corresponding to a time constant of 9 msec. This latter figure compares favorably with the time constant deduced for loudness measurements on periodic clicks (see Section 7.3.3, Chapter 7).

9.2.8 Other Close-Comparison Measures Related to Speech

A number of other psychophysical measurements relate, more or less strongly, to differential perception along speech dimensions. Several of these can be mentioned to illustrate the diverse nature of the data.

One experiment measured the perception of a single, time-varying formant (Brady *et al.* [1961]). A continuously-tunable resonant circuit was excited by five equi-spaced

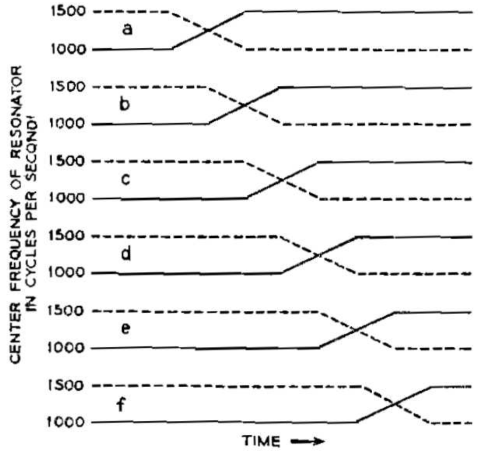


Figure 9.2: Frequency paths and excitation pattern for a simulated time-varying formant. Rising and falling resonances are used. The epochs of the five excitation pulses are shown. (After (Brady *et al.* [1961]))

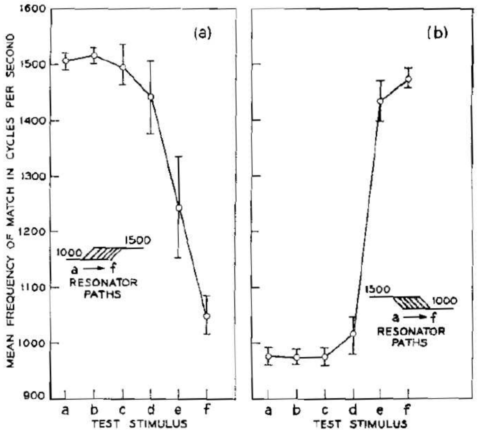


Figure 9.3: Results of matching a nontime-varying resonance to the time-varying resonances shown in Fig. 9.2. Mean values are plotted. The vertical lines indicate the standard deviations of the matches. (After (Brady *et al.* [1961]))

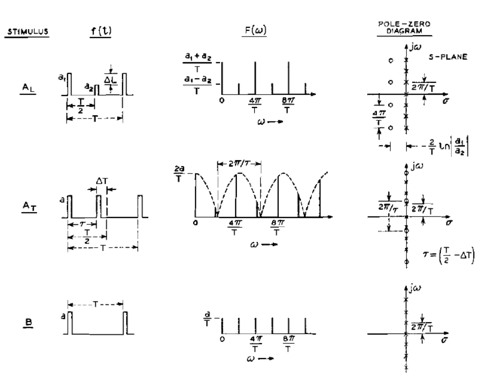


Figure 9.4: Periodic pulse stimuli for assessing the influence of amplitude and time perturbations upon perceived pitch. The left column shows the time waveforms of the experimental trains; amplitude variation (A_L), time variation (A_T), and the standard matching train (B). The second column shows the corresponding amplitude spectra, and the third column shows the complex-frequency diagram. (After (Flanagan *et al.* [1962b], Guttman and Flanagan [1962]))

pitch pulses. The pulses were produced at a rate of 100Hz. During the excitation, the tuning of the resonator was moved between 1000 and 1500Hz, according to the rising and falling trajectories shown in Fig. 9.2. The formant transitions were accomplished in 20 msec. To examine how the varying formant frequency is perceived, listeners were asked to adjust the frequency of a nontime-varying resonance until it sounded as much like the varying one as possible. Typical results of the matches are shown in Fig. 9.3. The test stimuli labelled a, b, c, d, e, and f correspond to those diagramed in Fig. 9.2. The data show a strong tendency to set the steady resonance to a frequency corresponding to the final value of the varying formant, particularly when the formant change occurs near the beginning of the sound. The tendency to match the final frequency appears somewhat stronger for stimuli in which the resonant frequency ascends.

In a different vein, a temporal fine structure is known to exist in the glottal source. The shape and periodicity of the glottal pulse is subject to various perturbations which may be significant to speech quality. In a condition known as diplophonia, for example, alternate glottal pulses may be of different size (Smith [1958]). Similarly, successive glottal periods may vary in duration. To quantify this effect, one study analyzed the durations of 7000 pitch periods of real speech (Lieberman [1961]). In successive samples of three periods each, the variation in period was greater than 0.1 msec in 86% of the cases. In 20% of the cases the duration difference between periods was greater than 0.6 msec, and in 15% it was greater than 1.0 msec. In 38% of the cases the periods were alternately long and short. Adjacent periods were not correlated, but alternate periods were highly correlated.

As one step toward trying to understand the possible perceptual correlates of these factors, a preliminary investigation has examined the effects upon perceived pitch of systematic differences in amplitude and timing of an otherwise periodic pulse train (Flanagan *et al.* [1962b], Guttman and Flanagan [1962]). Among the conditions considered were the pulse wave forms shown in the left-hand column of Fig. 9.4. Starting with exactly periodic trains (of period T12) , alternate pulses in the train were changed incrementally either in amplitude level (Stimulus A_L) or in time of occurrence (Stimulus A_T). The effect upon pitch was assessed by having listeners adjust the frequency of a uniform, periodic train (Stimulus B) until its pitch matched that of the perturbed train. As either the amplitude difference (ΔL) or the time difference (ΔT) increases, a point is soon reached where the pitch drops by an octave.

The second column of Fig. 9.4 shows the frequency spectra of the amplitude-varied stimulus (A_L) , the time-varied stimulus (A_T), and the standard matching stimulus (B). The third column of the figure shows the corresponding pole-zero diagrams for the three periodic trains. Notice that for the A_L signal the relative amplitudes of adjacent spectral lines are dependent only upon the pulse amplitudes a_1 and a_2. For A_T on the other hand, the spectral amplitudes are conditioned by the fundamental period T and by the cycloidal envelope which, in turn, is determined by the interval T.

Median matches made by a number of listeners for the ΔL and ΔT conditions are shown in Fig. 9.5a and 7.5b, respectively. In both plots the parameter is the pulse rate of the A stimulus (i.e., twice its fundamental frequency). The results in Fig. 9.5a for ΔL show that, over the frequency range appropriate to the human voice fundamental, an amplitude difference ΔL of about 6 to 8 db, or greater, will produce an octave reduction in the perceived pitch. In the same fashion, the ΔT data in Fig. 9.5b show that in the range of the voice fundamental (i.e., about 100Hz and above) a time shift ΔT on the order of 0.1 or more, will produce an octave reduction in pitch.

9.2.9 Differential Discriminations in the Articulatory Domain

The acoustic dimensions considered for the speech and speech-like signals in the preceding discussion have counterparts in the articulatory domain. However, the acoustic and articulatory relations do not generally exist in one-to-one correspondence. For example, a change in a constriction size, or in its location, alters not only one formant frequency, but in general all of them (see Fig. 4.36, Chapter 4). It is therefore difficult to interpret, say, limens for formant frequency and amplitude in terms of just-discriminable articulatory changes. One can, nevertheless, make some simple observations about the links between the domains.

The just-discriminable changes in formant frequency were found to be about three to five per cent. For a straight pipe the formants are approximately

$$F_n = \frac{(2n - l)c}{4l}, \quad n = 1, 2, \ldots$$

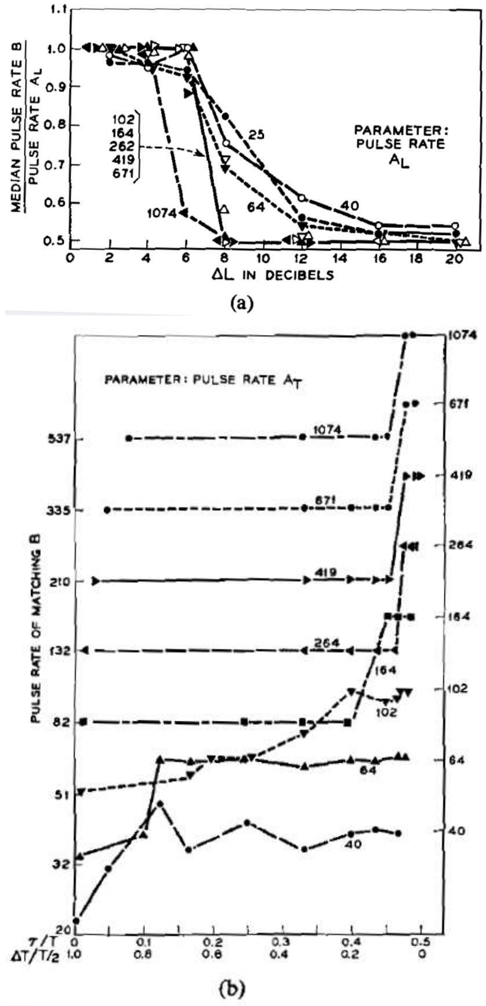


Figure 9.5: Results of matching the pitch of a uniform pulse train (B) to that of: (a) a periodic train (A_L) whose alternate pulses differ in amplitude by ΔL and (b) a periodic train (A_T) whose alternate pulses are shifted in time by ΔT. In both cases the parameter is the pulse rate of the A stimulus. (After (Flanagan *et al.* [1962b], Guttman and Flanagan [1962]))

The sensitivity of the mode frequencies to length changes is

$$\partial F_n/\partial l = -\frac{(2n-l)c}{4l^2}, \quad \text{or} \quad \frac{F_n}{\Delta F_n} = -\frac{l}{\Delta l},$$

so that a given percentage change in the tract length l produces the same percentage change in the formant frequencies. The DL for tract length might therefore be expected to be roughly comparable, percentage-wise, to the formant frequency DL. By referring to Fig. 4.36, Chapter 4, one can see other, more complex correspondences between formant changes and articulatory changes.

Another simple example is the sensitivity of the mode damping for a straight pipe to changes in the mean glottal area (see Eq. (4.41)]. Assume for simplicity that the equivalent glottal impedance is purely resistive and is produced only by kinetic factors, that is,

$$R'_g = \frac{(2\rho P_{s0})^{\frac{1}{2}}}{A_0}$$

[using the notation of see Eq. (4.18)]. The pole dampings (i.e., real parts) are given by

$$\sigma_n \approx -\left(\alpha c + \frac{Z_0 c}{l R_g}\right)$$

or

$$\sigma_n \approx -\left[\alpha c + \frac{c Z_0 A_0}{l(2\rho P_{s0})^{\frac{1}{2}}}\right]$$

[see Eq. (4.41)]. The sensitivity of the damping with respect to mean glottal area is then

$$\frac{\partial \sigma_n}{\partial A_0} \approx -\frac{c Z_0}{l(2\rho P_{s0})^{\frac{1}{2}}},$$

or the change in mode damping is approximately proportional to the change in mean glottal area.

9.3 Absolute Discrimination of Speech and Speech-Like Sounds

Most efforts to establish the acoustic cues for speech-sound recognition have been absolute identification experiments. The test stimuli have generally been synthetic versions of phoneme-length and syllable-length utterances. This approach presumably keeps the stimuli simplified to noncontextual situations where only the physical properties of the specific signal influence the percept. At the same time it may permit association of a linguistic structure, and the perceptual responses are usually interpreted within this frame of reference.

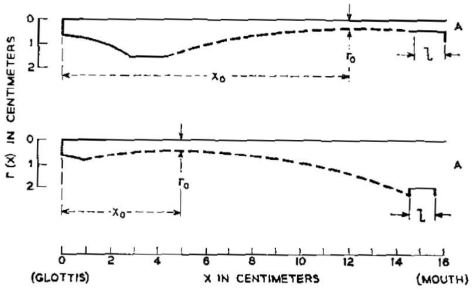


Figure 9.6: Three-parameter description of vowel articulation. r_0 is the radius of the maximum constriction; x_0 is the distance from the glottis to the maximum constriction; and A/l is the ratio of mouth area to lip rounding. (After (Stevens and House [1955]))

9.3.1 Absolute Identification of Phonemes

A relatively small number of experiments has dealt solely with isolated phonemes. One study–using a transmission-line vocal tract analog–investigated articulatory configurations appropriate to vowels. It tested a simple three-number articulatory description of vowel production (Stevens [1955], Stevens and House [1955], House [1956]). The three-number scheme for describing vowel articulation is illustrated for two configurations in Fig. 9.6. The three parameters used to describe the vocal shape are the radius of the maximum constriction, r_0; the distance from the glottis to the constriction, x_0; and the ratio of mouth area to lip rounding, A/l. The radius of the dashed portion of the tract is described by the function

$$r(x) = [0.025(1.2 - r_0)(x - x_0)^2 + r_0],$$

where the lengths are in centimeters.

An electrical transmission line simulated the configurations and synthesized the sounds. Isolated vowels,500 msec in duration, were judged absolutely by listeners and placed into nine English vowel categories. Pitch was monotonically inflected from 120 to 140Hz. The listener responses in terms of articulatory parameters are illustrated for one value of constriction in Fig. 9.7. The two response contours indicate agreement among 50% and 75% of the responses, respectively. The Peterson and Barney data for natural vowels uttered by men (see Fig. 5.10, Chapter 5), when transformed into the same articulatory coordinates, are given in Fig. 9.8. The two plots show that, except for small differences, the three number description does surprisingly well in providing a unique specification of the vowels.

A somewhat similar experiment on synthesis and perception has been carried out for Japanese vowels (Nakata and Suzuki [1959]). In this experiment, however, the sounds were produced by a terminal-analog synthesizer, and the idea was to find the synthetic formant patterns appropriate to the vowels.

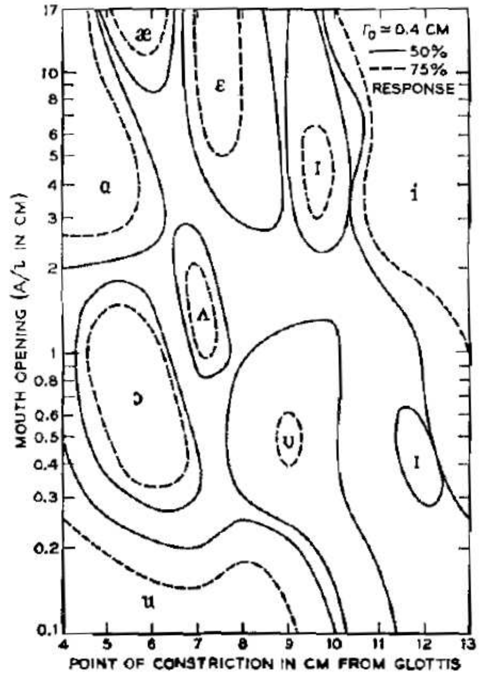


Figure 9.7: Listener responses to isolated synthetic vowels described by the 3-parameter technique. One value of constriction is shown. Two levels of response corresponding to 50 and 75% agreement among subjects are plotted. (After (House [1955]))

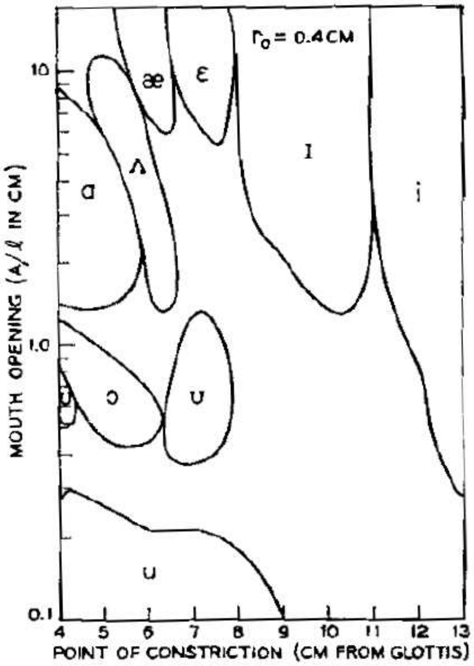


Figure 9.8: Formant frequency data of Peterson and Barney for 33 men transformed into the 3-parameter description of vowel articulation. (After (House [1955]))

The same transmission-line analog–but with attached nasal tract-has been used to study the perception of nasal consonants (House [1957]). Isolated, 500 msec representations of nasal consonants were synthesized and presented to listeners for absolute judgment. The permissible response categories were the three nasal consonants /m,n, and ŋ/. The articulatory description used for synthesis was similar to that described in the preceding discussion on vowels, but with the additional specification of the velar coupling. Typical confusion matrices of responses (to articulatory configurations which were determined by pre-tests to be representative nasal consonant stimuli) are shown in Table 9.1a.

While the responses to the synthetic nasal consonants do not look particularly decisive, they do compare favorably with similar measurements on natural nasal consonants (Malecot [1956]). A confusion matrix for the latter are shown in Table 9.1b. In this case the synthetic nasals are discriminated better than the natural ones! In view of the high functional load that nasals, particularly /n/, carry in connected speech (see Table 1.2, Chapter 1), the low discrimination scores suggest that transitions, both from and into adjacent sounds, may be highly important to nasal perception.

Table 9.1: Listener responses to synthetic and natural nasal consonants

a) Synthetic				b) Natural			
Stimulus	Response %			Stimulus	Response %		
	m	n	ŋ		m	n	ŋ
m	81	11	8	m	96	4	0
n	33	61	6	n	42	56	2
ŋ	20	18	62	ŋ	60	28	12

a) Synthetic: Mean correct response = 68%.
b) Natural: Mean correct response = 55%.

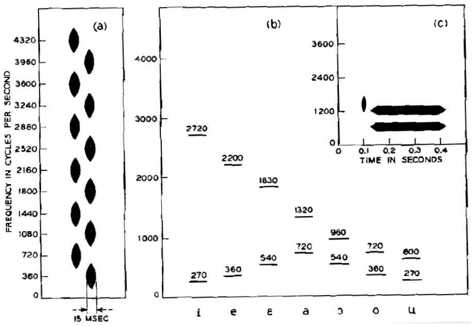


Figure 9.9: Stimulus patterns for determining the effect of noise-burst frequency on the perception of voiceless stop consonants: (a) frequency positions of the noise bursts, (b) formant frequencies of the two-formant vowels; (c) one of the synthetic consonant-vowel syllables formed by pairing a noise burst of (a) with a two-formant vowel of (b). (After (Cooper *et al.* [1952]))

9.3.2 Absolute Identification of Syllables

A substantial amount of research has considered the perception of isolated syllables. The effort has aimed mainly at discovering the acoustic cues important to phoneme recognition. Central to the objective is the determination of the separate contribution each acoustic variable makes to speech perception, as well as an understanding of how the contributions combine in the total percept. Much of the work points up the importance of acoustic environment upon perception; that is, the perception of a given phoneme can be strongly conditioned by its neighbors.

Among the leaders in this work has been the group at the Haskins Laboratories. Many of their experiments have used synthetic syllables generated by the pattern-playback machine. The operation of this synthesizer has been described in Chapter 11, and it is shown in Fig. 11.5. As explained in Section 11.3.1, the device synthesizes sound from data displayed as a conventional time-frequency-intensity spectrogram.

The nature of the experimentation is exemplified in consonant identification tests

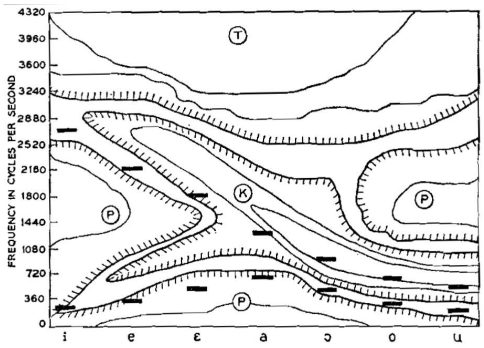


Figure 9.10: Listener responses to the synthetic consonant-vowel syllables shown in Fig. 9.9. (After (Cooper *et al.* [1952]))

on CV syllables. The consonant used is either a voiced or voiceless stop. If it is voiceless (i.e., /p,t,k/), one of the variables that seems to enable listeners to differentiate the sounds is the position along the frequency scale of the brief burst of noise constituting the stop release. To isolate this particular cue and to determine its role in perception, schematized stop-vowel syllables such as shown in Fig. 9.9c were synthesized (Cooper *et al.* [1952]). The noise burst (the small vertical ellipse in Fig. 9.9c) was constant in bandwidth and duration, and the vowel was a two-formant vowel that was maintained steady throughout the syllable. Combinations of noise bursts and vowel formants shown in Fig. 9.9a and b, respectively, produced the test ensemble.

The syllables were presented in isolation to listeners who were asked to judge the initial consonant either as /p,t or k/. The identifications, according to noise-burst location and vowel, are shown in Fig. 9.10. The contours indicate approximately equal response percentages, with the small contours representing the higher percentage response.

For these particular syllables, the one frequency variable (namely frequency of noise burst) appears adequate to distinguish the three consonants. High frequency bursts are heard as /t/ for all vowels. For /p/ and /k/ the identification depends not only upon frequency of burst but also on its relation to the vowel. Bursts on a level with the second formant, or slightly above, are heard as /k/; otherwise they are heard as /p/. The conclusion is advanced that the perception of these stimuli–and perhaps their spoken counterparts–requires the CV combination (that is, the syllable) as a minimal acoustic unit. Without information on the following vowel, the consonant percept may be equivocal.

A second cue important in the perception of stop-consonants is the stop-vowel formant transitions. One relevant question is how might this cue and the former one of burst position contribute singly, and how might they combine. To get some indication of the answer, the same voiceless-stop and vowel syllables were generated as before, except the noise burst was eliminated and the consonant cue was produced

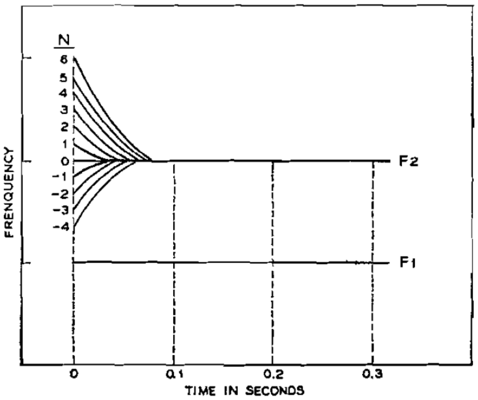


Figure 9.11: Second-formant trajectories for testing the contribution of formant transitions to the perception of voiceless stop consonants. (After (Cooper *et al.* [1952]))

solely by a transition of the second formant.

The ensemble of transitions tested is shown in Fig. 9.11. The transition numbers, N, ranging from -4 to +6, indicate the starting frequencies of the second formant. In terms of actualHz, the starting frequencies are given by $[F2 + N(120)]$ Hz, where $F2$ is the steady-state second formant frequency of the two-formant vowels shown in Fig. 9.9[4]. The first formant was maintained constant at the values given in Fig. 9.9. The fundamental frequency of the sound was also held constant at 120Hz. The durations of the transitions were 40 msec for 1, and 80 msec for +6. For transitions in between, the durations varied linearly. The form of the transition curve is unspecified except that an effort was made to approximate the transitions seen in spectrograms of real speech. In the experience of the authors, variations in the duration of the transition and its curvature do not cause the sound to change from one stop consonant to another.

The median /p,t,k/ responses of 33 listeners, for these transitions coupled with seven different vowels, are shown in Fig. 9.12. The lengths of the plotted bars show the quantile ranges of the responses. The results indicate that the second formant transition effectively cues the /p,t,k/ discrimination.

In extending this line of investigation to other consonants, the same authors found that the second formant cues also apply to the voiced cognates /b,d,g/. Distinctions between the voiced and unvoiced cognates are made by the first formant transition and by the voice bar. When vowel plus nasal-consonant syllables are generated in a similar manner, but with the formant transitions at the ends of the vowels and with an added, constant nasal resonance, the second formant transitions that serve to distinguish /p,t,k/ and /b,d,g/ also serve to distinguish /m,n,ŋ/ (Liberman *et al.* [1954])).

---

[4] An exception, apparently, was the negative F2 transitions of the vowels /o/ and /u/. This was $\left[F2 + N\left(\frac{120}{2}\right)\right]$ (Liberman *et al.* [1954]).

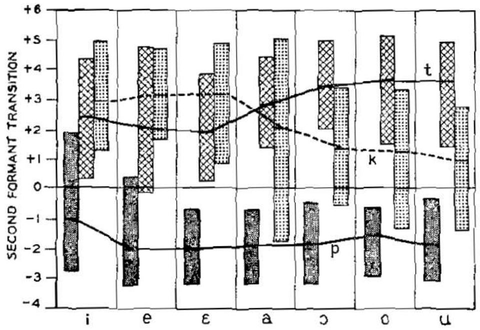


Figure 9.12: Median responses of 33 listeners to stop consonant and vowel syllables generated by the patterns shown in Fig. 9.11. The bars show the quartile ranges. (After (Cooper *et al.* [1952]))

Returning to the syllables composed of voiceless stop and vowel, several remarks can be made. The two sets of results show the individual contributions of the noise burst in the stop release and the formant transition in the following vowel. The results do not, however, suggest how these cues combine and how they may relate to each other. One might expect that identification would be improved by the combined burst and transition cues, and that they might complement each other; when one is weak, the other might be strong. In some syllables both cues may not be sufficient, and a still different factor, such as third formant transition, may be vital to the discrimination.

The dependence of consonant perception upon the following vowel suggests to the authors that listeners perceive speech in acoustic units of syllable length or perhaps half-syllable length[5]. A one-to-one correspondence between sound and phoneme is not found, and the phoneme may not exist in the speech wave in a free form. Clearly, one should not expect to find absolute acoustic invariants for the individual phoneme.

The experiments of the preceding discussion concerned sounds generated from abstracted spectrograms and by a particular synthesizer. Similar experiments have aimed to determine the perceptual adequacy of other synthesizers and to examine the influence of still different acoustic cues upon recognition. One of these has treated the synthesis of isolated fricatives and fricative-vowel syllables (Heinz and Stevens [1961]). Fricative consonants were generated by filtering noise with a single pole-zero electrical circuit. The frequency of the zero was always maintained an octave below that of the pole. The object was to determine whether such an idealized spectral representation can elicit fricative responses, and further, to establish the ranges of pole-zero locations associated with the particular responses. (Recall from Chapter 4 that the mode pattern of fricatives usually involves a number of poles and zeros. Recall, too, that the discussion in Section 9.2.7 suggests that many of the modes may not be perceptually significant.)

---

[5] This point, and other views on it, will be discussed further in Section 9.5.

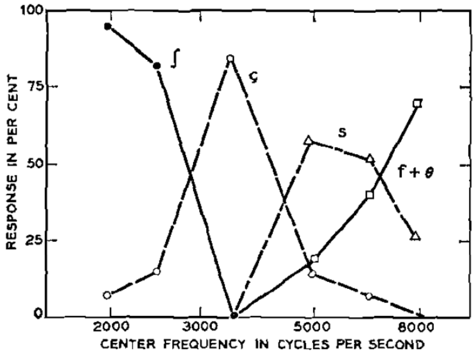


Figure 9.13: Listener responses in absolute identification of synthetic fricatives produced by a pole-zero filtering of noise. The frequency of the pole is indicated on the abscissa, and the frequency of the zero is approximately one octave lower. (After (Heinz and Stevens [1961]))

In one test, fricative consonants were generated and tested in isolation. A range of tuning and bandwidth was explored for the pole and zero. Identifications were made from an ensemble of five phonemes; namely, /ʃ,ç,s,θ,f/. The synthetic sounds were 200 msec in duration. The results show that different resonant bandwidths, ranging in Q from about 5 to 10, produce no significant changes in the fricative responses. Changes in tuning of the resonance, however, produce important differences in response. The effect is illustrated by the percentage response*vs* resonant frequency plotted in Fig. 9.13. The /f/ and /θ/ responses are combined.

Using the same synthetic fricatives, consonant-vowel syllables were synthesized with a terminal-analog synthesizer. The vowel used was always /ɑ/, and the syllable synthesized is illustrated by the schematic spectrogram in the upper part of Fig. 9.14. The timing sequence of control functions for the terminal-analog synthesizer is shown by the lower curves in Fig. 9.14. The first two curves show the build-up and decay characteristics of the noise (voiceless) and buzz (voiced) excitation. The third curve shows the timing of the formant transitions. The Fl vowel transition always started from 200Hz. The initial F2 value was either 900, 1700 or 2400Hz. Fricative resonances of 2500, 3500, 5000, 6500 and 8000Hz were used. Listeners were required to identify the initial consonant as /f,θ,s,ʃ/.

The consonant judgments–as functions of the fricative resonance frequency and second-formant transition-are plotted in Fig. 9.15. The results for two ratios of consonant-to-vowel intensity are shown, namely -5 db and -25 db. Two response contours are also shown. Inside the dashed lines the indicated fricative is responded in more than 90% of the presentations. Inside the solid lines the response is greater than 75%. The two consonant-to-vowel intensities dramatize the importance of relative level in the perception of /θ/ and /f/, and to a lesser extent, /s/. The responses also suggest that the fricative /f/ is distinguished from /θ/ largely on the basis of

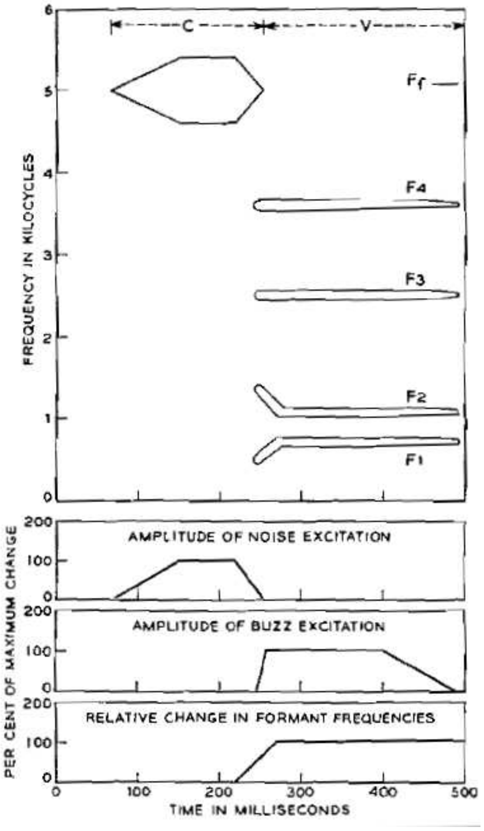


Figure 9.14: Abstracted spectrogram showing the synthesis of a syllable with fricative consonant and vowel. The single fricative resonance is F_f. The four-formant vowel is an approximation of /ɑ/. The lower three curves represent the temporal variation of the excitation and formant frequencies in the syllable. (After (Heinz and Stevens [1961]))

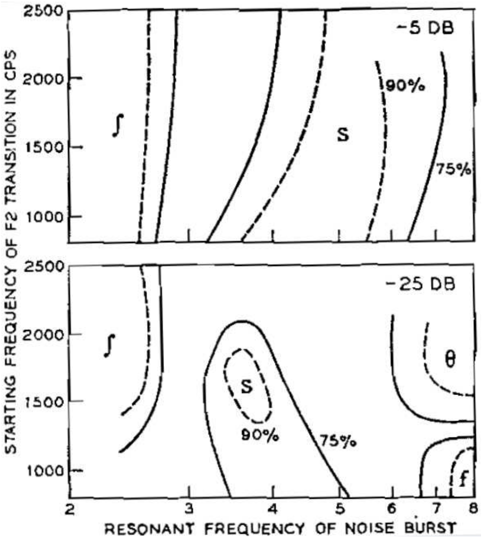


Figure 9.15: Absolute identifications of the initial consonant in the synthetic syllable schematized in Fig. 9.14. Two response contours are shown corresponding to 90 and 75% identification. Two consonantto-vowel intensities (-5 and -25 db) are shown. (After (Heinz and Stevens [1961]))

the F2 transition in the vowel. Contrariwise, the formant transition does not have much influence upon the /s/ and /ʃ/ discrimination, this being determined more by the frequency of the fricative resonance. Another study, closely related in form and philosophy to the present one, has been carried out for Japanese fricatives (Nakata [1960]).

In much the same vein, other experiments have studied formant transitions with a transmission-line analog (Stevens [1956]). The results show thatlow F2 loci (1000Hz or less) are generally associated with bilabial or labiodental articulatory configurations. On the other hand, F2 loci in the middle frequency range (1500 to 2000Hz) are associated with alveolar configurations, and F2 loci above 2000Hz are associated with palatal configurations.

A still different approach to synthesis and perception is exemplified by the generation of connected speech from individual, spectrallyconstant synthetic segments (Cohen and T'Hart [1962]). The segments are of phoneme length and are time-gated with prescribed build-up, decay and duration. From these results the suggestion is advanced that proper dimensioning of the time parameter makes it possible to neglect a number of details of formant information usually considered to be of paramount importance. It seems reasonably clear, however, that the ear accomplishes a short-time spectral analysis (see Chapter 5) and that it appreciates continuous variations both in frequency and intensity. The "time parameter" view implies a trading relation of a sort between spectral information and temporal detail. Such a trade may in fact exist, but the extent to which it can be exploited may be limited. It would appear unlikely that high-quality, high-intelligibility speech could be consistently synthesized without taking account of mode transitions within phoneme-length segments.

9.3.3 Effects of Learning and Linguistic Association in Absolute Identification of Speech-Like Signals

It was suggested earlier that at least two limitations exist in applying classical psychophysical data to speech recognition. First, the classical measures are generally restricted to differential discriminations. Second, they are usually made along only one dimension of the stimulus. Speech, however, appears to be a multidimensional stimulus. Its perceptual units, whatever they might be–and they probably vary according to the detection task–are presumably perceived absolutely. At least one experiment has attempted to measure the effects of learning and linguistic association in absolute discriminations. The tests treated several dimensions of complex, speech-like sounds (House *et al.* [1962]).

Four different groups of stimuli (A, B, C and D), varying in their similarity to speech, were used. The stimuli of each group were further divided into subgroups. The signals of each subgroup were coded in a given number of dimensions. Each member of the subgroup was designed to convey three bits of information per presentation. The signals of the A group, for example, were produced by filtering random noise with a simple resonant circuit. They could be coded along time, frequency and intensity dimensions. Stimuli in subgroup A1 were coded unidimensionally in terms of 8 frequency positions of the noise resonance. The center frequency of the resonance varied from 500 to 5000Hz, and its corresponding bandwidth varied from 300 to 3120 Hz. One intensity (namely, a reference intensity) and one duration (300 msec) were used. In contrast, stimuli of subgroup A7 were coded in terms of two frequency positions of the noise (820 or 3070Hz), two intensity values (8 db re A1), and two durations (150 or 450 msec). The subgroups A2 through A6 utilized different combinations of dimensions and quantizationsbetween these extremes.

The B stimuli were also rudimentary signals but with slightly more speech-like properties. They had temporal and spectral properties roughly analogous to vowel-consonant syllables. The vowel element was produced by exciting a single resonant circuit with 125Hz pulses. The center frequency of the resonator was 300Hz and its bandwidth was 60Hz. The consonant portion was produced by exciting a simple resonant circuit with white noise. The coded dimensions of the B signals were center frequency and bandwidth of the noise portion (center frequencies 500 to 5000Hz, bandwidths 100 to 1000Hz); intensity of noise (14 db); and duration of the silent interval (gap) between the vowel and consonant (10 to 180 msec). The total duration was always 350 msec. Like the A group, set B 1 was a one-dimensional coding and had eight frequency values, one intensity and one duration. Set B7 was a three-dimensional coding and had two frequencies, two intensities and two gap durations.

The C group was constructed to be still more similar to speech. It incorporated

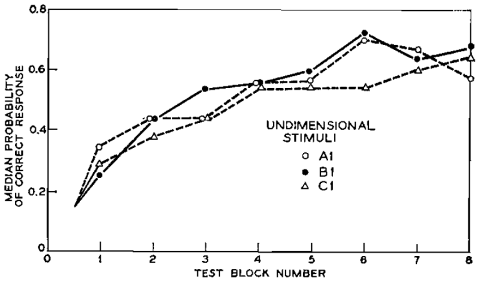


Figure 9.16: Median probability of correct response for frequency-coded, one-dimensional stimuli. (After (House *et al.* [1962]))

many of the characteristics of acceptable synthetic speech samples. Like B, the C stimuli were vowel-consonant syllables, but the vowel was generated from four resonators whose center frequencies were fixed at 500, 1500, 2500, and 3350Hz. Their bandwidths were approximately those of spoken vowels. The first formant was given a falling transition to the time gap, in analogy to the vowel-to-stop consonant transition. The consonant portion was generated by a single pole-zero filtering of noise, similar to the circuit described in the preceding section for producing fricative consonants (Heinz and Stevens [1961]). Voiced excitation during the vowel was inflected from 120 to 150 pps. The stimulus dimensions and the varied parameters were similar to those of the B signals. In set C1, the consonant resonance varied from 500 to 5000 in eight steps. The vowel duration was 250 msec, the gap 50 msec, and the consonant 100 msec. (Total duration was always 400 msec.) In set C7, the consonant dimensions of resonance, intensity and gap were all binary.

The D stimuli were real, monosyllabic speech utterances produced by one speaker. Only a single, three-dimensional subgroup was used. The eight syllables were composed of two vowels, /ɪ/ and /ʌ/, and four consonants /f,s,p,t/. Four of the eight syllables were monosyllabic English words, and four were nonsense syllables.

In the tests the stimuli were presented singly in isolation. Listeners were required to associate each with one of eight unlabelled buttons on a response panel. After the subject made his selection, one of eight lights on the panel flashed, indicating the correct button with which to associate the stimulus. The next sound was then presented. There was no speed requirement.

The results show how the median probability of correct identification increases with learning. Identification data from twelve listeners for the unidimensional, frequency-coded stimuli are shown in Fig. 9.16. Each test block involved the randomized presentation of sixteen items from a given (8-component) stimulus ensemble. The responses to the tri-dimensional stimuli are given in Fig. 9.17.

The two sets of results show that learning is more rapid for the tridimensional stimuli than for the one-dimensional items. Of the tridimensional signals, real speech (D7) is learned the fastest. The least speech-like artificial signal (A7) is learned the

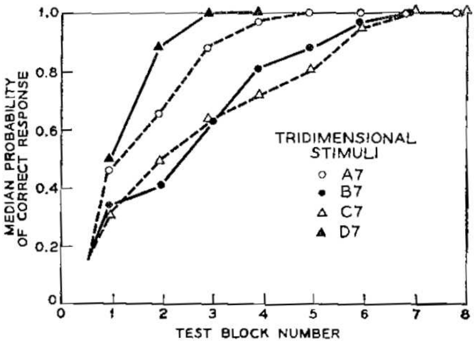


Figure 9.17: Median probability of correct response for time-frequency-intensity coded threedimensional stimuli. (After (House *et al.* [1962]))

next fastest. The results suggest two conclusions. First, performance during learning is better when the stimuli are coded in several physical dimensions than when they lie along a unidimensional continuum. Second, as the physical characteristics of the stimuli are made more similar to speech, there is a deterioration of performance, except for stimuli that are actual speech signals!

The explanation advanced for this latter and somewhat surprising result is that neither the A, B, nor C stimulus ensembles were sufficiently like speech to elicit a linguistic association. Hence, they had to be identified in a manner different from speech. Real speech sounds, however, are categorized with great facility by listeners, and presumably the subjects made use of linguistic categories in discriminating the D stimuli. The A, B, and C signals, lacking linguistic association, were probably identified in terms of what may be more "natural" basic dimensions in perception, namely, loudness, pitch and duration. Discrimination of these fundamental dimensions might be expected to be more clear cut for the A stimuli. The B and C signals apparently do not order well along these dimensions because of the fixed initial vowel segment.

The results are therefore interpreted to argue against the existence of a speech-like continuum. Although the signals may bear more or less resemblance to speech from a physical point of view, the subjective responses exhibit a sharp dichotomy. Either the sounds are associated with linguistic entities or they are not. In the present experiment presumably none of the synthetic sounds were associated with linguistic quantities. Within a linguistic frame, the tendency is to categorize a signal according to dimensions established by the language structure. Perception of the signal as a linguistic unit probably depends strongly upon nonperipheral processes. Small details of the signal, routinely preserved at the periphery of the ear, may not be of primary importance. For nonlinguistic signals, on the other hand, the tendency is to order them along what seem to be natural psychological dimensions. Their discrimination probably requires less central processing than does the perception of speech.

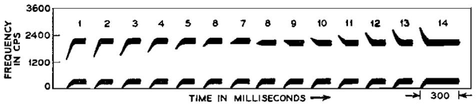


Figure 9.18: Synthetic two-formant syllables with formant transitions spanning the ranges for the voiced consonants /b,d,g/. The vowel is the same for each syllable and is representative of lei.(After (Liberman *et al.* [1957]))

9.3.4 Influence of Linguistic Association Upon Differential Discriminability

A listener's linguistic learning and experience provide an acute ability to categorize speech signals. In the experiment of the preceding section, listeners presumably resorted to linguistic associations for the D7 stimuli. They apparently did not for the other stimuli, either because the signals were not sufficiently speech-like, or because the listener's attention was not drawn to such an association by the instructions given him.

The results therefore raise a further question. Assuming that a linguistic association is made, is its effect reflected in the differential discriminations a listener can make? In other words, can the learning and discriminability acquired in linguistic experience carryover into a more classical differential comparison. At least one experiment suggests that it can (Liberman *et al.* [1957]). The objective was to demonstrate that the differential discriminability of formant motion in a synthetic speech syllable is more acute when the change traverses a phoneme boundary.

Consonant-vowel syllables were synthesized with the pattern playback device described in Section 11.3.1, Chapter 11. Two formants were used and the vowel was always /e/ (Fl=360, F2=2l60Hz). The consonants were various two-formant transitions spanning the known approximations to /b,d,g/. The set of synthetic syllables used is shown in Fig. 9.18. The positive first-formant transition is the same in all the syllables and is a necessary cue to voicing. The second formant transitions range from highly negative to highly positive. The duration is the same for all syllables, namely 300 msec.

Two tests were made. In one, the stimuli were presented singly for absolute judgment of the consonant. The allowed response categories were /b,d,g/. In the second, an ABX presentation was made. Stimuli A and B were different syllables from Fig. 9.18. They were separated by either one, two or three successive steps Shown in Fig. 9.18. Sound X was identical to either A or B. On the basis of any cues they chose to use, listeners judged whether X was most like A or B. The second test therefore gave a measure of relative discriminability at each step on the continuum described by the stimuli in Fig. 9.18.

The absolute identification results of the best subject in the experiment are shown in Fig. 9.19. This same subject's responses in the ABX test, when the step size

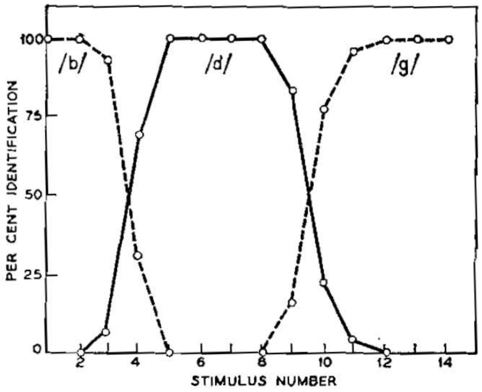


Figure 9.19: Absolute Consonant identifications of one listener for the stimuli of Fig. 9.18. (After (Liberman *et al.* [1957]))

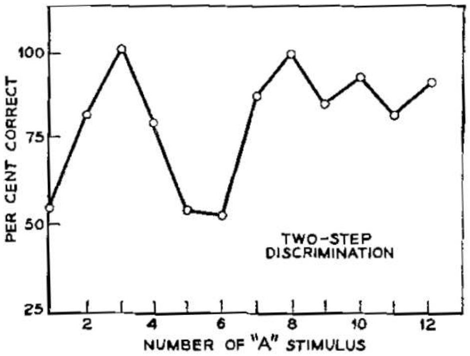


Figure 9.20: ABX responses of the listener whose absolute responses are shown in Fig. 9.19. The step size between A and B stimuli was two positions in the stimulus set of Fig. 9.18. (After (Liberman *et al.* [1957]))

between A and B is two (that is, the B stimulus number is A plus two in Fig. 9.18), are given in Fig. 9.20. Comparison of the data shows a clear diminution of differential discriminability of formant transition for the stimuli contained within the /b/ and /d/ response ranges. A corresponding drop for the /g/ range apparently is not obtained. The other subjects in the experiment did not give data with maxima and minima so well defined, but the indications are that somewhat similar variations exist. A rough approximation of differential discriminability can be made on the assumption that listeners can discriminate only so well as they can identify. This assumption tends to predict the relative variations in discriminability, but it underestimates the absolute level of discriminability. The difference may represent a so-called margin of true discrimination, that is, the ability of listeners to distinguish speech sounds not solely on the basis of phoneme labels, but also more directly by acoustic differences.

The suggestion is advanced that the inflection points in discrimination are not innately built into the human. Different languages have phoneme boundaries in different places. The case for acquired discriminability would of course be strengthened by demonstrating that native speakers of other languages exhibit maxima of differential sensitivity placed along the continuum in accordance with their languages. The crucial factor in the present experiment is the extent to which linguistic associations are elicited by the stimuli[6]. Lacking the ability to categorize, the differential discriminability might be expected to be monotonic along the stimulus continuum.

To inquire into this last point, a similar experiment was conducted on synthetic vowel sounds (Liberman *et al.* [1962]). No increase in discrimination was found at the phoneme boundaries. In addition, the differential discriminability lay considerably above that predicted simply on the basis that listeners can discriminate only so well as they can identify. (In other words, listeners can discriminate many within-phoneme differences.) The conclusion is that the perception of vowels tends to be continuous and is not as categorized as, for example, the stop consonants. A further experiment with two other phonemic distinctions, namely vowel length and tone in Thai, also failed to show sharpening at the phoneme boundary (Liberman *et al.* [1962]).

9.4 Effects of Context and Vocabulary Upon Speech Perception

The precision with which listeners identify speech elements is intimately related to the size of the vocabulary and to the sequential or contextual constraints that exist in the message. The percent correct response is higher the more predictable the message, either by virtue of higher probability of occurrence or owing to the conditional probabilities associated with the linguistic and contextual structure. This influence is apparent in intelligibility scores for various types of spoken material. Fig. 9.21

---

[6] The question is made more pointed, perhaps, by the results of the previous section where apparently no linguistic association was made with synthetic syllables.

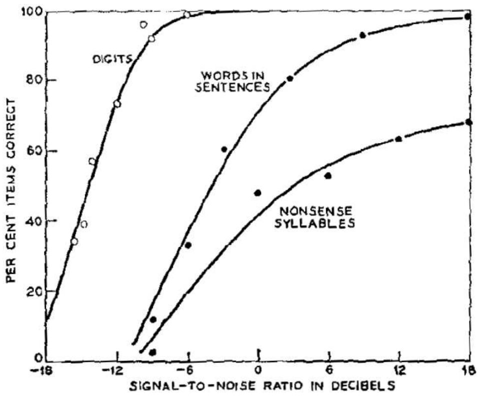


Figure 9.21: Intelligibility scores for different types of spoken material as a function of signal-to-noise ratio. (After (Miller *et al.* [1951]))

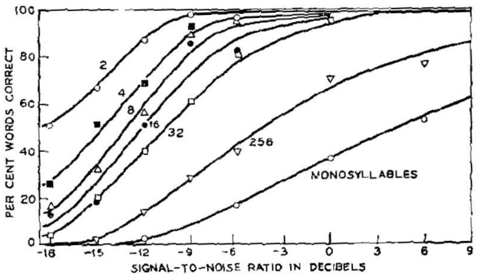


Figure 9.22: Effects of vocabulary size upon the intelligibility of monosyllabic words. (After (Miller *et al.* [1951]))

illustrates the effect in an experiment where speech was masked by varying amounts of noise (Miller *et al.* [1951]).

Three different types of test material were used. Articulation tests were made with the same subjects and experimental apparatus. One set of material was the spoken digits zero to nine. Another was complete sentences read and scored for the major words. A third was nonsense syllables which were pronounced and recorded using an abbreviated phonetic notation. As Fig. 9.21 shows, the signal-to-noise ratios necessary to produce 50 percent correct response are approximately -14 db for the digits, -4 db for the words in sentences, and +3 db for nonsense syllables. The discriminations among a small number of possibilities are obviously better than among a large number. The sequential constraints present in the sentences apparently result in higher intelligibility scores than for the nonsense material.

The effect of vocabulary size was examined in further detail. The same type of articulation tests were performed on monosyllabic word sets numbering 2, 4, 8, 16, 32, 256, or an unspecified number. For the restricted vocabularies, the listeners were informed of the alternatives. The results of the intelligibility tests are shown in Fig. 9.22. The results show clearly that as vocabulary size increases, the signal-to-noise ratio necessary to maintain a given level of performance also increases.

Semantic and syntactical constraints also influence the predictability of a speech utterance and hence its intelligibility. The grammatical rules of a given language prescribe allowable sequences of words. Semantic factors impose constraints upon those words which can be associated to form a meaningful unit. Experiments have demonstrated that the intelligibility of words is substantially higher in grammatically-correct, meaningful sentences than when the same words are presented randomly in isolation (Miller *et al.* [1951]). The sentence context reduces the number of alternative words among which a listener must decide, and the improvement in intelligibility is due, at least partially, to this reduction.

Reduction in the number of alternatives, however, is not the sole factor. Experiments have compared the intelligibility of words in grammatically-correct, meaningful sentences to the intelligibility in nongrammatical, pseudo-sentences (Miller [1962]). The pseudo-sentences were constructed so that the number of word alternatives was exactly the same as for the grammatical sentences. In the grammatical structures a listener apparently accomplishes perception in terms of phrases, or longer elements. He may delay decisions about words, rather than make them about each word as it occurs. The nongrammatical structures, on the other hand, cannot be processed this way. They must be perceived in terms of shorter temporal elements.

A somewhat different emphasis can be placed on context from the standpoint of acoustic environment and reference. Many perceptual evaluations seem to be made by a relative rather than absolute assessment of physical properties. That is, the physical surround establishes a frame of reference for the decoding operation. A simple example might be the pitch inflection of an utterance. The relative change, or pattern of inflection, is probably more significant perceptually than the absolute number of cycles per second.

Such acoustic "referencing" has been demonstrated in synthetic speech. It can be present to the extent that identification of a given monosyllabic word is strongly influenced by the time-frequency-intensity frame within which it is placed (Ladefoged and Broadbent [1957]). For example, a given synthetic vowel was produced as the central element of the synthetic word /b--t/. This word was used in synthetic sentences having different relative patterns of formant frequencies. Depending upon the acoustic reference established by the formant patterns in the rest of the sentence, the physically same synthetic word was variously identified as bit, bet or bat.

9.5 The Perceptual Units of Speech

The data in the preceding discussions suggest that speech perception is an adaptive process. It is a process in which the detection procedure probably is tailored to fit the signal and the listening task. If the listener is able to impose a linguistic organization upon the sounds, he may use information that is temporally dispersed to arrive at a decision about a given sound element. If such an association is not made, the decision tends to be made more upon the acoustic factors of the moment and in comparison to whatever standard is available.

The suggestion that a listener uses temporally spread information raises the question as to the size of the temporal "chunks" in which speech is perceived. Very probably the size of the perceptual element varies with the discrimination task, and the listener adjusts his processing rate to suit different types of speech information. He may, for example, attend to prosodic information while phonemic information is momentarily predictable. For nonspeech or nonlinguistically associated discriminations, the perceptual processing may be substantially different. In either case, however, the information must funnel through the same sensory transducer. As mentioned earlier, differential discriminations of "classical" psychoacoustic signals probably reflect the fundamental limitations of the transducer and the peripheral processing, whereas linguistically-connected discriminations probably reflect the storage and processing characteristics of the central mechanism.

Speech recognition presumably requires that sound elements be identified in absolute terms. For some sounds, however, distinctiveness is not so much an acoustic, or even articulatory factor, but a consequence of linguistic experience. A distinctiveness, which may be salient in connected speech, may be diminished or altogether lost in isolation. A case in point concerns the nasal consonants. These sounds carry a heavy functional load in connected speech (see Table 1.2, Chapter 1), but are poorly identified in isolation (see Table 9.1, Section 9.3.1).

A number of studies have aimed at determining the units in which perception occurs. For the most part the experiments arrive at disparate results, probably owing to the large differences in perceptual tasks and to the fact that there may be no single answer to the question. Perhaps exemplifying one extreme in perception is the task of speech "shadowing" (Chistovich [1962]). This approach aims to resolve whether, upon hearing the beginning of a speech sound, a listener immediately begins to make some preliminary decisions and corrects them as more information becomes available, or whether he stores long portions of data before interpreting them. The question was examined in two ways. First, the latency was measured for the articulatory movements of a listener who was repeating as rapidly as possible ("shadowing") the speech syllables he heard over earphones. The syllables were either vowel-consonantvowel or consonant-vowel. Second, the latency was measured for a written transcription of the consonant sounds in the syllables heard.

The results showed that in the vocal shadowing, the consonant latencies were on the order of 100 to 120 msec for the VCV syllables, and on the order of 150 to 200 msec for the CV's. In the VCV syllables the subject apparently anticipates the C before it is completely articulated, perhaps getting a good deal of information from the formant transitions in the initial V. He is often wrong initially, but generally corrects himself (on a running basis) by the end of the C. Because the subject reacts before he perceives the whole consonant–and even makes responses that are not possible in his language–the interpretation is advanced that the subject makes a number of simple decisions about the articulatory origin of the acoustic event (that is, whether the origin is dental, voiced, voiceless, nasal, etc.). The decisions are corrected as the sound proceeds, and a set of features are finally accumulated to form the phoneme. It is therefore suggested that shadowing is "phoneme creation" from simple decisions about articulatory parameters.

The latencies for the written mode of response were found to be very nearly the same as the latencies to the ends of the C's in shadowing (that is, the interval between ends of the original and the shadowed C's). The conclusion is therefore put forward that consonant writing is closely related to consonant shadowing.

It is difficult to say precisely how perception under these conditions relates to perception of running speech. The results may be strictly interpretable only within the frame of the task. If the task is made different, the measures are likely to indicate a different duration for the "unit." Another experiment perhaps illustrates the opposite extreme in evaluating the unit. It suggests that listeners are not only aware of large parts of an utterance at any moment, but actually may find it difficult to consider speech in terms of small segments, even when asked to make an effort to do so (Ladefoged [1958]).

The spoken word "dot" was superimposed on the recording of a complete sentence. Listeners were asked to note and report the precise moment in the sentence when the superimposed word commenced. The judgments were generally inaccurate, but it was not uncommon for subjects to report that the superimposed item occurred two or three words earlier in the sentence than was actually the case.

This behavior suggests that the mechanisms and times for processing on-going contextual information may be considerably different from those for isolated stimuli, even though the latter are speech sounds. It also suggests that continuous speech produces complex temporal patterns that are preceived as a whole. Items such as syllables, words, phrases, and sometimes even sentences, may therefore have a perceptual unity. In such an event, efforts to explain perception in terms of sequential identification of smaller segments would not be successful. As a consequence, attempts to build machines that recognize speech in terms of brief acoustic units may be of little or no profit.

It was suggested earlier (see Section 9.3.3) that "natural" auditory dimensions apparently include subjective attributes such as pitch, loudness, and temporal pattern, and that these dimensions appear useful in discriminating nonlinguistically associated sounds. These same dimensions may of course apply to continuous speech signals, but they may be assessed in different ways–perhaps in ways that are related to production. For example, there is some evidence that the loudness of speech is assessed more in terms of the respiratory effort required to produce the necessary subglottal pressure than it is in terms similar to, say, the loudness scale for sine waves (Ladefoged [1958]). If the "motor theory" of speech perception has validity, a listener may evaluate a speech signal in terms of the motor activity that produced it, as well as in terms of other acoustic factors not directly under motor control.

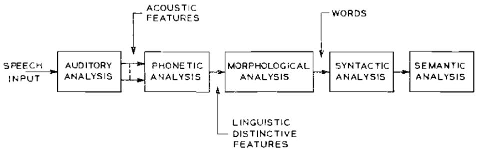


Figure 9.23: Block diagram model of stages in speech perception. (After (Bondarko *et al.* [1968]))

Many theorists in speech perception appeal to a link between production and perception. How tight this link is, is not known. If it is close, perception could conceivably occur in terms of "articulatory" segments rather than acoustic segments. In producing speech, the human has at least three kinds of feedback: auditory, tactile and proprioceptive. Blocking of one or more of these channels apparently causes some of its functions to be assumed–but generally less well-by one of the other channels. Speech attributes such as vowel quality, nasality and pitch seem highly dependent upon auditory feedback, while features such as lip and tongue movements in consonant articulation seem more dependent upon tactile and proprioceptive channels. If perception is linked to these processes, some speech properties might be identified by reference to acoustic factors, and others by reference to articulatory activity.

9.5.1 Models of Speech Perception

Much progress remains to be made in understanding and in modeling the mechanism of human speech perception. Not least is the problem of quantifying behavior in response to speech signals. Appeal to the mechanism of speech production is sometimes made on the basis that perceptual factors, at some level, must correspond to those necessary to speak the same message. This "motor theory of speech perception" has been the focus of considerable speculation and not little controversy (LIBERMAN et al.). If truly invoked by humans–which has not been shown–it has the advantage that motor commands to the vocal mechanism are more amenable to psychological study than are, say, electrical representations of speech signals in the human contex. Further, acoustic and linguistic correlates of the motor commands are more accessable for study.

At least one view (Bondarko *et al.* [1968]) has maintained that the development of a model of human speech perception is the same problem as the development of an automatic speech recognizer, and further, that present knowledge embraces only the most rudimentary aspects of such a model. The proposal for such a model involves the hierarchial structure shown in Fig. 9.23. The model is envisioned as a chain of transformations in which each stage acts as an information filter to reduce the dimensionality of the signal. For example, the first three blocks transform an acoustic signal into a succession of words where each word is described by a set of lexical and grammatical features and by prosodic characteristics. Syntax and finally semantic analysis complete the transformations necessary for message understanding. The natures of the transformations, if in fact they exist in identifiable and separable forms, are not known. Perceptual experiments do, however, suggest certain characteristics of the first two stages.

The peripheral auditory analysis made by the human cochlea is such that features of the short-time spectrum of the input signal are preserved. This analysis preserves temporal detail relevant to changes in spectral distribution, periodicity (or non-periodicity) and intensity. That this is true can be shown by psychoacoustic experiments on perception of changes in pitch, formants or intensity of speech and speech-like sounds. That this information is reduced in "dimensionality" for later processing is supported by experiments which show that consonant perception is influenced only by the direction and rate of change of formant transitions, and not by absolute values of their "loci" or initial frequencies. Similar perceptions of the direction and rate of change of fundamental frequency, or pitch, influence nasal-non-nasal discriminations in labial consonants (Chistovich [1955]).

The reduction of dimensionality performed in the phonetic analysis is likely to be one of feature analysis rather than one of comparison to a stored reference pattern. This view is supported by data on syllable recognition where features such as manner of production may be perceived correctly while, say, place of production is perceived incorrectly. Similarly, prosodic features may be perceived without discrimination of phonetic factors. Experiments on mimicking and shadowing (Chistovich *et al.* [1965]) are consistent with this in that some phonematic features can be recognized and produced even before a listener hears a whole syllable. This type of feature analysis also argues that the input to the phonemic analysis block of Fig. 9.23 may already be organized for parallel, multichannel processing.

Exactly what duration of signal may be subjected to such analysis in not clear, but data on short-term auditory memory provides some insight. In recall experiments with speech (Miller [1956], Nevelskü [1966]) a sequence of three vowels or three tones is recalled as a sequence of decisions regarding the stimuli and not as a sequence of acoustic descriptions (Chistovich *et al.* [1961]). The phonemic analysis must therefore work with speech segments shorter than average word length. Furthermore, experiments show that a man cannot remember sequences of nonsense syllables longer than 7 to 10 syllables (Miller [1956], Chistovich *et al.* [1965]). This fact bears on the size of the short-time storage and characterizes the "time window" through which the message is "seen" by the morphological analysis stage.

On the other side it is clear that a listener does not make separate decisions about every phoneme in running speech. The units with which he operates likely correspond to words, or to even longer segments. Information handed from the morphological analysis to the syntactic and semantic analysis can, consequently, be reduced in dimensionality to this extent. Auditory segments need not coincide with phonemes-i.e., each segment need not contain information about one and only one phoneme and the number of segments need not equal the number of phonemes.

Experiments on recall show that a listener remembers phonemes as a set of features (Wickelgren [1965, 1966], Galunov [1966]). Therefore, the phonemic information at the output of the phonetic analysis block should be represented by abstract, distinctive features. Several different acoustic (or auditory) features may contain information about one and the same distinctive feature.

9.6 Subjective Evaluation of Transmission Systems

9.6.1 Articulation Tests

A conventional technique for evaluating a communication facility is to determine the intelligibility of the speech it transmits. This is customarily done by counting the number of discrete speech units correctly recognized by a listener. Typically, a speaker reads a list of syllables, words, or sentences to a group of listeners. The percentage of items recorded correctly is taken as the articulation score. By choosing test material representative of the sound statistics of a language, a realistic test can be made of the transmission system. The development of the so-called phonetically-balanced (PB) test words has this objective (Egan [1944]). The techniques for administering and scoring articulation tests have been described in many places, and there is little need to repeat the procedures here (see for example, (Beranek [1954], Harris *et al.* [1957], Richardson [1953])).

An articulation score is not an absolute quantity. It is a function of parameters such as test material, personnel, training, and test procedure. It consequently should be treated as a relative measure. Usually the significant information is a difference between scores obtained with the same material, procedures and personnel. Syllable and word items can be scored in terms of the correctness of their written response. Sentences can be scored either in terms of their meaning conveyed, or in terms of key words in the sentence. Contextual constraints usually make the scores for sentences higher than those for isolated words. One relation that has been worked out between word articulation and sentence intelligibility (in terms of meaning conveyed) is shown in Fig. 9.24 (Egan [1944]).

Articulation tests are typically done without speed requirements, and the stimulus presentation rates are favorable for careful consideration of each item. More realistic articulation tests–so far as the informational capacity of a transmission system is concerned–should include time limitations. Some research into the design of such tests has been initiated (D'Eustachio [1960]). The philosophy of adding stress to the communication task is that "fragile" systems will fail before more robust systems with perhaps valuable redundancy. Time limitation is but one way stress can be introduced. Additional mental activities, such as required with simultaneous motor or visual tasks, also load the listener. The aim is to control the sensitivity of the test by varying the subjective load (Nakatani [1971], Moncur and Dirks [1967]).

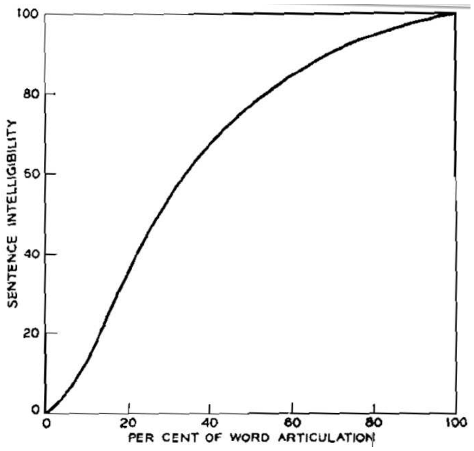


Figure 9.24: A relation between word articulation score and sentence intelligibility. Sentences are scored for meaning conveyed. (After (Egan [1944]))

9.6.2 Quality Tests

In the conventional articulation test, a listener is usually required to respond with the written equivalent of the speech he hears. The quality or naturalness of the signal is not specifically evaluated. Methods for quantitatively rating speech quality have not been well established, mainly because the physical correlates of quality are poorly understood. Various rating-scales and rank-order methods have been examined (Egan [1944]). However, generally applicable techniques for uniquely relating speech quality and acoustic factors are not presently available.

One proposal has suggested that speaker recognition is an important and measurable aspect of naturalness (Ochiai and Kato [1949], Ochiai [1958]). Results along these lines suggest that spectral distortions of a speech signal affect the accuracy of speaker identification much differently from the way they affect phoneme identification. Another proposal has been to consider voice quality as the "spectral remainder" after inverse filtering a prescribed number of formants out of the signal (Fujimura [1961]). A large contribution to what remains is then attributed to the source of vocal excitation.

Perhaps one of the most promising tools for assessing speech quality is Multidimensional Scaling (Shepard [1962], Kruskal [1964], Carroll [1972]). In this technique, non-metric data, corresponding to subjective judgments on a signal, are analyzed to reveal preferred rankings, and to show how individual subjects weight (in importance to preference) different attributes of the stimulus.

The technique assumes that observers use a common set of subjective factors (or coordinates) on which to base their judgements. The analysis indicates the number of such factors needed to account for prescribed amounts of variance in the subjective judgments. It does not, however, identify the physical correlates of the factors. This interpretation is a human one, and must rest upon knowledge of the physical properties of the stimuli.

The method is applicable to judgments made in close comparison (say, similarity or difference judgments on stimulus pairs) and to judgments made on an absolute basis (say, absolute assignment of quality ratings). Numerous variations of the method exist. An explanation of all would fill a book itself. The most expedient vehicle to illustrate the nature of the method is a specific example.

In one application, multidimensional scaling was used to assess the acceptability of amplitude-modulated, periodic pulses as an electronic telephone ringing signal (Bricker and Flanagan [1970]). Physical variables were pulse repetition frequency (f_0), harmonic content (c), modulation frequency (f_m) and modulation duty-factor (df)[7]. Listeners heard single presentations of each signal condition and assigned an absolute numerical rating chosen from an unbounded range of positive and negative integers. Positive ratings were assigned to signals that were liked and negative to those disliked. The assigned ratings of each subject were converted to standard scores having zero mean and unity standard deviation.

The normalized judgments of n subjects on m different signal conditions produce an $n \times m$ data matrix S. The multidimensional procedure factors this data matrix into an $n \times r$ matrix of subject vectors and an $r \times m$ matrix of stimulus coordinates in r-dimensional space. The product of the subject and stimulus matrices is an $n \times m$ matrix S^* which is, in a least-squares sense, the best approximation of rank r to the original data matrix S. In particular, the r-dimensional projections of the stimuli onto each subject's vector constitute the best approximation to that subject's original data vector. The r-dimensional projections of a subject's vector onto the r orthogonal coordinates indicate the relative weights assigned to the coordinates by that subject.

The goal is to find directions in r-space along which signals are ordered in a physically interpretable manner. These directions are then related to the common perceptual attributes assumed as the basis for judgment. The relation of the subject vectors to these directions indicate the weight (or importance) of the attributes in the individual subjective ratings.

The r-dimensions are ordered according to the size of their characteristic roots, or to the proportion of the variance they account for in the original data. In the present example 40 subjects rated 81 signal conditions, and three dimensions accounted for most of the variance ($r = 3$). The projections of the subject vectors onto the two most important dimensions are shown in Fig. 9.25a.

Each arrowhead is the endpoint of a unit vector in the 3-dimensional unit sphere generated by the program. The vector thus specified may be imagined as a line segment from the end point extending through the origin and an equal distance beyond; the arrow points in the direction of higher rating by that subject. The relative weights given to each of the three dimensions by a given subject, according to the assumptions of the technique, are reflected graphically by the perpendicular projections on the three axes of that subject's endpoint. Specifically, the squares of

---

[7] The modulation waveform was a half-wave rectified version of $(a + \sin 2\pi f_m t)$.

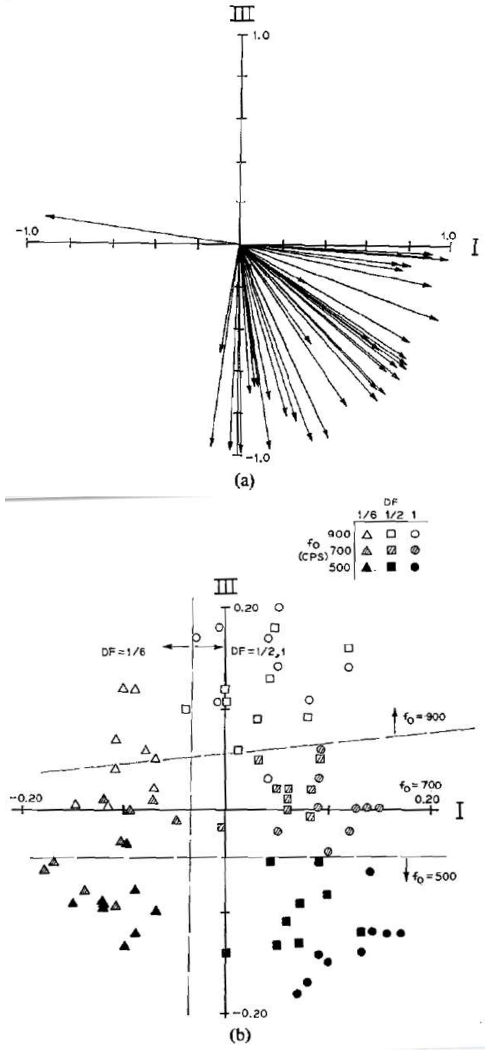


Figure 9.25: (a) Subject vectors obtained from a multi-dimensional scaling analysis projected onto the two most important perceptual dimensions I and III. The data are for a tone ringer experiment. (b) Preference judgments on 81 tone-ringer conditions, projected onto the two most important perceptual dimensions I and III. Direction of high preference is indicated by the vectors in Fig. 9.25a. (After (Bricker and Flanagan [1970]))

the projected values sum to 1.0 (by definition of the unit vector) and the subject weights are quantitatively related as the squares of the projected values. Thus, a subject whose endpoint is close to the end of one axis is described by the model as weighting that dimension heavily and the other two negligibly. One subject in Fig. 9.25a is seen to assign weights particularly different from the other 39.

The 81 stimulus coordinates of the preference judgments on the 81 signal conditions are shown projected onto the same factor plane in Fig. 9.25b. Each point represents a single signal condition. On this plane, a distinction is made between those signals differing only in duty factor (*df*) and fundamental frequency (f_0) (see insert key)[8]. The axes are scaled so that the variances of stimulus values on the two coordinates are equal. Dimension I can be associated with the physical attribute duty factor. Dimension III can be interpreted as fundamental frequency. The signal conditions can be divided according to duty factor and fundamental frequency, as shown by the dashed lines. Considering the direction of subject vectors in Fig. 9.25a, one sees there is a general preference for low duty factor and low fundamental frequency signals.

Multidimensional scaling in its many forms appears particularly promising for quality assessment of speech signals. Synthetic speech is a good case in point. Here the intelligibility may be made high, and the interest is in finding physical factors that relate to (and may be used to improve) naturalness. In other instances, multidimensional scaling has been valuable in assessing quality degradations due to nonlinear distortions in speech transmission systems.

9.7 Calculating Intelligibility Scores from System Response and Noise Level: The Articulation Index

Articulation tests, properly done to get stable and consistent results, are immensely time consuming. More desirable is the ability to estimate intelligibility from the physical transmission characteristics of the system; for example, from the frequency-amplitude response and the noise level. Under certain restrictive conditions, the well-known articulation index is a technique for making such an estimate (French and Steinberg [1947]). The concept has been extended and organized into graphical and tabular operations for rapid, practical application (Beranek [1947, 1954], Kryter [1962]).

The articulation index method is limited to particular distortions in systems using conventional "waveform" transmission. These distortions include relatively smooth variations and limitations in the transmission bandwidth, and the masking of the

---

[8] Each triangle, for example, represents nine different combinations of modulation rate and harmonic content.

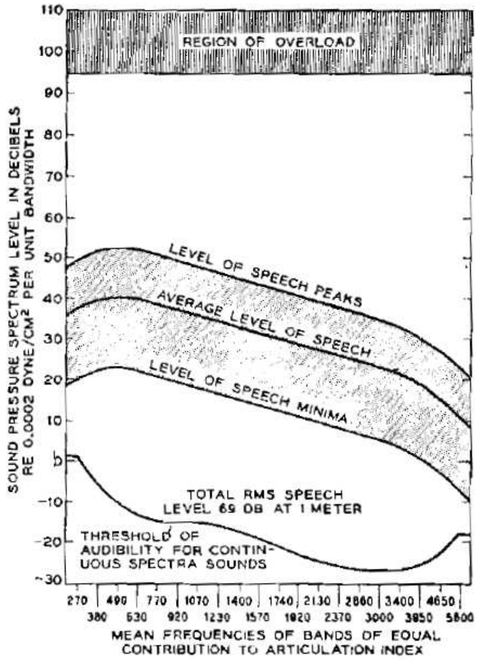


Figure 9.26: Diagram for calculating the articulation index. (After (Beranek [1954]))

transmitted signal by ongoing, continuous-spectra noises. Under certain conditions, interference caused by temporally interrupted noise, nonlinear amplitude distortion (peak clipping), and masking by reverberation can be accounted for. In general, however, the technique is not applicable to systems whose transmission bands exhibit many sharp peaks and valleys, to periodic line spectra masking noises, to intermodulation distortions and nonlinearities, and to transmission systems generally of the analysis-synthesis type (that is, where the speech information is coded in terms other than the facsimile waveform).

The technique for calculating the articulation index (AI) has been described in detail in many other places. The intent here is simply to recall its principles and, in a brief way, to indicate its applicability and utility. Its calculation is illustrated by the graph in Fig. 9.26 (Beranek [1954]). This plot shows several spectral densities laid off on a special frequency scale. The frequency scale is similar to the mel (pitch) scale. It is experimentally partitioned into twenty bands that contribute equally to intelligibility. The various spectral densities, or rms sound pressure levels per cycle, show: (a) the threshold of audibility for continuous spectra sounds, (b) the peak, average and minimum levels of speech for a man's raised voice at a distance of one meter (see Section 5.1.7, Chapter 5), and (c) an approximate overload spectrum level for the human ear.

In its simplest form, calculation of the articulation index proceeds as follows. The level and shape of the plotted speech spectrum is modified according to the amplifi-

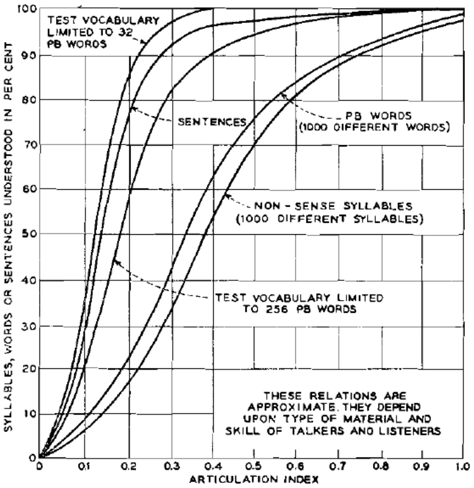


Figure 9.27: Several experimental relations between articulation index and speech intelligibility (After (Kryter [1962]))

cation and bandpass characteristics of the transmission system. The spectrum level of any added masking noise is plotted onto the graph. So long as the system response and noise level are such that all of the shaded "speech region" (between minima and maxima) lies above threshold, above the masking noise, and below overload, the intelligibility will be near perfect. In such a case the articulation index is 100%. If any of the speech region is obscured by noise, threshold or overload, the articulation index is diminished by the percentage of the area covered.

Having obtained a number for AI, it is necessary to relate it to intelligibility. The relation is an empirical one and is established from articulation tests. As mentioned earlier, articulation tests are subject to considerable variability and their results depend strongly upon testing technique and procedure. Absolute values of scores so derived must be used and interpreted with great discretion. Usually it is more relevant to consider differences in intelligibility scores, arrived at by the same technique, than to consider absolute values. Representative empirical relations between intelligibility score and articulation index for a range of test conditions are shown in Fig. 9.27 (Kryter [1962])).

9.8 Supplementary Sensory Channels for Speech Perception

Supplementary methods for speech communication are of great importance to persons either totally deafened or with partial auditory impairment. Not only is it difficult for them to hear the speech of others, but they cannot hear their own speech. It consequently is common that they also experience difficulty in speaking.

At least three avenues have been considered at the research level for providing supplementary perceptual channels and machine aids for speech communication. They include visual, tactile, and auditory approaches. The latter is oriented toward making use of whatever hearing ability may remain. Each approach can be illustrated briefly by a specific example. Other interests and efforts exist in the area.

9.8.1 Visible Speech Translator

One well-known technique for visually displaying speech information is the "Visible Speech" method (Potter *et al.* [1947]). A real time sound spectrograph, called a Visible Speech Translator, produces a running, continuous spectrographic display on a phosphor screen (Riesz and Schott [1946], Dudley and Jr. [1946]). The format is similar to the conventional sound spectrogram (shown in Section 5.1.4, Chapter 5) except that the pattern is "painted" continuously, either on a rotating cathode ray tube or on a phosphor belt. As the trace advances with time, a given duration of the past speech is retained and displayed by the persistence of the trace.

Some experiments have been made into the ability of viewers to "read" the direct-translator displays (Potter *et al.* [1947]). The results showed that after relatively lengthy training, trainees were able to converse among themselves by talking clearly and at a fairly slow rate. Within the limits of their vocabulary, they learned to carryon conversations with about the same facility as a similarly advanced class in a foreign language. The learning rates observed in the tests correspond roughly to 350 vocabulary words per one hundred hours of training.

Real-time spectrographic displays appear to have more promise for speech teaching, that is, articulatory training, than for speech reading. Some research has applied spectrographic methods in teaching articulation to deaf children (Stark *et al.* [1968], Risberg [1959], Pickett [1969]).

Because of the complex apparatus and important training procedures, visible speech techniques still remain in the realm of research. These and related methods–for example, the display of articulatory data and of formant data–are all valid problems for research and may hold potential for supplementary communication. Particularly promising are simple devices which signal rudimentary speech features, such as voicing, fabriction and stop gap (Upton [1968]). At present, however, much remains to be learned about modes of visual presentation of speech information.

9.8.2 Tactile Vocoder

The sense of touch offers another possibility for real-time communication. A filter bank analyzer, similar to that used in a vocoder, is one means for supplying cutaneous information about the short-time amplitude spectrum of speech (Pickett [1969]). The technique is shown in Fig. 9.28. Ten contiguous bandpass filters, spanning the frequency range 100 to 8000Hz, receive the speech signal. Their outputs are rectified

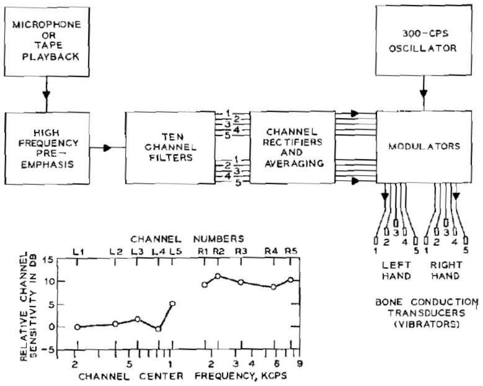


Figure 9.28: Block diagram of a tactile vocoder. (After (Pickett [1969]))

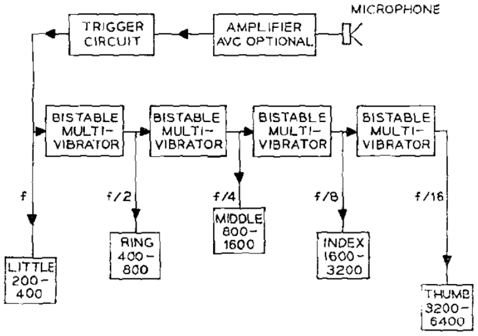


Figure 9.29: A frequency-dividing tactile vocoder. (After (Kringlebotn [1968]))

and smoothed to obtain values of the short-time spectrum at ten frequency positions. The ten time-varying voltages are used to amplitude-modulate individual sinusoidal carriers of 300Hz[9]. The modulated carriers are then applied to fingertip vibrators (actually bone conduction transducers). The analyzing channel of lowest frequency is led to the small finger of the left hand, and the channel of highest frequency connects to the small finger of the right hand.

After practice with the presentation, some subjects are able to make sound discriminations comparable to, and sometimes better than, that achieved in lip reading. When the tactile information is used in combination with lip reading, the ability to identify spoken words is considerably increased. For example, in one measurement of discrimination among 12 words, the lip reading response was about 60% correct. When supplemented by the tactile information, the response increased to 85% (Pickett [1969]).

As in the visible speech method, the vocoder apparatus for tactile display is relatively complex. A much simplified tactile device is shown in Fig. 9.29 (Kringlebotn

---

[9] This tactile "carrier" is used because the frequency range of the skin's vibratory sensitivity is limited to about 100 to 800Hz.

[1968]). This device employs only five vibrators applied to one hand. No filters are used, but stages of frequency division are arranged to divide five frequency ranges of the speech signal so as to vibrate the fingers individually. The vibrations on each finger are felt most strongly in the frequency range 200 to 400Hz. Because of the successive frequency divisions, this sensitivity range corresponds to successively higher frequency ranges in the input signal when distributed over the fingers, going from little finger to thumb. This method probably transmits some frequency information about the speech signal in terms of tactile frequency and other frequency information in terms of tactile location. Training tests with this system have been carried out with deaf children (Kringlebotn [1968]).

A number of other efforts in kinesthetic and tactile communication are in progress. Although many of these aim toward machine aids for the blind rather than for the deaf, the presentation of sensory information involves problems common to both areas (Bliss [1962], Linvill [1969]).

9.8.3 Low Frequency Vocoder

The conventional electronic hearing aid is an amplifying and frequency shaping device. It facilitates the use of whatever residual hearing a deafened person may have. In severe cases, however, the residual hearing is often confined to a very small bandwidth, usually at the low-frequency end of the audible spectrum. For example, a typical audiogram might show 60 to 80 db loss from 30 to 400Hz and 110 db ahove 500Hz.

One proposal is to make maximal use of such residual hearing. Slowly varying signals that describe the short-time speech spectrum (such as vocoder channel signals) are modulated either onto sinusoidal carriers of very low frequency, or onto distinctive complex signals of relatively small bandwith (Pimonow [1962]). In one implementation, seven spectrum channels extending to 7000Hz are used. The rectified, smoothed outputs amplitude modulate the same number of low-frequency, sinusoidal carriers. The carriers are spaced from 30 to 300Hz. The modulated carriers are summed and presented as an auditory signal. In an alternative arrangement, the modulated signals are non-sinusoidal and include a low-frequency noise band, a periodic pulse train, and a band of actual speech. In one series of experiments, deafened subjects who could not use ordinary hearing aids apparently learned to discriminate well among a limited ensemble of words (Pimonow [1962]).

Various devices for spectrum shifting, transposing or dividing have also been considered (Guttman and Nelson [1968], Levitt and Nelson [1970]). These devices generally aim to recode high-frequency information into a lower-frequency range where residual hearing exists. Like visible speech displays, their value appears to lie more in articulatory training than in speech reception. Like the other sensory aids discussed in this section, frequency scaling devices are still in the research stage. Extended experimentation and technical development will determine their potential as practicable aids to hearing.

Chapter 10

Automatic Speech Recognition

A human can listen to meaningful speech of a given language and set down a written equivalent of what he hears. He performs a transformation on the acoustic input signal wherein distinctive linguistic elements (phonemes) are recognized and re-encoded into a sequence of letter symbols. Recognition of the linguistic elements is based upon a knowledge of the contextual, grammatical and semantic constraints of the given language. It does not take much examination of sound spectrograms to convince oneself that a unique relation generally does not exist between a given segment of the acoustic signal and a linguistic clement. Neither are phonemic boundaries necessarily apparent in the acoustic signal.

Automatic recognition of speech implies phonemic analysis by machine. It is possible to simulate crudely the initial operations performed on the acoustic signal by the human (see the frequency analysis and neural encoding performed at the ear's periphery in Chapter 7) but, to date, not even the most elaborate mechanical recognizers have been able to apply linguistic constraints comparable in effectiveness to the human. This latter area represents an active field of research in theory of grammar, semantics, and mechanical translation.

The difference (or, more precisely, the gulf) between phoneme recognition for a given language and a straight-forward encoding of the acoustic signal, say in terms of vocal modes and excitation, cannot be overemphasized. The former implies complete linguistic knowledge, the latter only that the signal is produced by the human vocal mechanism. The latter is within the scope of present speech analysis techniques. The former, as yet, is not. If phoneme recognition ultimately proves possible, the import to efficient transmission is, of course, immense. (Recall it was suggested in Section 1.2, Chapter 1, that the information rate associated with the utterance of independent, equiprobable phonemes is on the order of 50 bits/sec. A coding exists for transmitting information at this rate over a channel of about 5Hz bandwidth and 30 db signal-to-noise ratio, with as small an error as desired.)

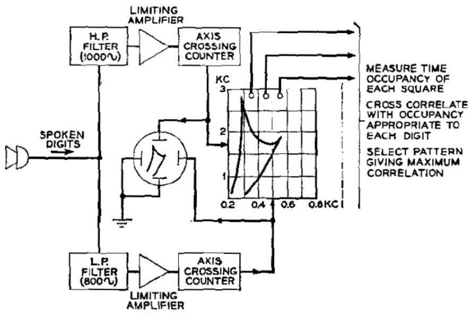


Figure 10.1: Principle of operation of a spoken digit recognizer. (After (Davis *et al.* [1952]))

10.1 Historical Approaches

A number of research investigations have treated machines which are capable of recognizing limited ensembles of speech sounds uttered by limited numbers of speakers (often only one). Generally these devices make decisions about either the short-time spectrum of the acoustic signal or about features of the time waveform. The constraints usually employed are ones more appropriate to the vocal mechanism (i.e., acoustical constraints) than to linguistic structure. Without attempting to be exhaustive, the state of the art can be outlined by several examples.

One effort toward a recognizer for a limited ensemble of sounds is a recognizer for spoken digits, called Audrey (Davis *et al.* [1952]). The principle of operation is to make a rough measure of the first and second formant frequencies as functions of time, and to compare the measured temporal patterns (in the F1-F2 plane) with a set of stored reference patterns. The stored pattern affording the best correlation is then chosen as the uttered digit.

The procedure is illustrated in Fig. 10.1. The speech signal is filtered into two bands, 900Hz low pass and 1000Hz high pass. Limiting amplifiers in both channels peak clip the signals. Axis-crossing measures approximate the frequencies of the first and second formants as functions of time. The first-formant frequency range (from 200 to 800Hz) is quantized into six lOOHz segments. The second-formant range (from 500 to 2500Hz) is quantized into five 500Hz steps. An F1-F2 plane with 30 matrix elements is thereby produced. For a given digit utterance, the time that the F1-F2 trajectory occupies each elemental square is determined.

A reference "time-occupancy" pattern for each digit is stored in the machine. The storage mechanism is 10 weighting resistors associated with each square. Through these resistors, charges are accumulated on 10 separate condensers during the time the square is occupied. A cross correlation of the stored and incoming patterns is effected by weighting the 10 conductances associated with each square according to the average time-occupancy of that square by the respective digits. That is, for each of the 30 squares, there are 10 relays which close charging paths to the 10 fixed

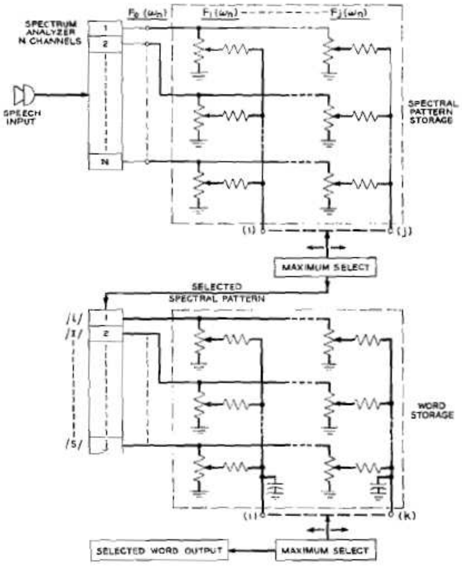


Figure 10.2: Scheme for automatic recognition of spectral patterns and spoken digits. (After (Dudley and Balashek [1958]))

condensers. The conductance of a given path is weighted proportional to the time occupancy of that square by a given digit. The condenser left with the greatest charge at the end of the utterance indicates the pattern affording the highest correlation, and hence the spoken digit.

The machine does not have provisions for automatically adjusting its stored patterns to a given speaker's voice. This must be done manually. When it is done, however, the accuracy in recognizing telephone quality uucrances of the digits ranges between 97 and 99% correct.

An extension of this technique is to correlate–on an instant-byinstant basis–a measured short-time amplitude spectrum with stored spectral patterns (Dudley and Balashek [1958]). Instead of the F1-F2 trackers, a set of bandpass filters (10 in this case, each 300 Hz wide) is used to produce a short-time spectrum. Stored spectral patterns (again, 10) are continuously cross-correlated with the short-time spectrum produced by the filters. The maximum correlation is taken as an indication of the particular speech sound being produced. The pattern-matching procedure is illustrated in Fig. 10.2. If $F_0(\omega_n)$ is the short-time amplitude spectrum produced by the n filter channels for a given speech input, and $F_j(\omega)$ the j-th stored pattern, the circuit, in principle, approximates the correlation quantity

$$\phi_{0j}(0) = \frac{1}{\Omega}\int_0^{\Omega} F_0(\omega)F_j(\omega)d\omega \quad j = 1, 2, 3, \ldots$$

by

$$\phi_{0j}(0) \approx \frac{1}{n}\sum_{n} F_0(\omega_n)F_j(\omega_n) \quad j = 1, 2, 3, \ldots$$

and selects the j that produces a maximum $\phi_{0j}(0)$. The 10 sound patterns stored in this particular development are all continuants and are /i,ɪ,ɛ,ɑ,o,u,n,r,f,s/.

A word recognizing device follows the spectral pattern recognizer to recognize the 10 digits. Similar to the Audrey device, each selected spectral pattern is weighted according to its duration in a given digit (see the lower part of Fig. 10.2). Again a maximum selection is made to recognize the uttered digit. The word indication is developed as follows. When a particular spectral pattern is energized, 10 charge paths are set up to 10 fixed condensers. The conductance of a given path is proportional to the average time for which that spectral pattern appears in a given digit. The 10 condensers therefore accumulate charges proportional to the correlation between the 10 stored word patterns and the measured pattern. At the end of the utterance, a maximum selection indicates the best-fitting word. This device–designed as an elaboration upon the previous one–provides digit recognition with good accuracy when set for a particular voice. In both devices the sequence of spectral patterns and the recognized digits are displayed on electrical panel lights. Despite its early date of conception and implementation, this device and the previously-described digit recognizer, Audrey, still reflect present limitations in automatic speech recognition; namely, one can achieve success if the vocabulary is isolated words, sufficiently small in number, and if the number of speakers is sufficiently constrained.

Another speech recognizing device also compares spectral patterns with stored patterns representative of specific speech phonemes (FRY and DENES). The comparison however, is made in a different way, and the machine types out the identification in terms of special symbols. Selection of a match is asynchronous and is initiated by the rate of change of the spectral patterns. More important, however, an attempt is made to exploit elementary linguistic constraints. A block diagram of the device is shown in Fig. 10.3.

A filter-bank analyzer (20 channels) produces a short-time amplitude spectrum. Spectral patterns appropriate to a given sound are produced by multiplying the outputs of two channels. The products are scanned by a selector, and the maximum is chosen. The choice is typed out by the machine and is remembered by a storage circuit. On the basis of the choice, the ensemble of stored patterns is biased according to digram statistics for the language. Selection of the next phoneme is biased in favor of its being the most probable one to follow the previous choice.

In the present machine 14 phonemes are recognized; four vowels, nine consonants and silence. A new selection is made whenever the product voltages have a rate of change greater than a given threshold value. With the machine adjusted for a given speaker, the spoken input and printed output have been compared. When the digram constraints are not used, the percentage correct response on individual sounds and on words is 60% and 24%, respectively. When the digram constraints are connected,

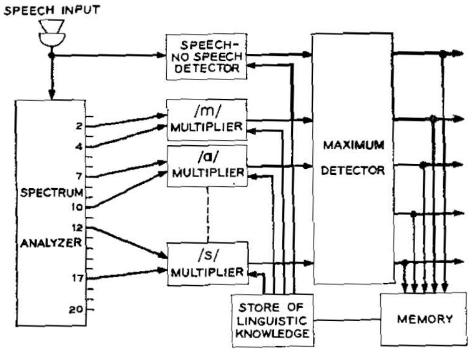


Figure 10.3: Block diagram of speech sound recognizer employing elementary linguistic constraints. (After (Fry and Denes [1958]))

these same scores rise to 72% and 44% for the single speaker. For a second and third speaker, without readjusting the machine, the sound articulation scores fall to about 45%.

The linguistic information clearly improves the recognition when scored to give all phonemes equal weight. If scored on the basis of information per phoneme, however, the digram constraints could, under certain conditions, be detrimental. The most probable phoneme is favored, but it is also the conveyor of the least information. The constraints also raise the question of sequential errors and how they might be propagated. A certain level of accuracy in the acoustic recognition is certainly necessary if the use of linguistic constraints is to lead to a decrease, rather than to an increase, in error rate. Sequential errors of course occur in the human listener. A listener, once embarked upon the wrong set of constraints in a particular sequence, may add one error to another for quite a long stretch. In the machine, severe restriction of vocabulary reduces this possibility.

If the linguistic constraints to be incorporated into the recognition process are at all realistic, the storage and processing functions become complex. Also if elaborate processings are to be carried out on the acoustic signal, large storage and rapid computation are requisite. The digital computer is adept at this, and a number of efforts have been made to capitalize upon its ability. One effort in this direction is the programming of a digit recognizer (Denes and Mathews [1960])). Short-time amplitude spectra are produced from a filter bank. The filter outputs are scanned sequentially, and the spectral data are read into and stored in the machine. A speech spectrogram–quantized in time, frequency and intensity–is laid down in the storage. Amplitude values are normalized so that the sum of the squares over all time-frequency blocks is unity. The measured time-frequency-intensity pattern is then crosscorrelated with stored spectrographic patterns. The correlation is effected by multiplying the amplitude values of corresponding time-frequency elements and summing the products over all elements of the time-frequency plane. The stored pattern yielding the maximum correlation is chosen.

Provisions are made to time-normalize the data if desired. The beginning and the end of the digit utterance are located, and the data are, in effect, stretched to fit a standard time duration (actually 60 scans of the filter bank at 70Hz) . Without time normalization only the beginning of each utterance is located, and the first 60 scans are used.

The reference pattern for each digit is obtained by averaging the spectral data for three utterances of that digit by five men. These patterns are used to recognize different utterances by the same and by different speakers. For different utterances by the same five speakers, the error rates are found to be 6% with time normalization and 13% without. When the reference patterns are set for a single speaker, the digits uttered by that speaker are recognized essentially with no error.

A more linguistically-based approach, using a large on-line computer facility, performs a feature analysis of segments of the speech waveform (Reddy [1967]). The wave is first divided into minimal segments, l0-msec in duration. Minimal segments which are acoustically similar are grouped to form larger segments representing either sustained parts or transitional parts. Features such as voiced-unvoiced, pitch, intensity, formant frequency and amplitude are used to classify each segment into four phoneme groups: stop, fricative, nasal-liquid and vowel. A very detailed algorithm is then used to assign a phoneme label to each segment of a phoneme group. The object, literally, is a speech to phoneme-like translation. This system, while recognizing the potential advantages of phonetic feature classification and language element probabilities, is nevertheless faced with the same problems of linguistic and semantic constraints that confront all recognizers. Its sophistication pays off, however, in enlarging the speaker population and vocabularly which can be successfully handled. The system has been demonstrated to yield 98% correct recognition on 500 isolated words spoken by one individual (Reddy [1969]).

At least one similar word-recognition experiment has been carried out for the Russian language (Velichko and Zagoruyko [1970]). In this case the energy-time-frequency dimensions of individually spoken words are quantized. A distance functional between the unknown word and the stored references for a word library of 203 words is computed. For two speakers, producing approximately 5000 utterances chosen from the 203 word library, the recognition accuracy was found to be about 95%. Computation time for each utterance was 2 to 4 sec.

The preceding discussion has attempted to indicate by example several stages of development in automatic speech recognition. A sizeable number of related efforts have not been mentioned (for example, (Smith [1951], Baumann *et al.* [1954], Olson and Belar [1961], Forgie and Forgie [1962], Frick [1962], Dreyfus-Graf [1962], Martin *et al.* [1964], Lindgren [1965a,b,c]). Most share a common point of departure, namely, the short-time spectrum. It is clear from the discussion that none of the schemes tells us very much about how the human processes speech information, nor about how he recognizes linguistic elements. None of the methods works well on an unrestricted number of voices, nor on a large contextual vocabulary. The human, however, is proficient at handling both. Nevertheless, the investigations do indicate what can be realized in the way of voiceactuated devices for special applications–specifically, applications where vocabulary and number of voices may be suitably restricted. It is clear, too, that for a given accuracy of recognition, a trade can be made between the necessary linguistic constraints, the complexity of the vocabulary, and the number of speakers.

Automatic speech recognition–as the human accomplishes it–will probably be possible only through the proper analysis and application of grammatical, contextual, and semantic constraints. These constraints, as yet, are largely unknown. Perhaps not surprisingly, research in speech synthesis seems to be providing more insight into linguistic constraints than is speech recognition work. One view of speech recognition (Pierce [1969]) makes the point that success will be very limited until the recognizing device understands what is being said with something of the facility of a native speaker.

10.2 Classification of Short-Time Spectra

10.2.1 Optimality Criteria for Classification and Training

A "statistical speech recognizer" is a recognizer which picks the phoneme or word that has the highest "probability" of matching the unknown utterance. In other words, if o_t is the observed spectrum at some time t, and if the possible phoneme hypotheses are λ_1 =/i/ and λ_2 =/a/, then a statistical speech recognizer chooses a "best hypothesis" $\hat{\lambda}$ according to the following rule:

$$\hat{\lambda} = \arg\max_{\lambda_i} p(\lambda_i|o_t) \tag{10.1}$$

There is no easy way to estimate $p(\lambda_i|o_t)$ directly, but using the definition of conditional probability, we can express $p(\lambda_i|o_t)$ in terms of things that *can* be estimated:

$$p(\lambda_i|o_t) = \frac{p(o_t|\lambda_i)p(\lambda_i)}{p(o_t)} \tag{10.2}$$

Since the denominator $p(o_t)$ doesn't depend on λ, it drops out of the classification rule:

$$\hat{\lambda} = \arg\max_{\lambda_i} p(\lambda_i|o_t) = \arg\max_{\lambda_i} p(o_t|\lambda_i)p(\lambda_i) \tag{10.3}$$

The probability $p(\lambda_i|o_t)$ is called the *a posteriori* probability of λ_i, and equation 10.1 is called the *Maximum a Posteriori* (or MAP) rule of classification. Notice that the MAP rule requires us to know two probabilities: the *a priori* probability $p(\lambda_i)$, and the conditional probability $p(o_t|\lambda_i)$.

$p(\lambda_i)$ is a measure of how probable I think it is that the next thing you say will be an λ_i, before I actually hear you say it. In speech recognition, $p(\lambda_i)$ is called the "language model," because it represents our knowledge of the sequences of words or phonemes which are likely in a particular language.

$p(o_t|\lambda_i)$ is the probability that a particular word or sequence of words (λ_i) will be represented by a particular sequence of spectra (o_t). Most of the technical machinery of speech recognition is aimed at estimating this probability in a computationally efficient manner.

Suppose we believe that all speech sounds have equal *a priori* probabilities (an absurd hypothesis, but sometimes this can be a useful simplification). In this case, equation 10.3 simplifies to the equation

$$\hat{\lambda} = \arg\max_{\lambda_i} p(o_t|\lambda_i) \tag{10.4}$$

Equation 10.4 is the rule which says that we should choose whichever class λ_i makes the observed data most "likely," so it is sometimes called "maximum likelihood" classification rule (ML).

10.2.2 Gaussian Models of the Speech Spectrum

In order to use either ML or MAP classification rules, we need to create a model of the probability $p(o|\lambda)$ for each of the different possible classes λ. In statistical classification, the way we do this is by gathering 10-1000 spectral vectors, $o_n = [o_n(1), \ldots]$, which we know for sure to be examples of λ, computing statistics, and using the statistics as parameters in some probability model.

For example, $p(o|\lambda)$ can be modeled using a Gaussian distribution. Given N training tokens in class λ, we can create a Gaussian model by just finding the sample mean μ_i, and the sample covariance matrix U_i:

$$\mu_i = \frac{1}{N}\sum_{n=1}^{N} o_n \tag{10.5}$$

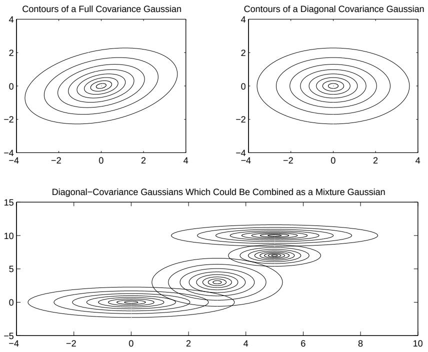
$$U_i = \frac{1}{N-1}\sum_{n=1}^{N}(o_n - \mu_i)'(o_n - \mu_i) \tag{10.6}$$

Then, if we want to know the probability that some new spectral vector o belongs to class λ, we calculate $p(o|\lambda)$ using the standard Gaussian formula:

$$p(o|\lambda) = \mathcal{N}(o; \mu_i, U_i) \tag{10.7}$$

where $\mathcal{N}(o; \mu, U)$ is notation for a Gaussian distribution with mean μ and covariance U:

$$\mathcal{N}(o; \mu, U) = \frac{1}{\sqrt{(2\pi)^p|U|}} \exp\left(-\frac{1}{2}(o-\mu)U^{-1}(o-\mu)'\right) \tag{10.8}$$


Contours of a Full Covariance Gaussian

Contours of a Diagonal Covariance Gaussian

Diagonal−Covariance Gaussians Which Could Be Combined as a Mixture Gaussian

Figure 10.4: Contour plots of Gaussian and mixture-Gaussian probability densities.

If o has only two dimensions ($c(1)$ and $c(2)$), it is possible to visualize the probability distribution $p(o|\lambda)$ by defining several "altitudes" θ_k, and drawing the contour lines

$$p(o|\lambda) = \theta_k \tag{10.9}$$

When $p(o|\lambda)$ is a Gaussian probability distribution, the resulting contour plot is always an ellipse in two dimensions, as shown in the upper left plot of figure 10.4.

A special case of particular interest is shown in the upper right plot of figure 2. If $c(1)$ and $c(2)$ are independent of each other — that is, if $u_i(1,2) = 0$ – then the major and minor axes of the ellipse are always parallel to the x_1 and x_2 axes. In the figure, for example, the major axis is x_1, and the minor axis is x_2.

10.2.3 Mixture Gaussian Models

A Gaussian distribution is only a good model if the data is really distributed in an ellipse. A better model for most distributions is a "mixture Gaussian model." The mixture Gaussian model can be thought of as a random choice between M different "Gaussian sub-classes." The probability of choosing sub-class G_{jk} is a constant, c_{jk}:

$$p(G_{jk}|\lambda_j) = c_{jk} \tag{10.10}$$

Then, once we have chosen a particular sub-class, the probability of the output o is calculated using the appropriate Gaussian model:

$$p(o|G_{jk}) = \mathcal{N}(o; \mu_{jk}, U_{jk}) \tag{10.11}$$

Putting these together, we get

$$p(o|\lambda_j) = \sum_{k=1}^{M} p(o|G_{jk}) p(G_{jk}|\lambda_j) = \sum_{k=1}^{M} c_{jk} \mathcal{N}(o; \mu_{jk}, U_{jk}) \tag{10.12}$$

An example of a series of Gaussian probability densities which might be added together to give a mixture Gaussian is shown in the bottom panel of figure 10.4. Notice that a mixture Gaussian can represent a non-elliptical probability density; in fact, if M is large enough, a mixture Gaussian can represent any probability density with arbitrarily good precision.

10.2.4 Sources of Error in Pattern Classification

10.2.5 Linear and Discriminant Features

10.2.6 Feedforward Neural Networks

Figure 10.5 shows the ith level-one neuron in a feed-forward neural network. The output of the neuron model, $y_i(x_1, x_2)$ is the result of linearly combining x_1 and x_2, shifting by some constant a_0, and passing the result through a sigmoidal (S-shaped) nonlinearity $f()$:

$$y_i(x_1, x_2) = f(a_0 + a_1 x_1 + a_2 x_2), \qquad f(x) \approx \begin{cases} 0 & x < 0 \\ 1 & x > 1 \end{cases} \tag{10.13}$$

$y_i(x_1, x_2)$ is a classification function, which is one over almost half of the (x_1, x_2) plane, zero over almost half of the plane, and between one and zero in some transition region (shaded in Fig. 2):

$$y_i(x_1, x_2) = \begin{cases} 0 & a_0 + a_1 x_1 + a_2 x_2 < 0 \\ 0 < y_i < 1 & \text{transition region} \\ 1 & a_0 + a_1 x_1 + a_2 x_2 > 1 \end{cases} \tag{10.14}$$

10.2.7 Talker Adaptation

10.3 Recognition of Words

Now suppose that, instead of a single spectrum, you are given a sequence of spectra, of the form

$$O = [o_1, \ldots, o_t, \ldots, o_T] \tag{10.15}$$

For simplicity, let's start by assuming that O is the spectrogram of a single word. Remember that, in order to do maximum likelihood speech recognition, we need "models" λ_i of every possible word. Each of these "models" must consist of a functional specification and a list of trainable parameters which will allow us to compute $p(O|\lambda_i)$. The functional specification must allow us to classify sequences of unknown length, since we don't know in advance how long T may be.

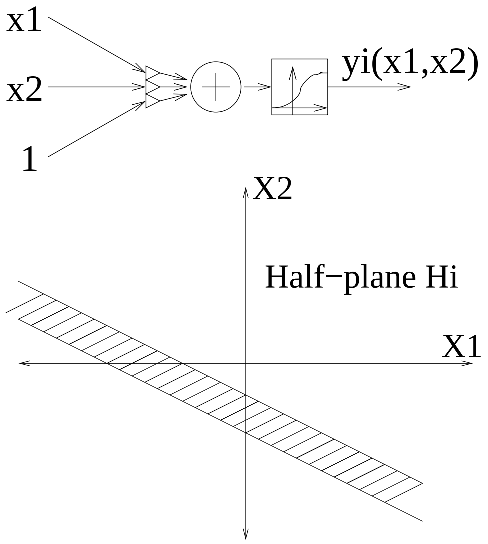


Figure 10.5: Flow-chart and classification space of a single-neuron neural network

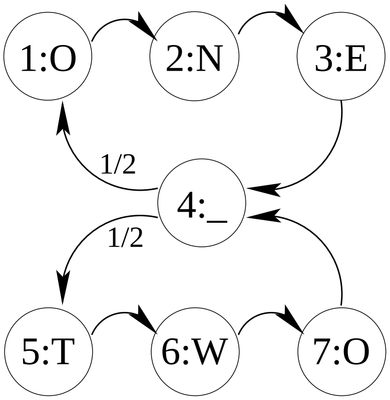


Figure 10.6: A model which generates a random sequence of ones and twos.

10.3.1 Linear Time Warping

10.3.2 Dynamic Time Warping

10.3.3 Hidden Markov Models: Testing

Imagine a process in which some person or computer is writing the words "ONE" and "TWO," in random order on a strip of paper. What is the probability of observing the sequence of symbols "ONE ONE TWO ONE TWO TWO ONE"?

The person or computer writing these symbols can be modeled as a simple finite state machine, as shown in figure 10.3.3.

Note the following facts:

- The model satisfies the Markov assumption. The Markov assumption states that the probability that the state at time $t+1$ is $q_{t+1} = j$ is a function only of q_t. This probability is called the "transition probability" a_{ij}:

$$P(q_{t+1} = j | q_t = i, q_{t-1} = h, \ldots) = P(q_{t+1} = j | q_t = i) \equiv a_{ij} \tag{10.16}$$

- Notice that we must make a distinction between the state q_t and the observed symbol o_t, because states 1 and 7 output the same symbol ("O"), but their transition probabilities are quite different.

- The state $q_{t+1} = j$ can be reached only if the model executes a transition from i to j, where i is whatever the current state happens to be. The probability $P(q_{t+1} = j)$ can be calculated by summing over i:

$$P(q_{t+1} = j) = \sum_{i=1}^{N} P(q_t = i) a_{ij} \quad \text{or, in matrix notation,} \quad P_{t+1} = P_t A \tag{10.17}$$

$$P_t = [P(q_t = 1), \ldots,\ P(q_t = N)], \qquad A = \begin{bmatrix} a_{11} & \ldots & a_{1N} \\ \vdots & \vdots & \vdots \\ a_{N1} & \ldots & a_{NN} \end{bmatrix} \tag{10.18}$$

- Given the transition probabilities a_{ij} and the initial state probabilities $\pi_i = P(q_1 = i)$, the probability of any particular sequence of states is

  $$P(q_1 = i, q_2 = i, q_3 = k, \ldots) = \pi_i a_{ij} a_{jk} \ldots, \qquad \pi_i \equiv P(q_1 = i) \tag{10.19}$$

  The probability of being in state m at time t is the sum over all intermediate states,

  $$P(q_t = m) = \sum_{i=1}^{N} \sum_{j=1}^{N} \ldots \sum_{l=1}^{N} \pi_i a_{ij} \ldots a_{lm}, \quad \text{or, in matrix notation,} \quad P_t = A^{t-1} \Pi \tag{10.20}$$

  $$\Pi = [\pi_1, \ldots, \pi_N], \qquad A = \begin{bmatrix} a_{11} & \ldots & a_{1N} \\ \vdots & \vdots & \vdots \\ a_{N1} & \ldots & a_{NN} \end{bmatrix} \tag{10.21}$$

- For example, using the process shown, the sequence of observations

  $$O = [o_1,\ o_2, \ldots,\ o_T] = \text{" ONE TWO"} \tag{10.22}$$

  corresponds to the sequence of states

  $$Q = [q_1,\ q_2, \ldots,\ q_T] = [4,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7] \tag{10.23}$$

  which occurs with a probability of

  $$P(Q | q_1 = 4) = a_{41} a_{12} a_{23} a_{34} a_{45} a_{56} a_{67} = (1/2)(1)(1)(1)(1/2)(1)(1) = 1/4 \tag{10.24}$$

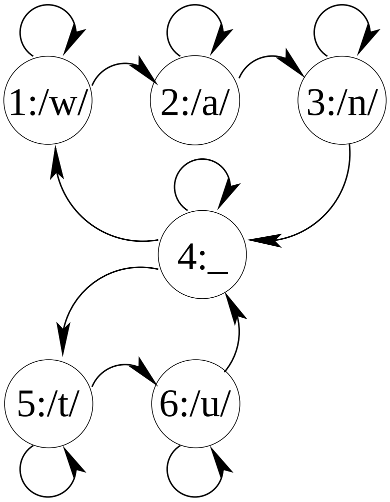


Figure 10.7: A model of a process which speaks the words "one" and "two" in random order

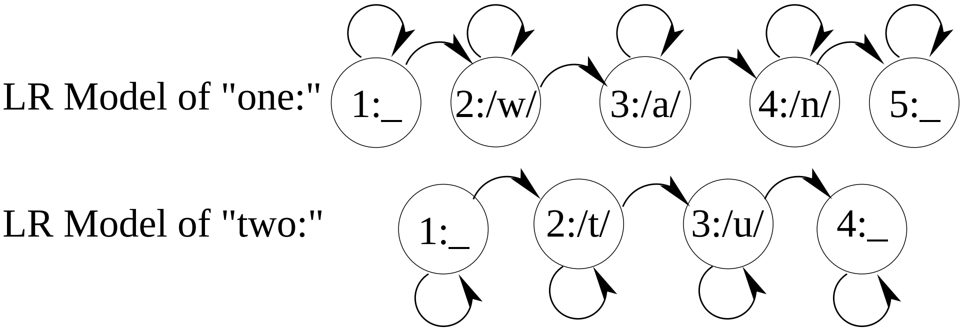


Figure 10.8: Left-to-right Markov models of the words "one" and "two"

The example above is "timed" such that there is one clock tick per symbol. In speech, there is no master clock to tell us when a new symbol has been spoken; instead, we need to analyze the speech signal in fixed time increments of, say, 10ms each. If each state now represents one phoneme, then it becomes necessary to introduce self-loop transition probabilities a_{ii} in order to model the variable duration of a phoneme. For example, if the words "one" and "two" are transcribed /w/a/n/ and /t/u/, the model above becomes something like the model shown in figure 10.7. Transition probabilities are not shown, but if the frame size is only 10ms, then $a_{ii} \gg a_{ij}$ for any $i \neq j$.

An "ergodic" process is a process in which any state can be reached (eventually) from any other state, as shown above. A process which generates one word and then quits can be modeled as a "left-to-right" process, as shown below. A "left-to-right" process is a process in which the transition matrix A is upper triangular, so that all transitions have to happen from left to right.

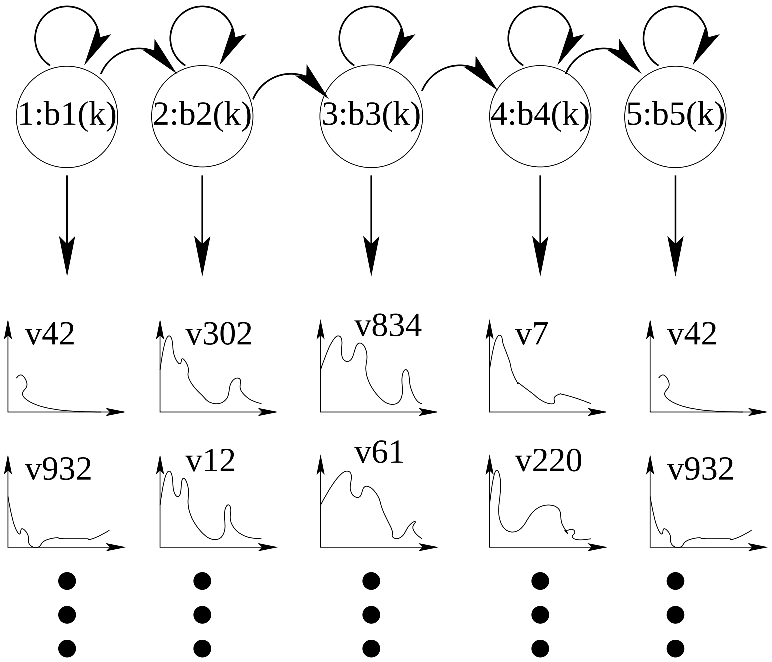


Figure 10.9: A hidden Markov model generates spectral vectors based on some internal state; the internal state of the model can never be known with certainty.

In the models above, we assumed that the state q_t is directly observable. In speech, the "states" are not observed — instead, all that we observe are the "outputs," which are vectors v_k containing the spectral, cepstral, or LPC information at time t.

Suppose that that there are only M possible spectral vectors, numbered from v_1 to v_M. Then a hidden Markov model is defined by the initial probabilities $\Pi = [\pi_i]$, the transition probabilities $A = [a_{ij}]$, and the discrete observation probabilities $B = [b_j(k)]$:

$$\begin{aligned}
\lambda &= (\pi_i, a_{ij}, b_j(k)), \qquad 1 \le i \le N,\ 1 \le j \le N,\ 1 \le k \le M && (10.25)\\
\pi_i &= p(q_1 = i) && (10.26)\\
a_{ij} &= p(q_t = j | q_{t-1} = i) && (10.27)\\
b_j(k) &= p(o_t = v_k | q_t = j) && (10.28)
\end{aligned}$$

Example 10.3.1 Discrete Hidden Markov Model

Consider an experiment in which your friend is tossing two coins behind a curtain, and yelling out the result of each coin toss. Your friend is switching back and forth between two coins, but he is not going to tell you when he switches. All you know is that the probability of a coin change on any given toss is always 25%:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{bmatrix} \qquad (10.29)$$

Furthermore, you do not know exactly *which* two coins your friend is using. You know that one of the coins is fair, but the other coin might be the "head-weighted" coin (which produces heads 75% of the time) or the "tail-weighted" coin (which produces tails 75% of the time). The two models you have available are:

$$B_1 = \begin{bmatrix} b_1(H) & b_1(T) \\ b_2(H) & b_2(T) \end{bmatrix} = \begin{bmatrix} 0.5 & 0.5 \\ 0.75 & 0.25 \end{bmatrix}, \quad B_2 = \begin{bmatrix} b_1(H) & b_1(T) \\ b_2(H) & b_2(T) \end{bmatrix} = \begin{bmatrix} 0.5 & 0.5 \\ 0.25 & 0.75 \end{bmatrix} \tag{10.30}$$

Finally, you are told that your friend always starts with the unfair coin, regardless of *which* unfair coin he is using:

$$\Pi_1 = [P(q_1 = 1), \quad P(q_1 = 2)] = [0, \ 1] \tag{10.31}$$

If your friend yells out the following sequence, is he using the head-weighted coin or the tail-weighted coin?

$$O = [o_1, \ldots, \ o_T] = [H, T] \tag{10.32}$$

- Case 1: Head-weighted coin. The probability of the sequence "HT" is

$$P(O|\lambda_1) = \sum_{j=1}^{2}\sum_{i=1}^{2} b_j(T)a_{ij}b_i(H)\pi_i = \sum_{j=1}^{2} b_j(T)a_{2j}b_2(H) = (0.5\times0.25\times0.75)+(0.25\times0.75\times0.75 \tag{10.33}$$

- Case 2: Tail-weighted coin. The probability of the sequence "HT" is

$$P(O|\lambda_2) = \sum_{j=1}^{2} b_j(T)a_{2j}b_2(H) = (0.5 \times 0.25 \times 0.25) + (0.75 \times 0.75 \times 0.25) = \frac{11}{64} \tag{10.34}$$

So the sequence "HT" is more likely to be produced if your friend starts with the head-weighted coin — model 1.

The observations in a hidden Markov model may be continuous random variables, distributed according to a mixture Gaussian distribution:

$$b_j(o) = p(o_t = o|q_t = j) = \sum_{k=1}^{M} c_{jk}\mathcal{N}(o; \mu_{jk}, U_{jk}), \qquad 1 \le j \le N \tag{10.35}$$

In this case, rather than specifying a set of discrete probabilities $b_j(k)$, a model is specified by finding the mixture weights, the means, and the covariance matrices:

$$\lambda = (\pi_i, a_{ij}, c_{jk}, \mu_{jk}, U_{jk}), \qquad 1 \le i \le N, \ 1 \le j \le N, \ 1 \le k \le M \tag{10.36}$$

Suppose that you have hidden Markov models of two words. The word "one" is represented by the model $\lambda_1 = (A_1, B_1, \pi_1)$. The word "two" is represented by the model $\lambda_2 = (A_2, B_2, \pi_2)$. Suppose, finally, that you observe a sequence of spectral vectors of the form:

$$O = [o_1, \ldots,\ o_t, \ldots,\ o_T] \tag{10.37}$$

Speech recognition (given that you have already somehow trained the models λ_1 and λ_2) boils down to the following problem: Given two models, λ_1 and λ_2, which model is most likely to have produced the observation sequence O? That is, which model maximizes the likelihood $p(O|\lambda)$?

Suppose we know for a fact that the model went through the following state sequence:

$$Q = [q_1,\ q_2, \ldots,\ q_T] \tag{10.38}$$

The probability of O given Q and λ is

$$p(O|Q, \lambda) = b_{q_T}(o_T) b_{q_{T-1}}(o_{T-1}) \ldots b_{q_1}(o_1) \tag{10.39}$$

The probability of Q is

$$p(Q|\lambda) = a_{q_{T-1}q_T} \ldots a_{q_1 q_2} \pi_{q_1} \tag{10.40}$$

Combining these two equations, we get the following:

$$p(O, Q|\lambda) = b_{q_T}(o_T) a_{q_{T-1}q_T} b_{q_{T-1}}(o_{T-1}) \ldots a_{q_1 q_2} b_{q_1}(o_1) \pi_{q_1} \tag{10.41}$$

Example 10.3.2 Phone Recognition Using an HMM

As a simple application of the HMM, consider a system which records a person saying "yes" in either Japanese or Swedish, and then identifies the language.

Being the polyglot that you are, you know that yes in Japanese is "hai," and yes in Swedish is "ja." In order to keep complexity down, let's assume that "hai" is approximately an /a/ sound followed by an /i/ sound, while "ja" is approximately an /i/ sound followed by an /a/ sound, as shown in figure 10.10. Suppose further that both models always start in the first state shown, i.e. "hai" always starts in /a/, and "ja" always starts in /i/.

The only spectral measurement available is the second formant frequency, F2. Means and standard deviations of F2 for /i/ and /a/ can be approximated by combining the data of Peterson and Barney (1952) for adult male and female speakers, yielding the following observation probability densities:

$$\begin{aligned} b_{/i/}(o) &= \mathcal{N}(o; 2540\text{Hz}, (350\text{Hz})^2) && (10.42) \\ b_{/a/}(o) &= \mathcal{N}(o; 1160\text{Hz}, (150\text{Hz})^2) && (10.43) \end{aligned}$$

Now suppose that we are given the following (very short!) unknown utterance:

$$O = [\ 1800,\quad 1500\ ] \tag{10.44}$$

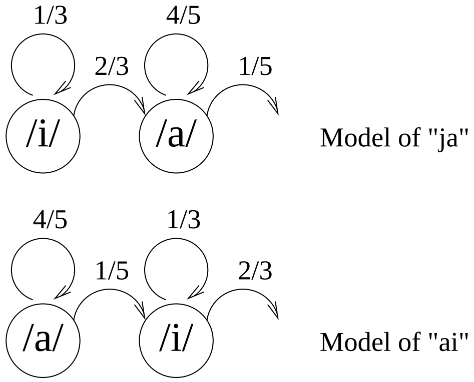


Figure 10.10: Simple Markov models of the words “hai” (/ɑi/, if we ignore the /h/) and “ja” (/ia/, if we pretend that /j/ and /i/ are the same). Transition probabilities are designed so that the /i/ states last an average of 1.5 frames, and the /ɑ/ states last an average of 5 frames.

Frame	o (kHz)	$\mathcal{N}(o; 2.54, 0.35^2)$	$\mathcal{N}(o; 1.15, 0.15^2)$
1	1.8	0.12	0.014
3	1.5	0.00022	0.17

Table 10.1: Column 3 is an estimate of the probability that the F2 values in column 2 are produced as part of an /i/ vowel. Column 3 is an estimate of the probability that the F2 values are produced as part of an /a/ vowel. Both columns 3 and 4 show probability density per kilohertz, assuming Gaussian distributions.

Which language is this person speaking?

The only information we have about the distribution of F2 is its mean and standard deviation, so let’s use Gaussian observation probability densities. The observation probability densities of the /i/ model in “ja” and the /i/ model in “hai” happen to be exactly the same (in speech recognition, we say that these two states have “tied” observation densities), as are the distributions of /a/ in both “ja” and “hai.” The observation densities $b_{/i/}(o)$ and $b_{/a/}(o)$ are given in table 10.1.

If the person is speaking Japanese, the model must have started in /a/; it might have stayed in /a/ for the second frame, or it might have transitioned to /i/. Adding up the likelihoods of both possibilities, we get:

$$p(O|\text{“hai”}) = b_{/a/}(1800)a_{/a/a/}b_{/a/}(1500) + b_{/a/}(1800)a_{/a/i/}b_{/i/}(1500) \quad (10.45)$$

$$= 0.00026 \quad (10.46)$$

Likewise, the likelihood that the person is speaking Swedish is

$$p(O|\text{“ja”}) = b_{/i/}(1800)a_{/i/i/}b_{/i/}(1500) + b_{/i/}(1800)a_{/i/a/}b_{/a/}(1500) \quad (10.47)$$

$$= 0.0101 \quad (10.48)$$

The observed falling F2 pattern seems much more likely to have come from the word “ja” than from the word “hai.” If the two words have equal *a priori* probabilities, then we can be pretty confident in declaring that the word was “ja.”

Maximum Likelihood Recognition: The Forward Algorithm

In order to do ML classification correctly, we need the probability $p(O|\lambda)$. $p(O|\lambda)$ can be obtained from $p(O,Q|\lambda)$ by summing over all possible state sequences Q:

$$\begin{aligned} P(O|\lambda) &= \sum_{Q} P(O,Q|\lambda) && (10.49) \\ &= \sum_{q_T} \cdots \sum_{q_2} \sum_{q_1} b_{q_T}(o_T) a_{q_{T-1}q_T} b_{q_{T-1}}(o_{T-1}) \ldots b_{q_2}(o_2) a_{q_1 q_2} b_{q_1}(o_1) \pi_{q_1} && (10.50) \end{aligned}$$

Suppose that we decide to add up all of the information from time $t = 1$ as soon as we have it, then add up all of the information from time $t = 2$, and so on. At time $t = 1$, we define

$$\alpha_1(j) = p(o_1, q_1 = j|\lambda) = b_j(o_1)\pi_j \tag{10.51}$$

At time $t = 2$, we have

$$\alpha_2(j) = p(o_1, o_2, q_2 = j|\lambda) = b_j(o_2) \sum_{i=1}^{N} a_{ij}\alpha_1(i) \tag{10.52}$$

Likewise, at every new time until $t = T$, we have

$$\alpha_t(j) = p(o_1, \ldots, o_t, q_t = j|\lambda) = b_j(o_t) \sum_{i=1}^{N} a_{ij}\alpha_{t-1}(i) \tag{10.53}$$

Finally, at time T, we see that

$$p(O|\lambda) = p(o_1, \ldots, o_T|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{10.54}$$

Equations 10.51 through 10.54 are called the “forward algorithm,” because the iteration moves forward in time; the same thing could also be done backward in time.

The Backward Algorithm

Equation 10.50 can also be broken down into a recursion which moves *backward* in time:

$$P(O|\lambda) = \sum_{q_1} \pi_{q_1} b_{q_1}(o_1) \sum_{q_2} a_{q_1 q_2} b_{q_2}(o_2) \ldots \sum_{q_T} a_{q_{T-1}q_T} b_{q_T}(o_T) \tag{10.55}$$

This recursion is written most simply if we define the backward variable $\beta_t(i)$:

$$\beta_t(i) \equiv P(o_{t+1},\ o_{t+2}, \ldots,\ o_T | q_T = i,\ \lambda) \tag{10.56}$$

Then equation 10.55 is calculated using:

1. Initialization
$$\beta_T(i) = 1, \quad 1 \leq i \leq N \tag{10.57}$$
2. Induction
$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \tag{10.58}$$
3. Termination
$$P(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i) \tag{10.59}$$

The calculation of this procedure is about the same as the forward algorithm, but it is not used as often in recognition, because the recursion works backward in time. However, this algorithm is often used in segmentation and training.

10.3.4 Approximate Recognition: The Viterbi Algorithm

Suppose that, for whatever reason, we don't want to do the addition in equation 10.53 at every time step for every possible combination of states. If this is the case, we can do a sort of approximate maximum likelihood classification. Instead of finding $p(O|\lambda)$, we find $p(O, Q|\lambda)$ for the best possible state sequence, which we can call $Q^*(O, \lambda)$:

$$\begin{aligned} Q^*(O, \lambda) &= \arg\max_Q p(O, Q|\lambda) && (10.60) \\ P^*(O, \lambda) &= \max_Q p(O, Q|\lambda) && (10.61) \\ &= \max_{q_T} \ldots \max_{q_2} \max_{q_1} b_{q_T}(o_T) a_{q_{T-1}q_T} b_{q_{T-1}}(o_{T-1}) \ldots b_{q_2}(o_2) a_{q_1 q_2} b_{q_1}(o_1) \pi_{q_1} && (10.62) \end{aligned}$$

Then, in order to decide which sequence of words is the correct sequence, we just look for the model which gives us the largest P^*.

Suppose we decide to do the $\max_{q_1}$ operation as soon as we have all of the information from time $t = 1$, and then do the $\max_{q_2}$ operation as soon as we have all of the information from time $t = 2$, and so on. At time $t = 1$, we can define

$$\delta_1(i) = \pi_i b_i(o_1) \tag{10.63}$$

Then, at every new time t until $t = T$, we find the best path which ends up in state j. $\delta_t(j)$ keeps track of the maximum probability, and $\psi_t(j)$ is a "back-pointer" which points backward from state j to the best previous state:

$$\delta_t(j) = b_j(o_t) \max_{1 \le i \le N} \delta_{t-1}(i) a_{ij} \tag{10.64}$$

$$\psi_t(j) = \arg \max_{1 \le i \le N} \delta_{t-1}(i) a_{ij} \tag{10.65}$$

Finally, at time T, we see that the best final probability is

$$P^*(O|\lambda) = \max_{1 \le i \le N} \delta_T(i) \tag{10.66}$$

and the best state in which to end up is:

$$q_T^* = \arg \max_{1 \le i \le N} \delta_T(i) \tag{10.67}$$

We can find out what state sequence yielded P^* by working our way backward in time, from time $t = T$ to time $t = 1$, following the "back-pointers" given by the $\psi_t(i)$ variables:

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \qquad 1 \le t \le T-1 \tag{10.68}$$

$$Q^*(O|\lambda) = [\ q_1^*,\ \ q_2^*, \ldots q_T^*\ ] \tag{10.69}$$

Local Recognition: The Forward-Backward Algorithm

Suppose we are only interested in finding the state at time t which maximizes:

$$\gamma_t(i) = P(q_t = i|O, \lambda) = \frac{P(O, q_t = i|\lambda)}{P(O|\lambda)} \tag{10.70}$$

This can be calculated as follows:

$$P(O, q_t = i|\lambda) = P(o_1,\ o_2, \ldots,\ o_t,\ q_t = i,\ o_{t+1},\ o_{t+2},\ \ldots,\ o_T|\lambda) \tag{10.71}$$

$$= P(o_1,\ o_2, \ldots,\ o_t,\ q_t = i|\lambda) P(o_{t+1},\ o_{t+2},\ \ldots,\ o_T|q_T = i, \lambda) \tag{10.72}$$

$$= \alpha_t(i)\beta_t(i) \tag{10.73}$$

therefore,

$$\gamma_t(i) = \frac{P(O, q_t = i|\lambda)}{\sum_{i=1}^N P(O, q_t = i|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^N \alpha_t(i)\beta_t(i)} \tag{10.74}$$

The most likely state q_t at time t is therefore the state which maximizes $\gamma_t(i)$.

Now suppose we are only interested in finding the most likely two-frame sequence of states. In other words, we would like to find i and j to maximize

$$\xi_t(i,j) = P(q_t = i, q_{t+1} = j|O, \lambda) = \frac{P(q_t = i, q_{t+1} = j, O|\lambda)}{P(O|\lambda)} \tag{10.75}$$

By calculations similar to the calculations for $\gamma_t(j)$, we can show that this probability is

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)} \tag{10.76}$$

So the most likely two-state sequence at times t and $t+1$ is the sequence which maximizes $\xi_t(i,j)$.

Notice that the likely two-frame sequence $q_t = i, q_{t+1} = j$ may not be part of a likely state sequence spanning the entire utterance. If you want to find a state sequence which spans the entire utterance, you have to use the Viterbi algorithm.

Duration Probabilities and Transition Probabilities

In the original HMM model, the probability of remaining in state i for d_i time steps is a geometric PMF:

$$p_i(d) = (1 - a_{ii}) a_{ii}^{d-1} \quad \text{if } a_{ii} \text{ independent of } d \tag{10.77}$$

The geometric PMF is a bad model of the distribution of real phonemes or phone-like units. For this reason, it is sometimes useful to train and use an explicit model of the duration PMF. Given an explicit model of $p_i(d)$, it is possible to calculate duration-dependent transition probabilities $a_{ij}(d)$ as follows:

$$a_{ij}(d) = \begin{cases} P(d_i > d | d_i \geq d) & j = i \\ P(d_i = d,\ q_{t+1} = j | d_i \geq d) & j \neq i \end{cases} \tag{10.78}$$

If we assume that the model still has no long-term memory, except that $a_{ii}(d)$ is a function of duration, then the following formulas result:

$$a_{ij}(d_i) = \begin{cases} \frac{1 - \sum_{d=1}^{d_i} p_i(d)}{1 - \sum_{d=1}^{d_i - 1} p_i(d)} & j = i \\ \breve{a}_{ij}(1 - a_{ii}(d_i)) & j \neq i \end{cases} \tag{10.79}$$

where the parameters $\breve{a}_{ij}$ are transition probabilities conditioned on a change in state:

$$\breve{a}_{ij} \equiv P(q_{t+1} = j | q_t = i, q_{t+1} \neq i) \tag{10.80}$$

Suppose that, in T time steps, an HMM sequentially visits S states:

$$q_t = r_s, \qquad t_s < t \leq t_s + d_s \tag{10.81}$$

Suppose that the model remains in state r_s for d_s time steps before transitioning to some other state. Then the event Q is the intersection of two events: a "transitions" event R, and a "durations" event D:

$$Q = R \cap D, \qquad R = [r_1, \ldots,\ r_S], \quad D = [d_1, \ldots,\ d_S] \tag{10.82}$$

Assuming that the various state durations and transitions are independent, the probabilities of these two events are

$$P(R|\lambda) = \prod_{s=1}^{S} \breve{a}_{r_{s-1}r_s}, \qquad \breve{a}_{r_0 r_1} \equiv \pi_{r_1} \tag{10.83}$$

$$P(D|\lambda, R) = \prod_{s=1}^{S} p_{r_s}(d_s) \tag{10.84}$$

The probability of any particular state sequence Q can be calculated in terms of the probabilities $p_i(d)$ and $\breve{a}_{ij}$:

$$\begin{aligned} P(O,Q|\lambda) &= \prod_{s=1}^{S} \left(\breve{a}_{r_{s-1}r_s} p_{r_s}(d_s) \prod_{\tau=1}^{d_s} b_{r_s}(o_{t_s+\tau}) \right) && (10.85) \\ &= P(D|\lambda, R) \prod_{t=1}^{T} \tilde{a}_{q_{t-1}q_t} b_{q_t}(o_t) && (10.86) \\ &= P(D|\lambda, R) P(O,Q|\tilde{\lambda}) && (10.87) \end{aligned}$$

where the pseudo-model $\tilde{\lambda}$ is

$$\tilde{\lambda} \equiv [\pi_i,\ \tilde{a}_{ij},\ b_j(o)], \qquad \tilde{a}_{ij} \equiv \begin{cases} 1 & i = j \\ \breve{a}_{ij} & i \neq j \end{cases} \tag{10.88}$$

The recognition probability $P(O|\lambda)$ is computed by adding up $P(O,Q|\lambda)$ over all possible Q:

$$P(O|\lambda) = \sum_{\text{all } Q} P(D|\lambda, R) P(O,Q|\tilde{\lambda}) \tag{10.89}$$

Approximate Duration Modeling using Viterbi and Forward Algorithms Equation 10.89 can be expressed as a recursion, similar to the forward algorithm, but the computational complexity of the resulting algorithm is so high that it is rarely used (essentially, the HMM is augmented to include ND states, where D is the maximum possible state duration). Instead, many recognizers use the Viterbi algorithm in parallel with the forward algorithm in order to compute an approximate recognition probability.

In the parallel approximation, $P(O|\lambda)$ is computed as follows:

$$P(O|\lambda) \approx P(D^*|\lambda, R^*) \sum_{\text{all } Q} P(O,Q|\tilde{\lambda}) = P(D^*|\lambda, R^*) P(O|\tilde{\lambda}) \tag{10.90}$$

The quantity $P(O|\tilde{\lambda})$ can be computed using the forward algorithm. $P(D^*|\lambda, R^*)$ is the probability of the state durations associated with the single maximum-likelihood state sequence Q^*, as returned by a Viterbi search:

$$Q^* = \arg\max_Q P(O,Q|\lambda), \qquad Q^* = D^* \cap R^* \tag{10.91}$$

10.3.5 Hidden Markov Models: Training

The goal of training a hidden Markov model is that the parameter π_i, for example, should be proportional to the number of times that the model started in state i out of all of the observed training tokens. In other words, we would like to have model parameters which look something like this:

$$\pi_i = \frac{\text{number of times in state } i \text{ at time } (t=1)}{\text{total number of training utterances}} \tag{10.92}$$

$$a_{ij} = \frac{\text{number of transitions from state } i \text{ to state } j}{\text{total number of transitions out of state } i} \tag{10.93}$$

$$c_{jk} = \frac{\text{number of times we choose Gaussian sub-class } k \text{ while in state } j}{\text{total number of times in state } j} \tag{10.94}$$

$$\mu_{jk} = \frac{\text{sum of } o_t \text{ for all frames spent in class } j\text{, sub-class } k}{\text{total number of times spent in class } j\text{, sub-class } k} \tag{10.95}$$

$$U_{jk} = \frac{\text{sum of } (o_t-\mu_{jk})'(o_t-\mu_{jk}) \text{ for all frames spent in class } j\text{, sub-class } k}{\text{total number of times spent in class } j\text{, sub-class } k} \tag{10.96}$$

The big problem with equations 10.92 through 10.96 is that we don't *know* how often the model is in state j — remember, the state transitions are "hidden"! There are (at least) two different ways to solve this problem: the segmental K-means algorithm (which is typically used to initialize the parameters of an HMM), and the Baum-Welch re-estimation procedure (which is typically used to refine a previously-estimated set of parameters).

Initializing the Observation Densities: Segmental K-Means

The segmental K-means algorithm is based on equations 10.92 through 10.96. In segmental K-means, we use the Viterbi algorithm to figure out which state the model is in at any given time, in a sort of boot-strapping procedure which goes like this:

1. Divide each of the training utterances into N segments, where N is the number of states in the model. Any arbitrary segmentation will work, although a phonetically motivated segmentation often leads to faster convergence.

2. In order to initialize state number j, gather observation vectors from the jth segment in each segmented training utterance. Call these training vectors y_{nj},

$$y_{nj} = o_t \text{ iff } t \text{ is in segment number } j \text{ of some training utterance} \tag{10.97}$$

3. For each state j, cluster the training vectors y_{nj} into M regions V_{jk} using a bottom-up clustering algorithm. Let N_{jk} be the number of vectors in V_{jk}, and let x_{jk} be the centroid of V_{jk}; then the new observation density parameters are

$$c_{jk} = \frac{N_{jk}}{N_j} \tag{10.98}$$

$$\mu_{jk} = x_{jk} \tag{10.99}$$

$$U_{jk} = \frac{1}{N_{jk}} \sum_{y_{n,j} \in V_{jk}} (y_{nj} - x_{jk})'(y_{nj} - x_{jk}) \tag{10.100}$$

4. Given the new parameter estimates, use the Viterbi algorithm to resegment each of the training utterances. If the new segmentation is different from the previous segmentation, go to step number 2.

Refining the Model: Baum-Welch Algorithm

There are several algorithms available for training hidden Markov model parameters, depending on the criterion which you want to optimize. The most common algorithm is the Baum-Welch algorithm, also called the Expectation-Maximization (EM) method. The algorithm works like this: suppose we define the function

$$f(O, Q, \lambda_2) = \log P(O, Q|\lambda_2) \tag{10.101}$$

A reasonable goal of parameter re-estimation would be to maximize the expected value of f over all possible Q, given the model λ_2:

$$E[f(O, Q, \lambda_2)|\lambda_2] = \sum_Q P(O, Q|\lambda_2) \log P(O, Q|\lambda_2) \tag{10.102}$$

Unfortunately, given the structure of an HMM, equation 10.102 can not be maximized in just one step. Instead, the Baum-Welch algorithm tries to maximize equation 10.102 iteratively: first we guess a model λ_1, then we find a new model λ_2 which maximizes

$$E[f(O, Q, \lambda_2)|\lambda_1] = \sum_Q P(O, Q|\lambda_1) \log P(O, Q|\lambda_2) \tag{10.103}$$

Iterating equation 10.103 several times moves the model λ_2 toward a local maximum of equation 10.102.

It can be shown that, given λ_1, the function in equation 10.103 is maximized if the parameters of λ_2 are as follows:

$$\bar{\pi}_i = \text{expected number of times in state } i \text{ at time } (t = 1) \tag{10.104}$$

$$\bar{a}_{ij} = \frac{\text{expected number of transitions from state } i \text{ to state } j}{\text{expected number of transitions out of state } i} \tag{10.105}$$

$$\bar{b}_j(k) = \frac{\text{expected number of times in state } j \text{ and observing symbol } v_k}{\text{expected number of times in state } j} \tag{10.106}$$

These expectations can be calculated:

$$E[\#\text{ times in state } i \text{ at time } t=1] = P(q_1 = i|O,\lambda) = \gamma_1(i) \tag{10.107}$$

$$E[\#\text{ times in state } i] = \sum_{t=1}^{T} P(q_t = i|O,\lambda) = \sum_{t=1}^{T} \gamma_t(i) \tag{10.108}$$

$$E[\#\text{ times in state } i \text{ and observing } o_t = v_k] = \sum_{t=1}^{T} P(q_t = i, o_t = v_k|O,\lambda) = \sum_{t \text{ s.t. } o_t=v_k} \gamma_t(i) \tag{10.109}$$

$$E[\#\text{ transitions from } i \text{ to } j] = \sum_{t=1}^{T} P(q_t = i, q_{t+1} = j|O,\lambda) = \sum_{t=1}^{T} \xi_t(i,j) \tag{10.110}$$

$$\tag{10.111}$$

Usually, we train an HMM in two steps. In the first step, the parameters of the mixture-Gaussian observation densities are initialized using the segmental K-means" algorithm — this is the algorithm used in the HTK program HInit, for example. The reason we don't stop there is that the segmental K-means algorithm assumes that the Viterbi algorithm will pick out all of the spectra which correspond to a particular state. In fact, the Viterbi algorithm only picks out the spectra which are *most likely* to have come from a particular state — the less likely spectra are usually assigned to the neighboring state. This means that the segmental K-means algorithm tends to underestimate the amount of variability which is really present in the training data.

In the second step, therefore, the Baum-Welch re-estimation procedure is used to refine the parameter estimates in order to find a "local maximum" of the conditional probability of the training data, $E[\log p(O|\lambda)]$. The words "local maximum" mean that, once you've run the Baum-Welch algorithm, if you then make a *small* change in the parameters, the performance of the recognizer will always get worse. It's entirely possible, however, that if you make a *big* change in the parameters, the performance might get better – in fact, this often happens! This odd and problematic behavior is the reason we initialize parameters first using the segmental K-means algorithm.

Gaussian Densities in a Hidden Markov Model

The observations in a hidden Markov model may be continuous random variables, distributed according to a Gaussian distribution:

$$b_j(o) = p(o|q_t = j) = \mathcal{N}(o; \mu_j, U_j), \qquad 1 \le j \le N \tag{10.112}$$

If we are given a model λ which includes initial estimates of μ_j and U_j for each state, it is possible to calculate the forward and backward variables $\alpha_t(j)$ and $\beta_t(j)$, and to multiply them to find

$$\gamma_t(j) = P(q_t = j|\ O, \lambda) = \frac{\alpha_t(j)\beta_t(j)}{P(O|\ \lambda)} \qquad 1 \le j \le N,\ 1 \le t \le T \tag{10.113}$$

Using the statistic $\gamma_t(j)$, the expected number of times that the model visits state j in T time steps is

$$E[N_j|\ O, \lambda] = \sum_{t=1}^{T} 1 \times P(q_t = j|\ O, \lambda) + 0 \times P(q_t \neq j|\ O, \lambda) = \sum_{t=1}^{T} \gamma_t(j) \qquad 1 \le j \le N \tag{10.114}$$

Suppose we wish to re-estimate the value μ_j, the mean of o in state j. A good estimate would be

$$\bar{\mu}_j \text{ is something like } \left(\frac{1}{N_j}\right) \sum_{t \text{ s.t. } q_t = j} o_t \qquad 1 \le j \le N \tag{10.115}$$

Unfortunately, both the numerator and the denominator are random variables, so we need to take expected values:

$$\bar{\mu}_j = \frac{E\left[\sum_{t \text{ s.t. } q_t = j} o_t | O, \lambda\right]}{E[N_j|\ O, \lambda]} = \frac{\sum_{t=1}^{T} o_t \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \qquad 1 \le j \le N \tag{10.116}$$

This turns out to be the maximum-likelihood re-estimation value $\bar{\mu}_j$. The maximum-likelihood updated estimate of the covariance, $\bar{U}_j$, is

$$\bar{U}_j = \frac{E[\sum_{t \text{ s.t. } q_t = j} (o_t - \mu_j)(o_t - \mu_j)' | O, \lambda]}{E[N_j | O, \lambda]} = \frac{\sum_{t=1}^{T} (o_t - \mu_j)(o_t - \mu_j)' \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \qquad 1 \le j \le N \tag{10.117}$$

Mixture Gaussian Densities

Re-estimation for a mixture Gaussian model depends on the training statistic:

$$\gamma_t(j, k) = P(q_t = j, G_t = k | O, \lambda) = \gamma_t(j) P(G_t = k | q_t = j, O, \lambda) = \gamma_t(j) \left[\frac{c_{jk} \mathcal{N}(o_t; \mu_{jk}, U_{jk})}{\sum_{k=1}^{M} c_{jk} \mathcal{N}(o_t; \mu_{jk}, U_{jk})}\right] \tag{10.118}$$

Notice that

$$\gamma_t(j) = \sum_{k=1}^{M} \gamma_t(j, k) \tag{10.119}$$

Consider the number N_{jk}, the number of times that the model moves from state $q_t = j$ to Gaussian $G_t = k$. The re-estimation probabilities for the mixture Gaussian parameters are then

$$\bar{c}_{jk} = \frac{E[N_{jk} | O, \lambda]}{E[N_j | O, \lambda]} = \frac{\sum_{t=1}^{T} \gamma_t(j, k)}{\sum_{t=1}^{T} \gamma_t(j)} \qquad 1 \le j \le N,\ 1 \le k \le M \tag{10.120}$$

$$\bar{\mu}_j = \frac{E\left[\sum_{t \text{ s.t. } q_t=j, G_t=k} o_t | O, \lambda\right]}{E[N_{jk}|O,\lambda]} = \frac{\sum_{t=1}^{T} o_t \gamma_t(j,k)}{\sum_{t=1}^{T} \gamma_t(j,k)} \qquad 1 \le j \le N,\ 1 \le k \le M \tag{10.121}$$

$$\bar{U}_j = \frac{E[\sum_{t \text{ s.t. } q_t=j, G_t=k} (o_t-\mu_j)(o_t-\mu_j)'|O,\lambda]}{E[N_{jk}|O,\lambda]} = \frac{\sum_{t=1}^{T}(o_t-\mu_j)(o_t-\mu_j)'\gamma_t(j,k)}{\sum_{t=1}^{T}\gamma_t(j,k)} \qquad 1 \le j \le N, \tag{10.122}$$

Multiple Observation Sequences

Suppose we want to train a model using data from K different waveform files. Each waveform file is a sequence of data vectors,

$$O^{(k)} = [o_1^{(k)}, \ldots,\ o_T^{(k)}], \qquad 1 \le k \le K \tag{10.123}$$

and the total dataset consists of the "sequence of sequences,"

$$\mathbf{O} = [O^{(1)}, \ldots, O^{(K)}] \tag{10.124}$$

The expected values needed for Baum-Welch re-estimation are:

$$E[\#\text{ times in state } i \text{ at time } t=1] = \sum_{k=1}^{K} P(q_1 = i|O^{(k)},\lambda) = \sum_{k=1}^{K} \gamma_1^{(k)}(i)$$

$$E[\#\text{ times in state } i] = \sum_{k=1}^{K}\sum_{t=1}^{T} P(q_t = i|O^{(k)},\lambda) = \sum_{k=1}^{K}\sum_{t=1}^{T} \gamma_t^{(k)}(i)$$

$$E[\#\text{ times in state } i \text{ and observing } o_t = v_k] = \sum_{k=1}^{K}\sum_{t=1}^{T} P(q_t = i, o_t = v_k|O^{(k)},\lambda) = \sum_{k=1}^{K}\left(\sum_{t \text{ s.t. } o_t=}\right.$$

$$E[\#\text{ transitions from } i \text{ to } j] = \sum_{k=1}^{K}\sum_{t=1}^{T} P(q_t = i, q_{t+1} = j|O^{(k)},\lambda) = \sum_{k=1}^{K}\sum_{t=1}^{T} \xi_t^{(k)}(i,j)$$

where

$$\gamma_t^{(k)}(i) = P(q_t = i|O^{(k)},\lambda) = \frac{\alpha_t^{(k)}(i)\beta_t^{(k)}(i)}{P(O^{(k)}|\lambda)} \tag{10.130}$$

$$\xi_t^{(k)}(i,j) = P(q_t = i, q_{t+1} = j|O^{(k)},\lambda) = \frac{\alpha_t^{(k)}(i)a_{ij}b_j(o_{t+1}^{(k)})\beta_{t+1}^{(k)}(j)}{P(O^{(k)}|\lambda)} \tag{10.131}$$

Essentially, the algorithm is exactly the same as it would be with one file, except that if you are only using one file, the formulas for $\bar{a}_{ij}$, $\bar{b}_j(o_k)$, and $\bar{\pi}_i$ simplify in ways which are not possible if you are using multiple files.

Probability Scaling in the Forward-Backward Algorithm

Remember that the induction steps for the forward-backward algorithm are

$$\alpha_t(i) = b_i(o_t)\sum_{j=1}^{N}\alpha_{t-1}(j)a_{ji}, \qquad 1 \le i \le N \tag{10.132}$$

$$\beta_t(i) = \sum_{j=1}^{N}a_{ij}b_j(o_{t+1})\beta_{t+1}(j), \qquad 1 \le i \le N \tag{10.133}$$

$$\tag{10.134}$$

If, for example, $b_j(o_t)$ is calculated using a Gaussian density with covariance matrix U_j, then

$$b_j(o_t) < \frac{1}{|U_j|^{1/2}(2\pi)^{p/2}} = b_{max} \tag{10.135}$$

and therefore

$$\alpha_t(i) < b_{max}^t, \qquad \beta_t(i) < b_{max}^{T-t} \tag{10.136}$$

If $|U_j| > 1$ (as is usually the case), then $\alpha_t(i)$ and $\beta_t(i)$ approach zero very quickly; within 5-10 time steps, they can easily be smaller than the floating point resolution of the computer.

The solution involves computing scaled forward and backward variables, $\hat{\alpha}_t(i)$ and $\hat{\beta}_t(i)$. The scaled forward algorithm essentially re-normalizes the αs at every time step so that

$$\sum_{i=1}^{N}\hat{\alpha}_t(i) = 1 \tag{10.137}$$

We can get this normalization by calculating a scaling constant c_t at each time step, as follows:

1. Initialization

$$\hat{\alpha}_1(i) = c_1\alpha_1(i), \qquad c_1 = \frac{1}{\sum_{i=1}^{N}\alpha_1(i)} \tag{10.138}$$

2. Induction

$$\hat{\hat{\alpha}}_t(i) = b_i(o_t)\sum_{j=1}^{N}\hat{\alpha}_{t-1}(j)a_{ji} \tag{10.139}$$

$$\hat{\alpha}_t(i) = c_t\hat{\hat{\alpha}}_t(i), \qquad c_t = \frac{1}{\sum_{i=1}^{N}\hat{\hat{\alpha}}_t(i)} \tag{10.140}$$

Recognition Using the Scaled Forward Algorithm In the scaled forward algorithm, the scaling factors accumulate over time, so that

$$\hat{\alpha}_t(i) = \alpha_t(i) \prod_{\tau=1}^{T} c_\tau \tag{10.141}$$

The termination step in the normal forward algorithm is

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) = \frac{\sum_{i=1}^{N} \hat{\alpha}_T(i)}{\prod_{t=1}^{T} c_t} \tag{10.142}$$

However, remember that the constants c_T are chosen so that

$$\sum_{i=1}^{N} \hat{\alpha}_T(i) = 1 \tag{10.143}$$

Therefore

$$P(O|\lambda) = \frac{1}{\prod_{t=1}^{T} c_t} \tag{10.144}$$

The Scaled Backward Algorithm It turns out that training works best if the backward algorithm uses the same scaling factors c_t as the forward algorithm, as follows:

1. Initialization

$$\hat{\beta}_T(i) = c_T, \qquad c_1 = \frac{1}{\sum_{i=1}^{N} \alpha_1(i)} \tag{10.145}$$

2. Induction

$$\hat{\beta}_t(i) = c_t \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \hat{\beta}_{t+1}(j), \qquad c_t = \frac{1}{\sum_{i=1}^{N} \hat{\hat{\alpha}}_t(i)} \tag{10.146}$$

Re-Estimation Using Scaled Parameters Remember that the original re-estimation algorithm for a_{ij} was

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{10.147}$$

where the training statistics ξ_t and γ_t are defined to be

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)} \tag{10.148}$$

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} = \sum_{j=1}^{N} \xi_t(i,j) \tag{10.149}$$

$$\tag{10.150}$$

The scaled and unscaled forward and backward parameters are related by products of c_t, as follows:

$$\hat{\alpha}_t(i) = \alpha_t(i) \prod_{\tau=1}^{t} c_\tau, \qquad \hat{\beta}_{t+1}(j) = \beta_{t+1}(j) \prod_{\tau=t+1}^{T} c_\tau, \qquad P(O|\lambda) = \frac{1}{\prod_{t=1}^{T} c_t} \tag{10.151}$$

By combining equations 10.148 and 10.151, it is possible to write $\xi_t(i,j)$ in terms of $\hat{\alpha}$ and $\hat{\beta}$:

$$\xi_t(i,j) = \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \prod_{t=1}^{T} c_t = \hat{\alpha}_t(i) a_{ij} b_j(o_{t+1}) \hat{\beta}_{t+1}(j) \tag{10.152}$$

Calculating $\gamma_t(j)$ is a little trickier. γ is still the sum of ξ, but it is not simply the product of $\hat{\alpha}$ and $\hat{\beta}$:

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) = \hat{\alpha}_t(i) \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \hat{\beta}_{t+1}(j) = \frac{\hat{\alpha}_t(i)\hat{\beta}_t(i)}{c_t} \tag{10.153}$$

10.3.6 Pronunciation Modeling

In very large vocabulary recognizers, it is hard to find enough training data to train models of every word. Often, a better approach is to train models of phone-like units, and then create each word model by stringing together appropriate phone models.

Recognition

Suppose you want to test the hypothesis that the sequence of words in a test utterance is

$$W = [w_1, w_2, \ldots, w_q, \ldots, w_Q] \tag{10.154}$$

In a phone-based recognizer, the first step in recognition is to find each of the words w_q in a look-up table called a "lexicon." The "lexicon" gives all of the known possible pronunciations of the word, as sequences of phone-like units $P_n(w_q) = [p_1, \ldots, p_r, \ldots, p_R]$. Since there are several possible pronunciations of each word, the lexicon also gives a probability for each of the possible pronunciations:

$$w_q \rightarrow P_n(w_q) \;\text{ with probability }\; p(P_n|w_q) \tag{10.155}$$

Normal pronunciations of a word are usually listed in the lexicon. Once a sentence is constructed, however, unusual phonemes in word w_{q-1} or w_{q+1} may cause unusual changes to the pronunciation of word w_q. For example, "did you" is often pronounced as D-IH-JH-AX ("didja"). Most such changes can be accounted for using a small set of phonological rules, which are applied to a sentence after all of the words have been converted into phone transcriptions using an appropriate table lookup.

Finally, each candidate sequence of phones, $P = [P(w_1), \ldots, P(w_Q)]$, is converted into a sequence of HMM states by looking up the phones in another lookup table. This results in a single giant HMM network of states, with three types of state transitions: transitions which remain inside a phone, transitions which cross from one phone to the next inside a word, and transitions from one word to the next. Transitions which remain inside a phone occur with a probability a_{ij} which is specified by the model

$$p(q_t = j|q_{t-1} = i) = a_{ij} \qquad \text{if } i, j \text{ both states in the model for phone } p_r \tag{10.156}$$

Transitions from one phone to the next phone occur with a probability specified by the lexical probability:

$$p(q_t = j|q_{t-1} = i) = p(q_t = j|p_r)p(p_r|p_{r-1}, w_q) \qquad \text{if } i \text{ in } p_{r-1},\ j \text{ in } p_r,\ \text{both in } w_q \tag{10.157}$$

$$= \frac{p([\ldots, p_{r-1}, p_r, \ldots]|w_q)}{p([\ldots, p_{r-1}, \ldots]|w_q)} \times \pi_{j|p_r} \tag{10.158}$$

Finally, transitions from one word to the next word occur with a probability specified by the language model.

Training

Few training databases are transcribed with the beginning and end times of individual phones, so phone models are trained using an "embedded re-estimation" procedure. In "embedded re-estimation," the word sequence in any training token is first parsed to generate a sequence of phones, then the sequence of phones is parsed to generate a sequence of Markov states, and then finally, the Markov states are matched to the utterance, and the state parameters are updated using the Baum-Welch algorithm.

Similarly, initialization of a phone model uses an "embedded" version of the segmental K-means algorithm:

1. Look up the phone transcription of each word in a training utterance.

2. If there are N phone in the phonemic transcription of a sentence, divide the training utterance into N equal-length segments.

3. Estimate model parameters using the K-means algorithm, or using the Baum-Welch re-estimation procedure within the specified segment.

4. Re-segment using the Viterbi algorithm.

5. If the segmentation has changed since the previous iteration, go back to step 3.

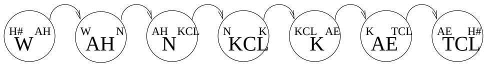


Figure 10.11: A network of triphone models representing the phrase "one cat." Phones are written in the TIMIT transcription system (Zue *et al.* [1990]).

10.3.7 Context-Dependent Recognition Units

In fluent speech, the articulators move smoothly from one phoneme target to another, stopping only briefly (if at all) at each phoneme target. As a result, most of the acoustic signal in continuous speech is composed of phoneme transitions.

Transitions can be modeled explicitly by creating diphone or triphone models instead of phone models. A triphone model $p_L - p + p_R$ is a model of the center phone, p, in the context of a particular phone on the left, p_L, and a particular phone on the right, p_R. For example, the phrase "one cat" might be expanded into the network of phone models shown in figure 10.11 (where the phonemes are written in the ARPABET transcription system, and the symbol /H#/ means silence):

The problem with diphone and triphone models is that there are often not enough data to robustly train a separate HMM for every phone in every context. If there are insufficient training data (at least one new speaker per ten HMM models, according to one study), then recognition accuracy may be very high on the training data, and very low on any test data independent of the training set.

If there is insufficient data to train all of the triphones, triphone models may be combined in one of several ways:

1. Triphones which are not well represented in the training data may be replaced by appropriate left-context or right context diphone models, either $p_L - p$ or $p{+}p_R$. If there is insufficient data to train a diphone model, then use an isolated phone model p.

2. Triphones may be clustered based on acoustic similarity. Initially, all triphones in the database are modeled separately; then triphones are combined, one at a time, in the order which causes the smallest decrease in $P(O|\lambda)$ measured on the training database.

10.3.8 Landmarks, Events, and Islands of Certainty

10.4 Recognition of Utterances

In connected word recognition, we seek to find a single word sequence $W(T)$ which maximizes the probability of the observation vectors O_T:

$$\begin{aligned}
O_T &= [o_1,\ o_2, \ldots,\ o_T], \quad T = \text{number of frames} && (10.159)\\
W(T) = W_Q &= [w_1,\ w_2, \ldots,\ w_Q], \quad Q = \text{number of words in time } T && (10.160)\\
P_A^* &= \max_{Q, W_Q} \left(P(O_T | W_Q) \right) && (10.161)\\
W^* &= \arg\max_{Q, W_Q} \left(P(O_T | W_Q) \right) && (10.162)
\end{aligned}$$

In order to calculate P_A^* and W^* in a practical system, it is useful to define two intermediate probabilities called $\alpha_t(i, v)$ and $P_A(t, v)$:

$$\begin{aligned}
\alpha_t(i|v) &= \max_{W(t-1)} P(O_t,\ q_t = i \mid W(t-1),\ w(t) = v) && (10.163)\\
P_A(t|v) &= \max_{W(t-1)} P(O_t \mid W(t-1),\ w(t) = v) = \sum_{i=1}^{N_v} \alpha_t(i|v) && (10.164)\\
& && (10.165)
\end{aligned}$$

If we assume that there are V different word models, then the optimum word sequence probability P_A^* can be written as

$$P_A^* = \max_{1 \leq v \leq V} P_A(T|v) = \max_{1 \leq v \leq V} \sum_{i=1}^{N_v} \alpha_T(i|v) \tag{10.166}$$

10.4.1 Static Search Graph: Finite State Methods

$\alpha_t(i|\ v)$ can be calculated using the following recursion, which is sort of a combination of the Viterbi and the forward algorithms. Many variations of this algorithm exist, and go by names such as the "one-pass algorithm" and the "frame-synchronous level-building algorithm (FSLB)."

1. **Initialization**

$$\begin{aligned}
\alpha_1(i|v) &= \pi_{iv} b_{iv}(o_1) && (10.167)\\
P_A(1|v) &= \sum_{i=1}^{N_v} \alpha_1(i|v) && (10.168)
\end{aligned}$$

where N_v is the number of states in word model λ_v, π_{iv} is the initial-state probability given $w_1 = v$, and $b_{iv}(o_1)$ is the probability of observing o_1 given $q_t = i$ and $w_t = v$.

2. **Recursion**

   At each time step, the model can either remain in the same word, in which case $w(t) = w_q$, or change to a different word, in which case $w(t) = w_{q+1}$. The accumulated word probability is the maximum of these two choices:

$$P_A(t|v) = \max\left(P_A(t|v = w(t-1)),\ P_A(t|v \neq w(t-1))\right) \tag{10.169}$$

   If the model remains in the same word, then the normal forward algorithm is used:

$$\alpha_t(i|v = w(t-1)) = b_{iv}(o_t) \sum_{j=1}^{N_v} \alpha_{t-1}(j|v) a_{ji} \tag{10.170}$$

   If the model changes words, then we require, by convention, that the model must make a transition from the last state of one word to the first state of the next word. In this case,

$$\alpha_t(i|v \neq w(t-1)) = \begin{cases} b_{1v}(o_t) \max\limits_{1 \le w_q \le V} \alpha_{t-1}(N_{w_q}, w_q) P(v|W_q) & i = 1 \\ 0 & i \neq 1 \end{cases} \tag{10.171}$$

   where $P(v|W_q)$ is the word transition probability, also known as the "language model:"

$$P(v|W_q) \equiv P(w_{q+1} = v|w_1, \ldots,\ w_q) \tag{10.172}$$

   In either case, we have that

$$P_A(t|v) = \sum_{i=1}^{N_v} \alpha_t(i|v) \tag{10.173}$$

   Combining equations 10.169 to 10.173, we obtain the following recursion:

$$P_A(t|v) = \max\left(\sum_{i=1}^{N} b_{iv}(o_t) \sum_{j=1}^{N_v} a_{ji}\alpha_{t-1}(j|v),\quad b_{1v}(o_t) \max_{1 \le w_q \le V} \alpha_{t-1}(N_{w_q}|w_q) P(v|W_q)\right) \tag{10.174}$$

$$\psi_t(v) = \arg\max\left(\sum_{i=1}^{N} b_{iv}(o_t) \sum_{j=1}^{N_v} a_{ji}\alpha_{t-1}(j, v),\quad b_{1v}(o_t) \max_{1 \le w_q \le V} \alpha_{t-1}(N_{w_q}, w_q) P(v|W_q)\right) \tag{10.175}$$

3. **Termination**

$$P_A^* = \max_{1 \le v \le V} P_A(T|v) \tag{10.176}$$

$$w_t^* = \arg\max_{1 \le v \le V} P_A(T|v) \tag{10.177}$$

4. **Backtracking**

$$w_t^* = \psi_{t+1}(w_{t+1}^*) \tag{10.178}$$

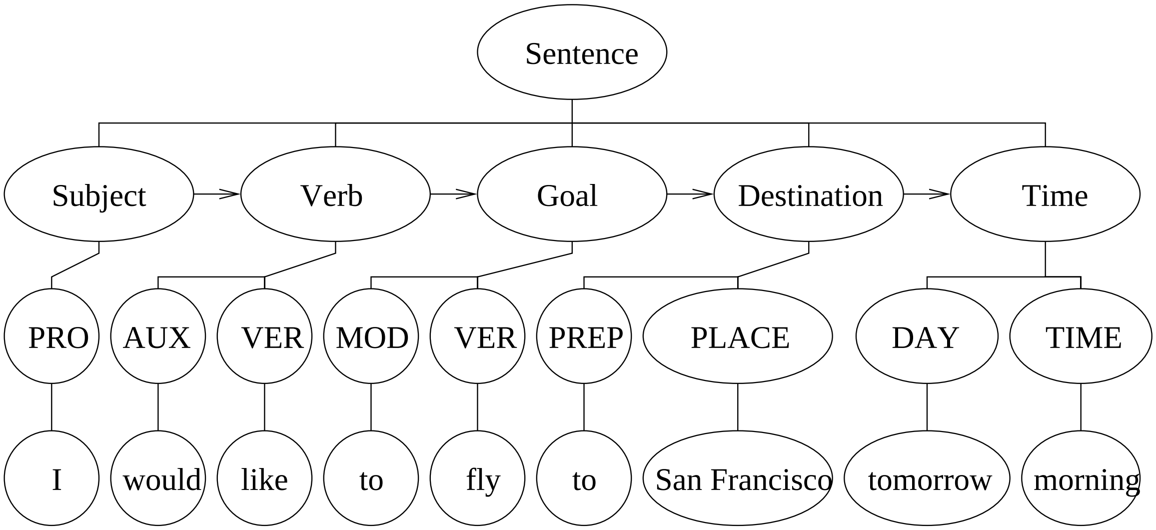


Figure 10.12: A detailed model of word transition probabilities can be created by parsing words into phrases, and phrases into complete sentences.

10.4.2 Regular Grammars for Dialog Systems

It is possible to create an intuitively pleasing language model by grouping words together into phrases. For example, the sentence

$$\text{“I would like to fly to San Francisco tomorrow morning.”} \tag{10.179}$$

Might be parsed as shown in figure 10.4.2.

In figure 10.4.2, the words have been parsed in several levels:

1. **Part of Speech**

   Each word w_q can be parsed as belonging to a particular part of speech c_q (e.g. PRO=pronoun, AUX=auxiliary verb, DAY, TIME, etc.), with the associated probability

   $$P(w_q|c_q) \tag{10.180}$$

2. **Phrase Composition**

   Each sequence of word class tokens, $C_n = [\ c_{n,1}, \ldots,\ c_{n,M}]$, can be parsed as being part of a particular phrase, ϕ_n. For example, in the figure above, the phrases are

   $$\Phi = [\phi_1, \ldots,\ \phi_N] = [\,\text{Subject, Verb Phrase, Goal, Place Phrase, Time Phrase}\,] \tag{10.181}$$

   Each of these phrases is mapped to a word-class sequence C with the associated probability

   $$P(C_n = [c_{n,1}, \ldots,\ c_{n,M}] \mid \phi_n) \tag{10.182}$$

Since the phrase structure of human languages is often recursive (phrases may contain phrases), the computation of $P(C|\phi)$ may also be designed in a recursive fashion; recursion complicates the Viterbi algorithm, but the system designer may decide that the added generality of the language model justifies the added complication. For example, if $C_n = [c_{n,1},\ \phi_{sub}]$, the probability $P(C_n|\phi_n)$ might be computed as:

$$P(C_n \mid \phi_n) = P([c_{n,1},\ \phi_{sub}] \mid \phi_n) P(C_{sub} \mid \phi_{sub}) \tag{10.183}$$

3. **Sentence Composition**

   The entire phrase sequence composes a "sentence" $\Phi = [\phi_1,\ \phi_2, \ldots,\ \phi_N]$ (since system users do not always speak in grammatical sentences, Φ may or may not be a complete sentence). If the probability of a particular phrase sequence is $P(\Phi)$, then the probability of the word sequence W is

$$P(W) = P(\Phi) \prod_{n=1}^{N} P(C_n|\phi_n) \prod_{m=1}^{M} P(w_{n,m}|c_{n,m}) \tag{10.184}$$

Training

The language model includes three probabilities: $P(\Phi)$, $P(C|\Phi)$, and $P(w_q|c_q)$. All three probabilities can trained from data using a version of the segmental K-means algorithm:

1. The system designer must make initial estimates of the probabilities $P(\Phi)$, $P(C|\phi)$, and $P(w_q|c_q)$:
   - Since there is a potentially infinite set of phrase sequences Φ and word-class sequences C, the designer must specify the possible sequences at each of these two levels, and assign approximate probabilities for each possible sequence.
   - The word-class probabilities $P(w_q|c_q)$ are difficult to control during training, so these probabilities should be set conservatively. For any given word w_q, $P(w_q|c_q)$ should be non-zero for the minimum possible number of word classes (often only one).
2. Based on the initial probabilities, the Viterbi algorithm is used to find the most likely phrase-level and class-level transcriptions, Φ^* and C^*, of each sentence.
3. The model probabilities are re-estimated based on frequency of occurrence.
4. If there has been a change in any of the model probabilities, repeat from step 2.

5. The final phrase-level transcriptions should be checked by hand, to identify mistakes. Mistakes in the final transcription often result from the speaker's use of a grammatical structure which the designer didn't consider possible. Such problems are usually easy to spot, because they will usually prevent the Viterbi algorithm from finding a path with non-zero probability.

10.4.3 N-Grams and Backoff

The most common language model is a model in which the probability of each word depends only on the $N-1$ preceding words:

$$P(w_q|W_{q-1}) = P(w_q|w_{q-N+1}, \ldots, w_{q-1}) \tag{10.185}$$

The N-gram probabilities are typically stored in a lookup table. If N is too large, the lookup table becomes impossible to use in a practical situation. Most coders therefore use either a bigram ($N = 2$) or trigram ($N = 3$) language model.

The N-gram model may be trained from either speech data or, if you have a text database which reflects the kinds of utterances you expect people to say, it may be trained from text. Training involves counting $N(w_1, w_2, w_3)$, the number of times that word w_3 follows words w_1 and w_2:

$$\bar{P}(w_3|w_1, w_2) = \frac{N(w_1, w_2, w_3)}{N(w_1, w_2)} \tag{10.186}$$

If the number of possible words is large, many valid trigram combinations will be rare or nonexistant in the training data. If the language model is trained using 10.186, trigrams which do not exist in the training data will be marked as impossible, and will never be recognized correctly if they occur in the test data. In order to avoid this problem, the trigram probabilities may be estimated by interpolating the relevant trigram, bigram, and unigram frequencies, as follows:

$$\bar{P}(w_3|w_1, w_2) = p_1 \frac{N(w_1, w_2, w_3)}{N(w_1, w_2)} + p_2 \frac{N(w_2, w_3)}{N(w_2)} + p_3 \frac{N(w_3)}{\sum_{w_3} N(w_3)} \tag{10.187}$$

The interpolation probabilities may vary depending on the word frequencies $N(w_1, w_2, w_3)$, but they should always be normalized so that

$$p_1 + p_2 + p_3 = 1 \tag{10.188}$$

Perplexity

A speech recognizer with a vocabulary of V words does not need to consider V different possibilities for every new word w_q. Intuitively, the recognizer only needs to consider words for which the language model probability $P(w_q|W_{q-1})$ is large. This intuition can be formalized by defining the entropy of the language model:

$$H_Q = -E[\log_2 P(w_Q|W_{Q-1})] = -\sum_{w_Q=1}^{V} P(w_Q|W_{Q-1}) \log_2 P(w_Q|W_{Q-1}) \qquad (10.189)$$

where Q is chosen so that, for the particular language model in question,

$$H_Q = \lim_{q\to\infty} H_q \qquad (10.190)$$

The entropy H_Q can be interpreted as a measure of the difficulty encountered by the recognizer in trying to identify each new word. In particular, a recognition task with entropy H_Q can be said to be as difficult as a task in which words are chosen randomly from a vocabulary of B words, where

$$B = 2^{H_Q} \qquad (10.191)$$

B is called the “perplexity” or “branching factor” of the language model. It is possible to build quite accurate speech recognizers with very large vocabularies if the task is constrained in such a way that the branching factor B is low.

Class N-Gram

The N-gram model is the most direct language model possible, but unless given infinite training data, it will often miss important language constraints and possibilities. For example, if the training data contains the sentences “cats drink milk” and “dogs swim in water,” one might hope that the language model will assign a non-zero probability to the word string “dogs drink water,” even if that sentence is not found in the training data.

This kind of generalization is possible if every word w_j is assigned to a word class $C(w_j)$ before training. The language model probabilities are then

$$\bar{P}(w_3|C(w_1), C(w_2)) = \frac{N(C(w_1), C(w_2), C(w_3))}{N(C(w_1), C(w_2))} \qquad (10.192)$$

The classes might be syntactic (“noun” and “verb”), or, for a more precise model, they can combine syntactic and semantic information (“animal,” “drinkable liquid”). If the classes are designed to fit the application, language models built in this way can be extremely accurate.

10.4.4 Dynamic Search Graph: Stack-Based Methods

10.4.5 Dynamic Search Graph: Bayesian Networks

10.4.6 Multi-Pass Recognition

10.4.7 System Combination

10.5 Automatic Recognition and Verification of Speakers

The previous discussion pointed up the notion that the spectral patterns of one speaker are not always adequate to recognize the speech of another. This fact suggest that spectral data might be used to recognize or identify different speakers. A number of efforts along these lines have been made–mainly with the use of digital computers. By way of illustration, one study produced quantized time-frequency-intensity (spectrographic) patterns from a 17-channel filter bank scanned at a rate of 100Hz (Pruzansky [1963]). Ten key words were excerpted from context for 10 different speakers (three women, seven men). For each talker, three utterances of the 10 key words were used to establish the reference patterns for that individual.

For talker identification, the spectrographic pattern of a different key-word utterance by an unknown speaker of the ten-member group was cross-correlated with the reference patterns (again by multiplying amplitudes at each time-frequency element of the spectrogram), and the maximum correlation was taken. Because the utterances varied in length, alignment of patterns was done by matching them at the maximum overall amplitude points. Results showed that among the 10 speakers for whom the reference library was formed, the identification was correct in 89% of the cases.

In the same study, the three dimensional time-frequency-intensity patterns were reduced to two dimensions by summing over the time of the utterance for each filter channel. The summation produces a graph of integrated intensity-versus-frequency for each utterance. It was found that this operation still afforded a recognition score of 89%.

It is of course difficult to draw conclusions about human recognition of speakers from such an experiment. Again, however, for a limited, specific application, where speaker ensemble and vocabulary are restricted, such a technique could be effectively applied.

A few experiments have measured human recognition of speakers from visual inspection of speech spectrograms. In one of these (Kersta [1948, 1962a]) a group of speakers (either 5, 9 or 12) was asked to utter 10 key words four times. Conventional bar spectrograms and contour spectrograms were made of their utterances (see Section 5.1.4). For each word a randomized matrix of spectrograms consisting of four utterances of each speaker was displayed. Subjects were asked to identify the utterances of each individual speaker. The errors in grouping the prints according to speaker ranged from 0.35% to 1.0% for bar prints and from 0.37% to 1.5% for contour spectrograms. When the test words were excerpted from context, the error was still about the same order of magnitude.

A second experiment was modeled after fingerprint identification procedures, although the analogy is a tenous one. A file of "voice prints" of five key words was compiled for 12 speakers. Subjects then identified a different set of utterances by an unknown member of the group through comparisons to the reference sets. Using the groups of five cue words, the misidentifications were less than 1%. Identifications based upon two 5-word groups in tandem gave errors of about one-half percent. Preliminary investigations were also made into the ability to recognize disguised voices. The results suggest that adults have certain invariant linguistic and physiological characteristics which the spectrograph may display even when an effort is made to alter the voice.

These experiments, through a combination of publicity and private development, captured the notice of various law-enforcing organizations, who saw in the method a new means for identifying criminals. Several efforts were made to introduce the technique into legal proceedings with controversial results. Independent experiments were conducted to test the method, and the findings were at variance with the original experiments (Young and Campbell [1967]). Most opinion holds that more research is needed to accurately establish the utility of human recognition of speakers from sound spectrograms (Bolt [1970]). Subsequent efforts continue in this direction (TOSI). These latter experiments have treated a variety of experimental conditions (for example, closed sets versus open sets) and the error rates in visual identification vary from 1% to 30%, depending upon the experimental constraints. This error range, when analyzed in terms of experimental conditions, appears consistent with previous data.

A problem perhaps more interesting and presently more tractable than speaker recognition is automatic verification of speakers (Doddington [1971], Lummis [1971], Das and Mohn [1969]). In the usual context of this problem one has a restricted population of "customers" who want to be verified (i.e., a cooperative situation), and they are willing to state a prearranged phrase (secret if desired) chosen to be advantageous for the machine. (The voice banking, and voice validation of credit cards are applications in point). In the verification situation unknown caller, x, claims to be customer, C_i. The machine must decide to accept or reject x as C_i. The decision can be weighted according to the importance of the verification (for example, whether the sum charged is large or small) and a predetermined mix of error types (i.e., rejecting a true speaker versus accepting a false speaker) can be specified.

The most important aspect of the verification problem, and the one which distinguishes it from the recognition problem, is that no matter what the size of the impostor population, the average percent correct verification tends to be constant. The performance is determined by the average consistencies of the known speakers and by how each of them differs from the average of the impostor population. In a recognition situation, on the other hand, where the unknown must be identified by successive comparisons to all members of a known set, the probability of error is monotonely related to the number of speakers in the set, and the probability of a recognition error approaches unity as the user population becomes large.

One experiment on verification (Doddington [1971]) has made use of pitch, formant and intensity data to form reference patterns for the known speakers. Frequency data (i.e., formants and pitch) were considered attractive because they are resistant to variations in the amplitude-frequency characteristics of a voice communication link. A novel non-linear time-warping of the utterance of an unknown speaker was used to compare (register) it with a stored reference pattern corresponding to the claimed identity. The non-linear warp was achieved on digital computer by a steepest-ascent algorithm. The algorithm warped the pattern of the unknown speaker to maximize its correlation with the stored reference pattern. A mean square error measure was then made for the registered patterns and the speaker was accepted or rejected depending upon whether the mean square error was less than greater than a threshold chosen for a specified mix of errors (i.e., reject true versus accept false).

Fig. 10.13 shows how the formant, pitch and intensity (gain) data are compared for a verification phrase; namely, the voiced sentence “We were away a year ago.” In Fig. 10.13a the unknown utterance (solid curve) has been given a linear time stretch to make its duration equal to the reference (dashed curve). Poor internal registration is evident. In Fig. 10.13b, the non-linear warp has been applied to register the second formant tracks with maximum correlation. The registration of the other parameters is similarly improved. The remaining differenc and the amount of non-linear warp applied are indicative of the similaritie of the two patterns. A square error measure is formulated to indicat a “distance” between the registered patterns.

Using this technique, with a population of 40 male speakers, correct verification was achieved 98.5% of the time on the verification phrase “We were away a year ago” used by all subjects. Identical twins included in the experiment were differentiated 100% of the time.

If more sophisticated “distance measures” are used to characterize the differences between the registered patterns for the unknown and reference, a comparable performance can be obtained on simple measures, easily made in real time. A subsequent experiment on the population of 40 speakers, and using more elaborate distance measures on only intensity, pitch and non-linear warp, achieved 99% correct verification (Lummis [1971]).

A natural query is “How well would human listeners perform in the same task?” To answer this, a completely parallel auditory experiment was conducted with the same 40 speakers, but using human listeners instead of a computer to make the verification decision. The listeners performed with greater error rate than the machine and achieved approximately 96% correct verification (Rosenberg [1971a]).

Results of these and related verification experiments suggest that automatic ma-

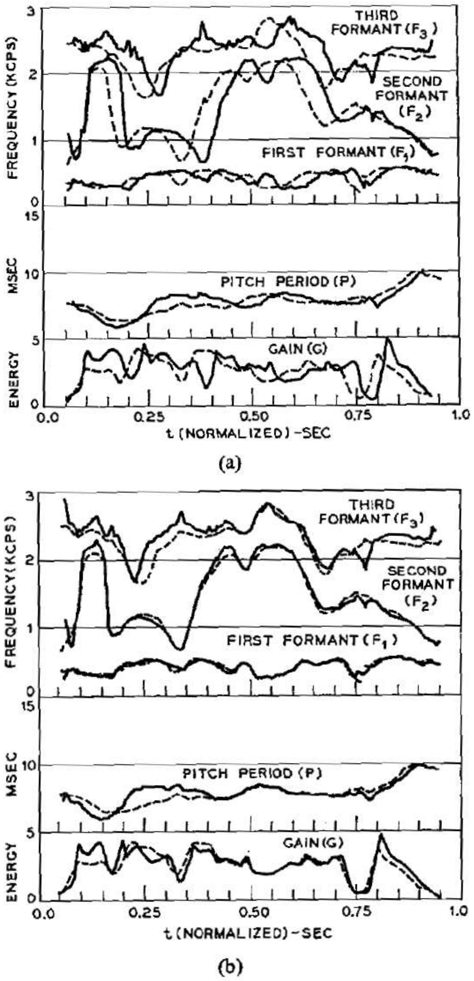


Figure 10.13: Effects of nonlinear warp in registering speech parameter patterns. The dashed curves are reference data for an individual. The solid curves are a sample utterance from the same individual. (a) Linear stretch to align end points only. (b) Nonlinear warp to maximize the correlation of the F2 patterns. (After (Doddington [1971]))

chine verification may have practical value. An obvious and further question is how easily might accomplished mimics deceive the machine and be erroneously accepted. Continuing research is aimed at this question.

A number of features seem to distinguish one speaker from another. The size and shape of the vocal tract vary considerably among persons. Characteristic damping, mouth and glottal dimensions also vary. Individual nasal coupling, size and damping of the nasal tract are other relevant features. Temporal patterns of intensity (stress) and pitch (inflection) are still others. Vocal obstructions and variations in dental work may contribute still further differences. Some or all these factors might be used to recognize or verify a speaker. It is probable that machine and human do not use the same features to equal effect. The machine, for example, might make use of data the human ear cannot assimilate.

As suggested earlier, the speech-recognition and speaker-identification experiments described here tell us little about the perceptual processing which the human accomplishes. They do not, for example, suggest the temporal span of the recognition unit used by the human. Neither do they indicate subjective techniques for measuring whether the unit is the phoneme, word, sentence, or something larger. The automatic machine methods deal mainly with advantageous processings of essentially the acoustic signal, and not with perception as the human practices it.

The mechanism of human perception of speech is difficult to analyze and present understanding is meager. The discussion of Chapter 7 showed that for signals with simple temporal and spectral structure, reasonably close correlations can be made between subjective behavior and the known physiology of the peripheral ear. To a modest extent, similar relations can be established for speech signals. (For example, one can identify features such as voice pitch, formant frequency and voiced-unvoiced excitation in terms of the basilar membrane motion.) But how the neural data are stored and processed alter leaving the periphery is a completely open question. Continued research on the electrophysiology of the auditory tract, and on human response to meaningful speech signals, will hopefully provide some of the answers.

10.6 Homework

Problem 10.1

Days are either Good (G) or So-so (S). The probability that today is a good day depends only on whether or not yesterday was a good day:

$$P(q_t = G|\ q_{t-1} = G) = 3/4, \qquad P(q_t = G|\ q_{t-1} = S) = 1/4 \tag{10.193}$$

Unfortunately, you have no way of directly measuring whether a given day is Good or So-so. You have noticed, however, that on Good days, the cafeteria is more likely to serve your favorite lunch (filet mignon with fresh asparagus, truffles, and slivered almonds):

$$P(o_t = \text{filet} \mid q_t = G) = 3/4, \qquad P(o_t = \text{filet} \mid q_t = S) = 1/4 \tag{10.194}$$

You have also noticed that the first day of a new quarter is always a good day:

$$P(q_1 = G) = 1 \tag{10.195}$$

Given this model, what is the probability that the cafeteria will serve your favorite lunch for the first two days of a semester?

Problem 10.2

Write a program that uses a Gaussian model of each spectral frame in order to classify a waveform. Use a perceptually-motivated cepstral feature vector, such as PLP (Hermansky [1990]) or MFCC (Davis and Mermelstein [1980]).

Record examples of your own voice saying the words "yes" and "no," twenty times each.

Create a simple push-to-talk interface that allows you to press a button to start recording, then say a word. Create a "voice activity detector" that examines the energy of the signal in each 10ms frame, and chops off leading and trailing silences.

After chopping off leading and trailing silences, your user interface should pass waveforms to the Gaussian classifier. The mixture Gaussian classifier should compute the values of $p(\text{“yes”}|x_1, \ldots, x_T)$ and $p(\text{“no”}|x_1, \ldots, x_T)$. Assume that the words "yes" and "no" have equal *a priori* probability, and that the observation vectors are independent given class label, i.e.,

$$p(x_1, \ldots, x_T|\text{“yes”}) = \prod_{t=1}^{T} p(x_t|\text{“yes”}) \tag{10.196}$$

How accurate is your recognizer? Find an example of a waveform that was mis-classified by the recognizer, and plot the cepstral distances $d_2(\text{test}, \text{“yes”})$ and

d_2(test, “no”) as a function of time. Line up these distance measures with the spectrogram. Can you figure out why the recognizer made a mistake?

Problem 10.3

Write a program that uses dynamic time warping to compute the cepstral distance between two waveforms. Use a perceptually-motivated cepstral feature vector, such as PLP (Hermansky [1990]) or MFCC (Davis and Mermelstein [1980]).

Record examples of your own voice saying the phrases “supercalifragilistic,” “soup magic,” and “supertragic.”

Experiment with different constraints on the warping costs. For very low warping costs, can you convince the recognizer that “supertragic” is closer to “supercalifragilistic” than it is to “soup magic?”

Chapter 11

Speech Synthesis

Ancient man often took his ability of speech as a symbol of divine origin. Not unnaturally, he sometimes ascribed the same ability to his gods. Pagan priests, eager to fulfill great expectations, frequently tried to make their idols speak directly to the people. Talking statues, miraculous voices and oracles were well known in the Greek and Roman civilizations–the voice usually coming to the artificial mouth via cleverly concealed speaking tubes. Throughout early times the capacity of "artificial speech" to amaze, amuse and influence its listeners was remarkably well appreciated and exploited.

As the civilized world entered the Renaissance scientific curiosity developed and expanded. Man began to inquire more seriously into the nature of things. Human life and physiological functions werefair targets of study, and the physiological mechanism of speech belonged in this sphere. Not surprisingly, the relatively complex vocal mechanism was often considered in terms of more tractable models. These early models were invariably mechanical contrivances, and some were exceedingly clever in design.

11.1 Mechanical Speaking Machines

One of the earliest documented efforts at speech synthesis was by Kratzenstein in 1779. The Imperial Academy of St. Petersburg offered its annual prize for explaining the physiological differences between five vowels, and for making apparatus to produce them artificially. As the winning solution, Kratzenstein constructed acoustic resonators similar in shape to the human vocal tract. He activated the resonators with vibrating reeds which, in a manner analogous to the human vocal cords, interrupted an air stream.

A few years later (1791), von Kempelen constructed and demonstrated a more elaborate machine for generating connected utterances (Apparently von Kempelen's efforts antedate Kratzenstein's, since von Kempelen purportedly began work on his device in 1769 (Kempelen [1791], Tarnoczy [1950])). Although his machine received

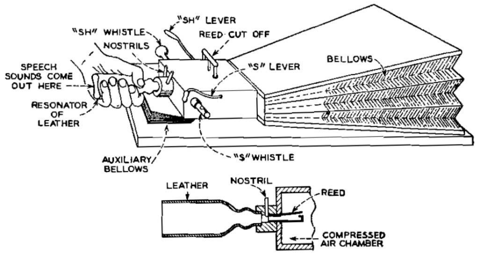


Figure 11.1: Wheatstone's construction of von Kempelen's speaking machine

considerable publicity, it was not taken as seriously as it should have been. Von Kempelen had earlier perpetrated a deception in the form of a mechanical chess-playing machine. The main "mechanism" of the machine was a concealed, legless man–an expert chess player.

The speaking machine, however, was a completely legitimate device. It used a bellows to supply air to a reed which, in turn, excited a single, hand-varied resonator for producing voiced sounds. Consonants, including nasals, were simulated by four separate constricted passages, controlled by the fingers of the other hand. An improved version of the machine was built from von Kempelen's description by Sir Charles Wheatstone (of the Wheatstone Bridge, and who is credited in Britain with the invention of the telegraph). It is shown in Fig. 11.1.

Briefly, the device was operated in the following manner. The right arm rested on the main bellows and expelled air through a vibrating reed to produce voiced sounds. (See the lower diagram in Fig. 11.1.) The fingers of the right hand controlled the air passages for the fricatives /ʃ/ and /s/, as well as the "nostril" openings and the reed on-off control. For vowel sounds, all the passages were closed and the reed turned on. Control of vowel resonances was effected with the left hand by suitably deforming the leather resonator at the front of the device. Unvoiced sounds were produced with the reed off, and by a turbulent flow through a suitable passage. In the original work, von Kempelen claimed that approximately 19 consonant sounds could be made passably well.

Von Kempelen's efforts probably had a more far-reaching influence than is generally appreciated. During Alexander Graham Bell's boyhood in Edinburgh, Scotland (latter 1800's), Bell had an opportunity to see the reproduction of von Kempelen's machine which had been constructed by Wheatstone. He was greatly impressed with the device. With stimulation from his father (Alexander Melville Bell, an elocutionist like his own father), and his brother Melville's assistance, Bell set out to construct a speaking automaton of his own.

Following their father's advice, the boys attempted to copy the vocal organs by making a cast from a human skull and molding the vocal parts in gutta-percha. The lips, tongue, palate, teeth, pharynx, and velum were represented. The lips were a framework of wire, covered with rubber which had been stuffed with cotton batting.

Rubber checks enclosed the mouth cavity, and the tongue was simulated by wooden sections–likewise covered by a rubber skin and stuffed with batting. The parts were actuated by levers controlled from a keyboard. A larynx "box" was constructed of tin and had a flexible tube for a windpipe. A vocal cord orifice was made by stretching a slotted rubber sheet over tin supports.

Bell says the device could be made to say vowels and nasals and could be manipulated to produce a few simple utterances (apparently well enough to attract the neighbors). It is tempting to speculate how this boyhood interest may have been decisive in leading to U.S. patent No. 174,465, dated February 14, 1876–describing the telephone, and which has been perhaps one of the most valuable patents in history.

Bell's youthful interest in speech production also led him to experiment with his pet Skye terrier. He taught the dog to sit up on his hind legs and growl continuously. At the same time, Bell manipulated the dog's vocal tract by hand. The dog's repertoire of sounds finally consisted of the vowels /ɑ/ and /u/, the diphthong /oʊ/ and the syllables /mɑ/ and /gɑ/. His greatest linguistic accomplishment consisted of the sentence, "How are you Grandmamma?" The dog apparently started taking a "bread and butter" interest in the project and would try to talk by himself. But on his own, he could never do better than the usual growl. This, according to Bell, is the only foundation to the rumor that he once taught a dog to speak.

Interest in mechanical analogs of the vocal system continued to the twentieth century. Among those who developed a penetrating understanding of the nature of human speech was Sir Richard Paget. Besides making accurate plaster tube models of the vocal tract, he was also adept at simulating vocal configurations with his hands. He could literally "talk with his hands" by cupping them and exciting the cavities either with a reed, or with the lips made to vibrate after the fashion of blowing a trumpet.

Around the same time, a different approach to artificial speech was taken by people like Helmholtz, D. C. Miller, Stumpf, and Koenig. Their view was more from the point of perception than from production. Helmholtz synthesized vowel sounds by causing a sufficient number of tuning forks to vibrate at selected frequencies and with prescribed amplitudes. Miller and Stumpf, on the other hand, accomplished the same thing by sounding organ pipes. Still different, Koenig synthesized vowel spectra from a siren in which air jets were directed at rotating, toothed wheels.

In 1937, Riesz (Riesz and Watkins [1939]) developed the mechanical talker shown in Fig. 11.2. Air under pressure is brought from a reservoir at the right. Two valves, V_1 and V_2 control the flow. Valve V_1 admits air to a chamber L_1 in which a reed is fixed. The reed vibrates and interrupts the air flow much like the vocal cords. A spring-loaded slider varies the effective length of the reed and changes its fundamental frequency. Unvoiced sounds are produced by admitting air through valve V_2. The configuration of the vocal tract is varied by means of nine movable members representing the lips (1 and 2), teeth (3 and 4), tongue (5, 6, and 7), pharynx (8), and velar coupling (9).

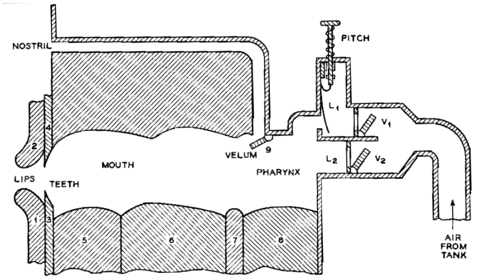


Figure 11.2: Mechanical vocal tract of Riesz

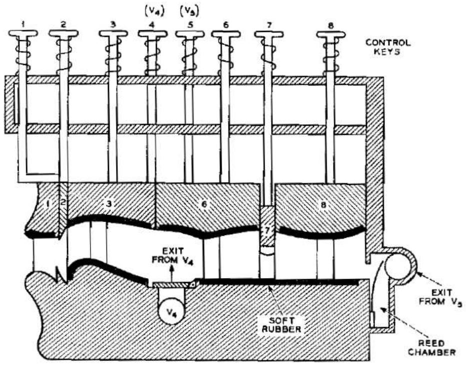


Figure 11.3: Key control of Riesz's mechanical talker

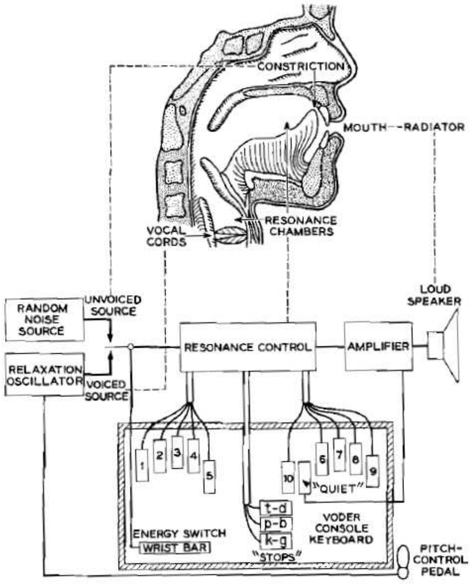


Figure 11.4: Schematic diagram of the Voder synthesizer (After (Riesz and Watkins [1939])

To simplify the control, Riesz constructed the mechanical talker with finger keys to control the configuration, but with only one control each for lips and teeth (i.e., members 1-2 and 3-4 of Fig. 11.2 worked as opposing pairs). The simplified arrangement with control keys is shown in Fig. 11.3. The dark surface regions indicate soft rubber linings to accomplish realistic closures and dampings. Keys 4 and 5 operate excitation valves V_4 and V_5, arranged somewhat differently from V_1 and V_2 in Fig. 11.2. Valve V_4 admits air through a hole forward in the tract (below element 6) for producing unvoiced sounds. Valve V_5 supplies air to the reed chamber for voiced excitation. In this case pitch is controlled by the amount of air passed by valve V_5. When operated by a skilled person, the machine could be made to simulate connected speech. One of its particularly good utterances was reported to be "cigarette"[1].

Probably the first electrical synthesizer which attempted to produce connected speech was the Voder (Riesz and Watkins [1939]). It was basically a spectrum-synthesis device operated from a finger keyboard. It did, however, duplicate one important physiological characteristic of the vocal system, namely, that the excitation can be voiced or unvoiced. A schematic diagram of the device is shown in Fig. 11.4.

The "resonance control" box of the device contains 10 contiguous band-pass filters which span the speech frequency range and are connected in parallel. All the filters receive excitation from either the noise source or the buzz (relaxation) oscillator. The wrist bar selects the excitation source, and a foot pedal controls the pitch of the buzz oscillator. The outputs of the band-pass filters pass through potentiometer

---

[1] Personal communication, R. R. Riesz.

gain controls and are added. Ten finger keys operate the potentiometers. Three additional keys provide a transient excitation of selected filters to simulate stop-consonant sounds.

This speaking machine was demonstrated by trained operators at the World's Fairs of 1939 (New York) and 1940 (San Francisco). Although the training required was quite long (on the order of a year or more), the operators were able to "play" the machines–literally as though they were organs or pianos–and to produce intelligible speech.[2].

11.2 Unit Selection Synthesis

The sanatorium in Heliopolis, during the time of the Roman Empire, included an underground chamber where worshippers could sleep under the protection of the gods. It was said, at the time, that worshippers would be visited there by gods during the night, and that the gods would whisper, into their ears, instructions that they must follow in order to improve their health. In fact, however, the "voices of the gods" were supplied by over-zealous priests and acolytes, who spoke to sleeping clients through a network of cleverly concealed speaking tubes (THIS STORY IS MARK'S MEMORY – NEEDS TO BE VERIFIED).

In some ways, modern telephone dialog systems rely on a technology remarkably similar to that of Roman Heliopolis: the only way we are able to generate completely natural speech is by bringing an actor to a recording studio, and asking him or her to record every utterance that the dialog system is expected to produce. Storage is cheap. The cost of storing ten or twenty hours of recorded speech is inconsequential—far less, in fact, than the cost of hiring the actor to sit in a recording studio for twenty hours, reading prompt sentences from cue cards.

Pre-recorded utterances are fine for a dialog system with a limited repertoire of utterances, but one would imagine that this technology is useless for a system that requires infinite flexibility: a system that requires, for example, the ability to read books or newspapers. The most economically successful speech synthesis technology of the twentieth-century was a set of methods that convert a corpus of pre-recorded speech waveforms into an infinitely flexible speech synthesizer. The basic idea of a "corpus-based synthesizer" is that utterances should be generated by cutting and pasting waveform segments from a large database of recorded speech. Waveform units are chosen to be as long as possible—entire sentences, if possible. The corpus itself is as large as possible—typically ten to twenty hours—and is designed to provide complete coverage of the vocabulary and common utterances of the intended dialog

---

[2] H. W. Dudley retired from Bell Laboratories in October 1961. On the completion of his more than 40 years in speech research, one of the Voder machines was retrieved from storage and refurbished. In addition, one of the original operators was invited to return and perform for the occasion. Amazingly, after an interlude of twenty years, the lady was able to sit down to the console and make the machine speak.

system. Automatic speech recognition is applied in two ways: first, to label phone boundaries and word boundaries in the recorded corpus, and second, to select the sequence of waveform units that will synthesize a desired utterance with the fewest possible cut points.

11.2.1 Search Algorithms for Unit Selection

11.2.2 Unit Selection Criteria for Affective and Expressive Speech

11.2.3 Text Analysis

11.3 Spectrum Reconstruction Techniques

Section 11.2 discussed unit-selection synthesis: a set of methods capable of generating arbitrary utterances by cutting and pasting units from a large recorded corpus, with no additional signal processing. Corpus-based synthesis is arguably the most financially successful speech synthesis algorithm ever, but it has important limitations. First, a corpus-based synthesizer can only simulate the voice of the person who recorded the speech corpus. It is impossible, using corpus-based synthesis, to synthesize the voices of all characters in an animated film; likewise, it is impossible to create a personalized speech synthesizer using the voice of a person who has just undergone tracheotomy. Second, notwithstanding the methods described in Sec. 11.2.2, it is very difficult to synthesize expressive speech using unit selection. A skilled actor expresses emotion through a wide range of voice qualities and intonational contours; there has never yet been an actor who was willing to sit in a recording studio long enough to record every word in the English language with every possible combination of intonational contours and voice quality characteristics. Third, a corpus-based synthesizer is only useful on a computer with enough disk space to store a large corpus of recorded speech. Hand-held devices typically require a speech synthesizer with much lower memory requirements. All of these problems can be solved by applying the signal processing techniques from Chapter 5 in order to modify recorded speech waveforms. Signal modifications usually reduce the perceived naturalness of the synthesized speech; most current commercial research in the field of speech synthesis aims to develop signal processing techniques that generate natural-sounding speech.

11.3.1 Short-Time Spectral Reconstruction Techniques

Investigators such as Helmholtz, D. C. Miller, R. Koenig and Stumpf had earlier noted that speech-like sounds could be generated by producing an harmonic spectrum with the correct fundamental frequency and relative amplitudes. In other words, the signal could be synthesized with no compelling effort at duplicating the vocal system, hut mainly with the objective of producing the desired percept. Among the first to demonstrate the principle electrically was Stewart, who excited two coupled resonant electrical circuits by a current interrupted at a rate analogous to the voice fundamental. By adjusting the circuit tuning, sustained vowels could be simulated. The apparatus was not elaborate enough to produce connected utterances. Somewhat later, Wagner devised a similar set of four electrical resonators, connected in parallel, and excited by a buzz-like source. The outputs of the four resonators were combined in the proper amplitudes to produce vowel spectra.

Speech analysis by the sound spectrograph was described at some length in Chapter 5. Since–as Helmholtz and others observed–intelligibility is largely preserved in the short-time amplitude spectrum, speech synthesis from spectrographic plots is immediately suggested. Coupled with this notion is the question of the extent to which spectrograms of real speech might be abstracted or "caricatured" without destroying intelligibility. Several devices for automatically "playing" sound spectrograms have been designed. One uses a line source of light, parallel to the frequency axis of the spectrogram, to illuminate a variable density spectrographic pattern (Riesz and Schott [1946], Schott [1948]). Contiguous photocells behind the pattern develop amplitude control signals for a set of band-pass filters (such as in the Voder). Voiced-unvoiced selection and pitch control information are represented in additional tracks. A similar scheme has been used to control a Voder-type synthesizer in an arrangement called Voback (Borst [1956]).

A somewhat different type of spectrogram playback has been used in extensive studies on speech synthesis (Cooper [1950], Liberman and Borst [1951]). The speech wave is effectively simulated by a Fourier series $\sum_n A_n \cos(n\omega_0 t + \Phi_n)$. The coefficients A_n are time varying and are determined by the spectrogram intensity at a given instant. The sound generation is accomplished by the arrangement shown in Fig. 11.5a.

The regular time-frequency-intensity pattern is illuminated by 50 contiguous light spots. The spots are sinusoidally modulated in intensity at harmonically related frequencies. The contiguous spots are produced by illuminating a "tone-wheel" with a line source. The tone wheel has 50 concentric, variable-density bands. The innermost band has four sinusoidal cycles, the next 8, the next 12, and on up to 200 for the 50th band. The tone wheel is rotated at 1800 rpm so the fundamental frequency is 120Hz. Light from the tone wheel can be either reflected from the spectrographic pattern or transmitted by it. The reflected (or transmitted) light is sensed by a collector and photocell which effectively sums the fifty terms of the Fourier series. The collected components are amplified and transduced.

Because of the constant rotation of the tone wheel, the pitch is monotone. Also, the phase relations of the harmonic components are fixed by the tone-wheel bands. Unvoiced sounds are simulated from a random time and intensity modulation of the frequency componentssimilar to the spectrographic representation of a noise burst. Spectrograms of both real speech and its abstracted version can be played on the

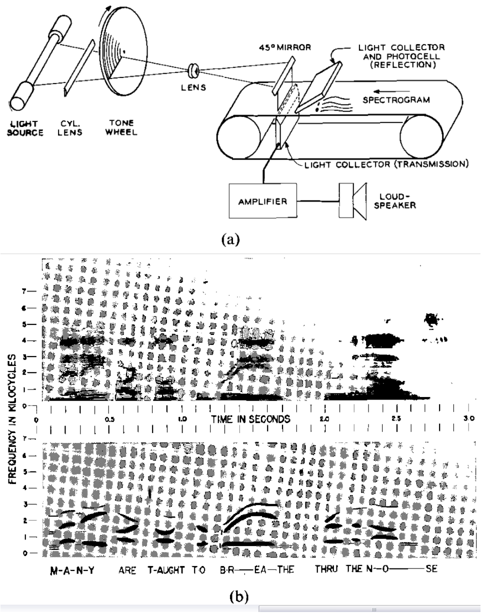


Figure 11.5: (a) Functional diagram of a spectrogram play-back device. (After (Cooper [1950])) (b) Spectrograms of real speech and an abstracted, hand-painted version of the same. Both displays can be synthesized on the pattern play-back machine. (After (Borst [1956]))

machine. A sample of each is shown in Fig. 11.5b. In the abstracted spectrogram, in the lower part of the figure, the dark bars represent the speech formants, and the patches of fine, irregular dots produce the noise bursts. Intelligible monotone speech can be produced by the machine, and it has been used in extensive perceptual studies. Some of these results will discussed in Chapter 9.

11.3.2 Unit-Concatenative Synthesis for Embedded Applications

11.3.3 Signal Modification for Affective and Expressive Speech

11.3.4 Talker Morphing

11.4 "Terminal Analog" Synthesizers

In Chapter 4 linear circuit theory was applied to the acoustic analysis of the vocal tract. The results show that for simple geometries the transmission properties can be stated in a straightfoward form. Complex geometries, on the other hand, may be approximated by quantizing the vocal tube as short, abutting cylindrical sections. Effects of losses and yielding walls can be included as discussed in Section 4.7.3.

The tract behavior can be considered either in terms of its over-all transmission, or in terms of its detailed distributed properties. Speech synthesis may be based upon either view. The former approach attempts to duplicate–usually with a unilateral electrical circuit–the transmission properties of the tract as viewed from its input and output terminals. Synthesizers designed in this manner have, for lack of a better term, been named "terminal-analogs" (Flanagan [1957b]). The second view attempts to duplicate, on a one-for-one basis, the geometry and distributed properties of the tract. Electrical synthesizers designed according to this approach are bilateral, nonuniform transmission-line models of the system. The present section proposes to discuss the terminal analog approach, while the following section will treat the transmission-line device.

Both approaches to synthesis must take account of sound radiation and the vocal sources of excitation. These factors, common to both modellings of speech production, will be discussed subsequently.

11.4.1 Terminal Properties of the Vocal Tract

The unconstricted, glottally-excited tract can be approximated as a straight pipe, closed at the vocal cords ($Z_g = \infty$) and open at the mouth ($Z_r = 0$). For such a case the results of Chapter 4 show that the ratio of mouth and glottal volume velocities has a frequency-domain representation

$$\frac{U_m}{U_g} = \frac{1}{\cosh \gamma l}, \tag{11.1}$$

where l is the length of the tube, $\gamma = (\alpha + j\beta) = [(R_a + j\omega L_a)(G_a + j\omega C_a)]^{\frac{1}{2}}$, and R_a, L_a, G_a and C_a are the per-unit-length acoustical parameters of the pipe (see Fig. 4.19 and Eq. (4.28)]. It will be convenient in the subsequent discussion to treat frequency as a complex variable. Let $j\omega \to s = \sigma + j\omega$ and rewrite γ as

$$\gamma(s) = [(R_a + sL_a)(G_a + sC_a)]^{\frac{1}{2}}$$

which for low-loss conditions is

$$\gamma(s) \approx \left(\alpha + \frac{s}{c}\right)$$

where $c = 1/\sqrt{L_a C_a}$ is the sound velocity [see Eq. (3.9)].

Since the vocal tract is a distributed system, its transmission characteristics involve transcendental functions. However, to represent the terminal behavior by lumped-constant electrical networks, it is necessary to describe the vocal transmission in terms of rational, meromorphic functions. Because the transcendental transfer functions for the vocal tract are meromorphic, and because their numerator and denominator components are generally integral functions (i.e., analytic for all finite values of the complex variable), it is possible to approximate the transmission by rational functions. A relation in function theory (Titchmarsh [1932]) says that if $f(z)$ is an integral function of the complex variable z, and meets certain restrictions, it can be represented by the product series

$$f(z) = f(0)e^{z\frac{f'(0)}{f(0)}} \prod_{m=1}^{\infty} \left(1 - \frac{z}{z_m}\right) e^{z/a_m} \tag{11.2}$$

where the a_m's are the ordered, simple zeros of $f(z)$. For the vocal transmission (11.1), the zeros of the denominator (or the poles of the transmission) occur for

$$\gamma(s) = \pm j\frac{(2n-1)\pi}{2l}, \quad n = 1, 2, \ldots{}^{3}$$

or

$$\gamma^2(s) = -\frac{(2n-1)^2\pi^2}{4l^2} = (R_a + sL_a)(G_a + sC_a)$$

or, dropping the subscript a's,

$$\begin{aligned} s_n &= -\left(\frac{R}{2L} + \frac{G}{2C}\right) \pm j\left[\frac{(2n-1)^2\pi^2}{4l^2LC} - \left(\frac{R}{2L} - \frac{G}{2C}\right)^2\right]^{\frac{1}{2}}, \quad n = 1, 2, \ldots \\ &= -\sigma_n + j\omega_n \end{aligned} \tag{11.3}$$

For small loss

$$s_n \approx -\alpha c \pm j\frac{(2n-1)\pi c}{2l}, \quad n = 1, 2, \ldots \tag{11.4}$$

which [except for the change to $(2n - l)$, $n = 1, 2, \ldots$] is the same as Eq. (4.30) in Chapter 4. Substituting the result (11.3) in (11.2) gives

$$\cosh z = \prod_{n=1}^{\infty} \left[1 - \frac{z}{\pm j\frac{(2n-1)\pi}{2}}\right], \tag{11.5}$$

where $z = \gamma(s)l$. [The initial two terms of (11.2) yield unity, and the final term multiplies to unity because the roots of $f(z)$ are conjugate imaginaries.] For small loss $\gamma(s)l \approx (\alpha + s/c)l$ and

$$\frac{1}{\cosh \gamma(s)l} = \prod_n \frac{\pm j(2n-1)\pi c/2l}{s + \alpha c \pm j\frac{(2n-1)\pi c}{2l}} \tag{11.6}$$

---

[3] In Chapter 4 this result was written

$$\gamma = \pm j\frac{(2n+l)\pi}{2l}, \quad n = 0, 1, 2, \ldots$$

(see Eq. (4.29)). For the present discussion it will be convenient to write $(2n-1)$, $n = 1, 2, \ldots$. This has the mnemonic nicety that n may also represent the formant number.

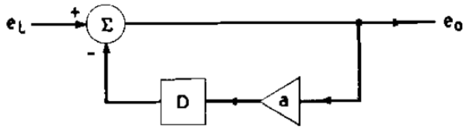


Figure 11.6: Feedback circuit for producing a transmission having uniformly spaced complex conjugate poles

$$
\begin{aligned}
&= \prod_n \frac{\omega_n^2}{(s - s_n)(s - s_n^*)} \\
&\approx \prod_n \frac{s_n s_n^*}{(s - s_n)(s - s_n^*)}
\end{aligned}
$$

which is Eq. (4.31) in Chapter 4.

As (11.4) indicates, the poles for the straight pipe are uniformly spaced at $\pi c/l$ intervals along the $j\omega$-axis. In this particular case, a very simple electrical circuit will realize the transmission function, namely the feedback circuit shown in Fig. 11.6. Its transmission is

$$\frac{e_0}{e_1} = H(s) = 1 - ae^{-sD} + a^2e^{-2sD} - \ldots = \frac{1}{1 + ae^{-sD}}, \tag{11.7}$$

where a is a positive-real gain less than unity, and D is a simple delay equal to twice the sound transit time through the pipe. The impulse response therefore simulates the multiple reflections, with some loss, that occur at the ends of the pipe. The poles of $H(s)$ occur at

$$s_n = -\frac{1}{D}\ln\frac{1}{a} \pm j\frac{(2n-1)\pi}{D}, \quad n = 1, 2, \ldots \tag{11.8}$$

If $D = 2l/c$ and $a = e^{-2\alpha l}$, the poles are identical to (11.4).

For a nonuniform pipe, the transmission (11.6) will generally have its poles spaced nonuniformly in frequency. In such a case, one simple way to realize the vocal transmission with electrical circuits is by "building up" the function in terms of the individual pole-pairs. This can be done by cascading individual, isolated electrical resonators, suitably tuned. This approach has the advantage of a one-to-one relation between speech formants and resonator poles, and it provides for non-interacting control of the resonances.

11.4.2 Spectral Contribution of Higher-Order Poles

On perceptual grounds it is usually sufficient to simulate only the first several (three to five) modes of the tract. The remaining modes can be accounted for by a single multiplicative term representing their summated influence upon the amplitude (magnitude) spectrum (Fant [1960]). This factor, following the technique of Fant, then becomes simply a frequency-equalizing network. Assuming the higher modes to be approximately those of a straight pipe, the nature of the equalizer can be set down directly.

Write Eq. (11.6) as two product series:

$$P(s) = \prod_{n=1}^{k} \frac{s_n s_n^*}{(s - s_n)(s - s_n^*)} \cdot \prod_{n=k+1}^{\infty} \frac{s_n s_n^*}{(s - s_n)(s - s_n^*)} \tag{11.9}$$

$$= P_k(s) \cdot Q_k(S),$$

where $s_n = (-\sigma_n + j\omega_n)$. For $s = j\omega$,

$$Q_k(j\omega) = \prod_{n=k+1}^{\infty} \frac{\omega_{0n}^2}{(\omega_{0n}^2 - \omega^2) + j2\sigma_n\omega} \tag{11.10}$$

where $\omega_{0n}^2 = (\sigma_n^2 + \omega_n^2)$.

Taking the magnitude,

$$|Q_k(j\omega)| = \prod_{n=k+1}^{\infty} \frac{\omega_{0n}^2}{[(\omega_{0n}^2 - \omega^2)^2 + (2\sigma_n\omega)^2]^{\frac{1}{2}}} \tag{11.11}$$

For low loss $\sigma_n \ll \omega_n$, and

$$|Q_k(j\omega)| \approx \prod_{n=k+1}^{\infty} \frac{1}{\left(1 - \frac{\omega}{\omega_n^2}\right)} \tag{11.12}$$

Taking the logarithm of both sides gives

$$\ln|Q_k(j\omega)| = -\sum_{n=k+1}^{\infty} \ln\left(1 - \frac{\omega}{\omega_n^2}\right)$$

Expanding the logarithm as a series and taking only the first term (to approximate the behavior at frequencies $\omega < \omega_n$) yields

$$\ln|Q_k(j\omega)| \approx \omega^2 \sum_{n=k+1}^{\infty} \frac{1}{\omega_n^2}$$

where

$$\omega_n = (2n-1)\omega_1 = \frac{(2n-1)\pi c}{2l}, \quad n = 1, 2, \ldots$$

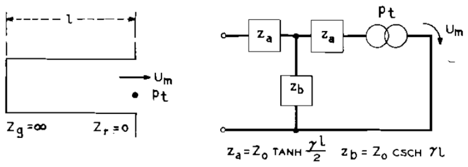


Figure 11.7: Front excitation of a straight pipe by a pressure source

(that is, the modes for the straight pipe of length l). Alternatively, the logarithm may be written

$$\ln |Q_k| \approx \left(\frac{\omega}{\omega_1}\right)^2 \sum_{n=k+1}^{\infty} \frac{1}{(2n-1)^2} \tag{11.13}$$

But

$$\sum_1^{\infty} \frac{1}{(2n-1)^2} = \frac{\pi^2}{8},$$

and the sum in (11.13) may be written

$$\sum_1^{\infty} \frac{1}{(2n-1)^2} = \frac{\pi^2}{8} - \sum_1^{k} \frac{1}{(2n-1)^2} \tag{11.14}$$

Therefore,

$$\ln |Q_k| \approx \left(\frac{\omega^2}{\omega_1^2}\right) \left[\frac{\pi^2}{8} - \sum_1^{k} \frac{1}{(2n-1)^2}\right] = \left(\frac{\omega^2}{\omega_1^2}\right) [R(k)]$$

or

$$|Q_k| \approx e^{(\omega/\omega_1)^2 R(k)} \tag{11.15}$$

where $R(k)$ is a positive-real function of k, the highest pole accounted for on an individual basis.

11.4.3 Non-Glottal Excitation of the Tract

The discussion of Chapter 4 showed that if the vocal excitation occurs at some point other than at the end of the tract, the transmission function will exhibit zeros as well as poles. This can be simply illustrated for front excitation of a straight pipe by a pressure source, as shown in Fig. 11.7. The ratio of mouth current to the source pressure is simply the driving point impedance of the mouth, or

$$\begin{aligned} \frac{U_m(s)}{p_t(s)} &= \frac{1}{Z_0} \tanh \gamma(s) l \qquad (11.16) \\ &= \frac{1}{\cos \gamma(s) l} \cdot \frac{\sinh \gamma(s) l}{Z_0} \\ &= P(s) \cdot Z(s). \end{aligned}$$

Since $P(s)$ has no zeros, the zeros of the transmission are the zeros of $Z(s)$ and occur for

$$(e^{2\gamma l}-1)=0 \tag{11.17}$$

$$\begin{aligned}\gamma &= \pm j\frac{m\pi}{l}, \quad m=0,1,2,\ldots\\ \gamma^2 &= \frac{-m^2\pi^2}{l^2}=\left[(R+sL)(G+sC)\right].\end{aligned}$$

The zeros therefore lie at

$$s_m=-\left(\frac{R}{2L}+\frac{G}{2C}\right)\pm j\left[\frac{m^2\pi^2}{l^2LC}-\left(\frac{R}{2:}-\frac{G}{2C}\right)^2\right]^{\frac{1}{2}},$$

or, again for small losses,

$$s_m\approx\left(-\alpha c\pm j\frac{m\pi c}{l}\right)\quad m=0,1,2,\ldots. \tag{11.18}$$

The poles of the transmission are the same as given in Eq. (11.4), and the poles and zeros in this instance alternate in the $j\omega$-direction.

Applying the product series formula in Eq. (11.2) gives

$$\sinh z=z\prod_{m=1}^{\infty}\left(1-\frac{z}{\pm jm\pi}\right),$$

where

$$z=\gamma l\approx\left(\alpha l+s\frac{l}{c}\right). \tag{11.19}$$

Then

$$\sinh\gamma l=\left(\alpha l_s\frac{l}{c}\right)\prod_{m=1}^{\infty}\left(1-\frac{\alpha l+s\frac{l}{c}}{\pm jm\pi}\right) \tag{11.20}$$

$$\begin{aligned}&= \frac{l}{c}(\alpha c+s)\prod_{m=1}^{\infty}\left(\frac{s+\alpha c\pm jm\pi c}{\pm j\frac{m\pi c}{l}}\right)\\ &\approx \frac{l}{c}(s+s_0)\prod_{m=1}\infty\frac{(s-s_m)(s-s_m^*)}{s_ms_m^*},\end{aligned}$$

where $s_0=-\alpha c$.

11.4.4 Spectral Contribution of Higher-Order Zeros.

The series for the zero terms can be "truncated" as described previously for pole terms, and a spectral correction factor can be obtained for higher-order zeros. Following the technique of Eq. (11.9),

$$Z(S) \approx \frac{1}{cZ_0}(s+s_0)\prod_{m=1}^{k}\frac{(s-s_m)(s-s_m^*)}{s_m s_m^*}\cdot|Y_k(s)|,$$

where

$$\ln|Y_k(j\omega)| \approx -\sum_{m=k+1}^{\infty}\frac{\omega^2}{\omega_m^2} \tag{11.21}$$

$$\approx -\frac{\omega^2}{\omega_1^2}\sum_{m=k+1}^{\infty}\frac{1}{m^2}$$

and where $\omega_1 = \pi c/l$.

The summation may be rewritten as

$$\ln|Y_l(j\omega)| \approx -\frac{\omega^2}{\omega_1^2}\left[\frac{\pi^2}{6}-\sum_{m=1}^{k}\frac{1}{m^2}\right],$$

or

$$|Y_k(j\omega)|^2 \approx e^{-(\omega^2/\omega_1^2)T(k)}, \tag{11.22}$$

where $T(k)$ is a positive-real function of the zero number k. Except for the sign of the exponent, this is the same form as (11.15). The factor $|Y_k(j\omega)|$ can therefore be realized by a frequency-equalizing network in conjunction with the variable poles and zeros of a formant synthesizer.

This simple example of front excitation illustrates that the vocal transmission, in general, involves poles $[P(s)]$ as well as zeros $[Z(s)]$. In the example, the zeros (like the poles) are uniformly distributed in frequency. For the nonuniform vocal tract, the mode frequencies will generally be irregularly distributed. Besides being dependent upon source location, zeros of transmission can also arise from side-branch paths coupled to the main transmission path. Cases in point are nasal consonants, nasalized vowels and perhaps liquids such as /l/[4]. In all cases where the sound radiation is from a single port (i.e., either mouth or nostril), the vocal transmission is minimum phase. For simultaneous radiation from mouth and nostril (as in a nasalized vowel) the transmissions to individual ports are minimum phase, but the combined response at a fixed point in front of the speaker may be nonminimum phase.

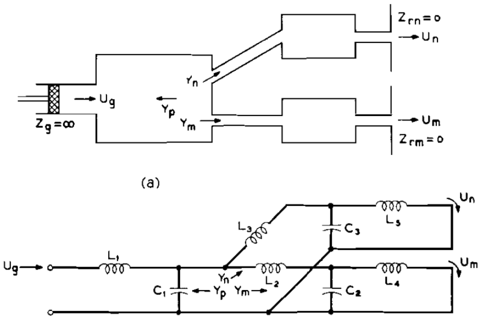


Figure 11.8: Simplified configuration illustrating coupling between oral and nasal cavities

11.4.5 Effects of a Side-Branch Resonator

The effect of a nasal or oral side branch can be simply illustrated by the circuit of Fig. 11.8a. For very low frequencies the circuit may be treated in terms of lumped-constant approximations to the major cavities and constrictions, as illustrated in Fig. 11.8b. The poles occur at frequencies where the sum of the admittances at any network node is zero. The velar junction is a convenient point to consider. Neglecting losses, the respective admit- tances for the low-frequency approximation are

$$Y_n = \frac{s^2 + \frac{1}{L_5 C_3}}{sL_3\left[s^2 + \frac{1}{C_3}\left(\frac{1}{L_3} + \frac{1}{L_5}\right)\right]} \tag{11.23}$$

$$Y_m = \frac{s^2 + \frac{1}{L_4 C_2}}{sL_2\left[s^2 + \frac{1}{C_2}\left(\frac{1}{L_2} + \frac{1}{L_4}\right)\right]}$$

$$Y_p = sC_1$$

or for real frequencies $s \to j\omega$,

$$Y_n = \frac{\omega_{n0}^2 - \omega^2}{j\omega L_3(\omega_{np}^2 - \omega^2)} \tag{11.24}$$

$$Y_m = \frac{\omega_{m0}^2 - \omega^2}{j\omega L_2(\omega_{mp}^2 - \omega^2)}$$

$$Y_p = j\omega C_1$$

where ω_{n0} and ω_{m0} are the zeros of the nasal and mouth admittances respectively, and ω_{np} and ω_{mp} are the poles of the nasal and mouth admittances.

The poles of the system occur at frequencies for which

$$\sum Y = Y_n + Y_m + Y_p = 0$$

---

[4] The cul-de-sac formed by the tongue can act as a side-branch resonator.

or

$$\omega^2 C_1 = \frac{\omega_{n0}^2 - \omega^2}{L_3(\omega_{np}^2 - \omega^2)} + \frac{\omega_{m0}^2 - \omega^2}{L_2(\omega_{mp}^2 - \omega^2)} \tag{11.25}$$

The low-frequency zero of U_n/U_m is ω_{mp}, and the zero of U_m/U_g is ω_{np}.

It is instructive to consider the loci of the low frequency modes for a highly simplified situation. Suppose the pharyngeal, oral and nasal cavities (C_1, C_2, C_3) are held fixed in size, and the mouth and velar constrictions (L_2, L_3, L_4) are varied. Suppose the velar areas are such that $(A_n + A_m) = A_0 =$ constant, so that L_2 and L_3 are inversely related. Assume that all tube lengths are held fixed so that area variations alone constitute the lumped element variation. Consider the low frequency mode behavior corresponding to the sequence: vowel → nasalized vowel → nasal, as in /ɑm/. The simplified articulatory sequence is: vowel, with the nasal tract decoupled and sealed off and the mouth open; nasalized vowel, with the velum partially open and the mouth still open; and nasal, with the velum full open and the mouth closed.

For the vowel, the nasal coupling is nil and $L_3 \approx \infty$. The frequencies ω_{n0} and ω_{np} are equal (i.e., the pole and zero are coincident) and $Y_n = 0$. The poles of the glottis-to-mouth transmission occur at frequencies where $Y_m = Y_p$. As the vowel is nasalized, the velum opens, L_3 diminishes and L_2 increases. ω_{n0} remains fixed, but ω_{np} parts from ω_{n0} and moves up in frequency. ω_{n0} becomes the zero of glottis-to-mouth transmission. In a similar manner ω_{m0} remains fixed, but ω_{mp} moves down. The exact trajectories of the system modes depend upon the relative sizes or the nasal and oral cavities, but, in general, the original vowel poles move up in frequency. A new pole is introduced in the region above ω_{n0} by the parting of ω_{n0} and ω_{np}.

As the mouth closes to produce the nasal, L_4 becomes infinite and all sound radiation transfers to the nostril. The closed oral cavity now acts as a side branch resonator for the glottis-to-nostril transmission. ω_{m0} now goes to zero, and ω_{mp} becomes lower. ω_{mp} is the zero of glottisto-nostril transmission. The first system pole is relatively low in frequency, and the second resides in the vicinity of ω_{mp}. The third is generally somewhat higher than ω_{np}. A more detailed computation, using an idealized vocal configuration, has been given previously in Fig. 4.34. Representative frequency positions for a nasal such as /m/ are approximately 250, 1100, 1350 and 2000Hz for the first four poles and 1300Hz for the zero. More extensive analyses of nasals can be round in the literature (Fujimura [1962]).

So long as the radiation is from a single port, the *dc* transmission to that port is essentially unity. For simultaneous radiation from mouth and nostril, the sound energy divides according to the oral and nasal admittances, and the *dc* transmission to a single port is determined by the respective branch losses.

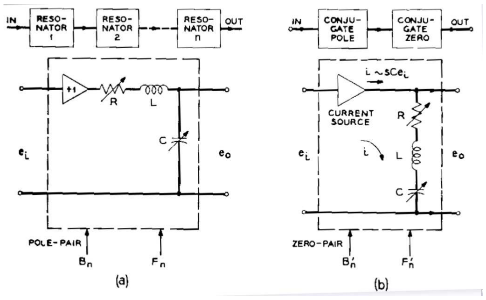


Figure 11.9: (a) Cascade connection of isolated RLC resonators for simulation of vocal transmission for vowel sounds. Each pole-pair or vocal resonance is simulated by a series circuit. (b) Cascaded pole and zero circuit for simulating low frequency behavior of a side branch resonator. The zero pair is approximated by the transmission of a simple series circuit

11.4.6 Cascade Type Synthesizers

The intent of these elementary considerations is to indicate that for all configurations and excitations, the vocal transmission $T(s)$ may be approximated in terms of its first few (low-frequency) poles and zeros, that is, the first several roots of $P(s)$ and $Z(s)$. A straightforward means for simulating the vocal transmission electrically is to build up the product functions in terms of the individual poles and zeros by cascading individual electrical resonators. As the preceding discussion showed, the transmission function for a vowel sound can be written

$$T(s) = P(s) = \prod_n \frac{s_n s_n^*}{(s - s_n)(s - s_n^*)}.$$

Such a function can be represented in terms of its individual poles by the isolated, cascaded, series RLC resonators shown in Fig. 11.9a. Here the transmission of a single resonant circuit is

$$\frac{e_0(s)}{e_i(s)} = \frac{\frac{1}{LC}}{s^2_{\frac{R}{L}} s_{\frac{1}{LC}}} = \frac{s_n s_n^*}{(s - s_n)(s - s_n^*)},$$

where

$$\omega_n = \sqrt{\frac{1}{LC} - \frac{R^2}{4L^2}}, \quad \sigma_n = \frac{R}{2L} \tag{11.26}$$

and

$$s_n = -\sigma_n + j\omega_n.$$

Control of the formant tuning is effected by changes in the tuning capacitor C. Control of formant bandwidth is accomplished by variation of R. For the serial connection of resonators, specification of the pole frequencies s_n implies specification of the spectral peaks, or formant amplitudes, as well. This point has been treated in some detail in the literature (Fant [1956], Flanagan [1957b]). The results of Chapter 4 and the preceding discussion (Fig. 11.8) suggest that sounds such as unvoiced consonants, nasals, nasalized vowels, and perhaps liquids, may have at least one low-frequency zero which might be perceptually significant[5]. In particular, a pole-zero pair additional to the usual vowel formants is commonly associated with nasals and nasalized vowels. The transmission of the vowel resonator string of Fig. 11.9a can be simply modified to accomodate this condition. A resonance and an antiresonance–as shown in the upper part of Fig. 11.9b–can be included in the synthesizer circuit (Flanagan *et al.* [1970]). So long as a pure vowel is to be produced, the added pole and zero are made coincident in frequency, and their transmission is unity. For nasal production they are pulled apart and set to appropriate values corresponding to the relations for the side branch resonator.

Practically, the complex conjugate zero can be approximated by the electrical circuit shown in the lower part of Fig. 11.9b. Its transmission is

$$\frac{e_0(s)}{e_i(s)} = LC\left(s^2 + s\frac{R}{L} + \frac{1}{LC}\right) \tag{11.27}$$

which is the reciprocal of the conjugate pole. As in the pole-pair resonator, the low frequency (dc) gain is made unity–which is proper so long as radiation occurs from a single port, and is approximately correct for the mouth radiation of nasalized vowels.

The front-excited voiceless sounds can also be approximated in terms of their poles and zeros. Following the results of the previous discussion and of Chapter 4, a reasonable approximation is given by

$$T(s) = P(s)\cdot Z(s) = K\cdot s\cdot \frac{\prod_m (s-s_m)(s-s_m^*)}{\prod_n (s-s_n)(s-s_n^*)}, \tag{11.28}$$

where an m and n of order 1 or 2 often suffice perceptually (in addition to higher-order pole and zero corrections). The zero at zero frequency arises because of the essentially closed back cavity (see Fig. 3.31). The amplitude scale factor K is accounted for by an over-all amplification.

11.4.7 Parallel Synthesizers

The vocal tract transmission has been represented as a ratio of product series which, when truncated, produce rational meromorphic functions. Because the poles are simple, the transmission can be expanded as a partial fraction with first-degree terms

$$T(s) = P(s)Z(s) = \sum_n \frac{A_n}{(s-s_n)} + \frac{A_n^*}{(s-s_n^*)}, \quad n = 1, 2, \ldots$$

---

[5] The perceptual effects of spectral zeros–both of the excitation and of the system–have not been thoroughly established. The extent to which the quality of synthetic speech depends upon these factors is a current problem in research. It will be discussed further in a later section.

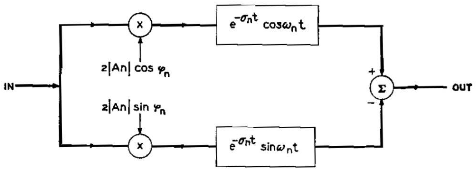


Figure 11.10: Circuit operations for simulating the time-domain response of Eq. (11.30)

$$= \sum_n \frac{2a_n s + 2(\sigma_n a_n + \omega_n b_n)}{s^2 + s\sigma_n s + (\sigma_n^2 + \omega_n^2)}, \tag{11.29}$$

where $s_n = (-\sigma_n + j\omega_n)$, and $A_n = (s - s_n)T(s)|_{s \to s_n} = (a_n + jb_n)$ is the residue in the n-th pole and is a function of all the poles and zeros. The inverse transform is

$$h(t) = \sum_n 2|A_n|e^{-\sigma_n t}\cos(\omega_n t + \phi_n),$$

where

$$A_n = |A_n|e^{j\phi_n}$$

Expanding the cosine term, $h(t)$ may be rewritten

$$h(t) = \sum_n 2|A_n|e^{-\sigma_n t}\left[\cos\phi_n \cos\omega_n t - sin\phi_n \sin\omega_n t\right]. \tag{11.30}$$

Each term of the latter expression can be realized by the operations shown in Fig. 11.10, where the filters represented by the boxes are simple resonant circuits.

If the transmission function is for pure vowels, $Z(s) \to 1$ and $T(s) \to P(s)$ and the transmission has only poles. Its numerator is not a function of s, but only of the s_n, that is, $\prod_n s_n s_n^* = f(s_n)$. The residue in the q-th pole is then

$$A_q = \frac{f(s_n)}{2j\omega_q \prod_{n\neq q}\left[(\sigma_n - \sigma_q)^2 + (\omega_n^2 - \omega_q^2) + 2j\omega_q(\sigma_q - \sigma_n)\right]}. \tag{11.31}$$

If the σ's are essentially equal (a reasonable approximation for the lower modes of the vocal tract), then

$$A_q \approx \frac{f(s_n)}{2j\omega_q \prod_{n\neq q}(\omega_n^2 - \omega_q^2)},$$

or

$$A_q \approx \frac{f(s_n)}{2j\omega_q(-1)^{q-l}} \frac{1}{\prod_{n\neq q}|\omega_n^2 - \omega_q^2|}. \tag{11.32}$$

The residues are therefore pure imaginary (i.e., $\cos\omega_n = 0$) and their signs alternate with pole number. The inverse transform (impulse response) for this transmission

$$h(t) = \sum_n (-1)^{n-1} 2|A_n|e^{-\sigma_n t}\sin\omega_n t, \tag{11.33}$$

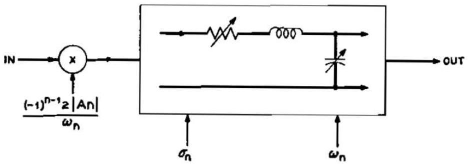


Figure 11.11: Circuit for simulating the vowel function impulse response [see Eq. (11.33)]

where each term can by synthesized by the electrical operations in Fig. 11.11. This circuit is essentially the lower branch of the previous circuit where now $-sin\phi_n = -\sin\left[(-1)^n(\pi/2)\right] = (-1)^{n-1}$, and the RCL resonator has an impulse response $(\omega_n e^{-\sigma_n t}\sin\omega_n t)$. Summation of the outputs of similar circuits, one for each n, produces the response (11.33).

The magnitude of the residue bears a simple approximate relation to the spectral magnitude at the formant frequency. Recall the residue magnitude is

$$|A_n| = \left|(s-s_n)T(s)\right|_{s\to s_n},$$

which for small damping $\sigma \ll \omega_n$ is approximately

$$\left|(s-s_n)T(s)\right|_{s\to j\omega_n} = \left|(j\omega_n - s_n)T(j\omega_n)\right| \approx |a_n|,$$

or

$$\sigma_n\left|T(j\omega_n)\right| \approx |A_n|. \tag{11.34}$$

If the transmission function exhibits zeros, as exemplified by Eq. (11.28), the residues are then

$$A'_q = (s-s_q)T(s)\big|_{s\to s_q} = Z(s)\cdot(s-s_q)P(s)\big|_{s\to s_q} \tag{11.35}$$

$$\begin{aligned} &= Z(s_q)\cdot A_q = Ks_q\left[\prod_m (s_q-s_m)(s_q-s_m^*)\right]A_q \\ &= A_qKs_q\cdot\prod_m\left[(\sigma_q-\sigma_m)^2+(\omega_m^2-\omega_q^2)+j2\omega_q(\sigma_q-\sigma_m)\right]. \end{aligned}$$

Again, if the σ's are nearly the same,

$$A'_q = A_qKs_q\cdot\prod_m(\omega_m^2-\omega_q^2), \tag{11.36}$$

and the sign of A'_q is determined by the relative magnitudes ω_m and ω_q. Or,

$$A'_q = A_q(-1)^pKs_q\prod_m\left|\omega_m^2-\omega_q^2\right|, \tag{11.37}$$

where p is the number of zeros lying below the pole ω_p. Or, substituting for A_q from Eq. (11.32),

$$A'_q = \frac{f(s_n)(-1)^p K s_q \prod_m |\omega_m^2 - \omega_q^2|}{2j\omega_q(-1)^{q-1} \prod_{n\neq q} |\omega_n^2 - \omega_q^2|}, \tag{11.38}$$

and the net sign of the residue is determined by the difference between the numbers of poles and zeros lying below the q-th pole. Again the residue bears a simple approximate relation to the realfrequency spectrum evaluated at the pole frequency. That is,

$$A_n = (s - s_n)T(s)|_{s\to s_n},$$

but for low damping $s_n \to j\omega_n$,

$$A_n \approx (j\omega_n - s_n)T(j\omega_n) \tag{11.39}$$

$$\begin{aligned} An &\approx \sigma_n T(j\omega_n) = \sigma_n |T(j\omega_n)| \angle T(j\omega_n) \\ A_n &= |A_n| e^{j\phi_n} \end{aligned}$$

A number of terminal-analog synthesizers, both of the parallel and cascade types, have been constructed and operated. (See for example, (Fant [1959b], Bastide and Smith [1955], Lawrence [1953], Stead and Jones [1961], Campanella *et al.* [1962], Chang [1956], Flanagan [1956a, 1960b]).) Most of the devices utilize one or more of the relations discussed–either by overt recognition of the principles or by qualitative implication. The transmission relations commonly exploited involve the formant frequency and the magnitude of the residue, or the formant frequency and amplitude.

At least one study has considered use of the complex residue, that is, the angle or sign of the residue. In this case, analysis of the short-time phase spectrum of speech[6]–in conjunction with the short-time amplitude spectrum–is used to gain naturalness. Specification of the complex residues, as implied by Eq. (11.29), is equivalent to specification of spectral zeros. A parallel formant synthesizer, implemented as described by Eq. (11.30) and using pitch-synchronous spectral analysis to obtain formant frequency and complex residue, produced speech of improved quality (Flanagan [1965]).

11.4.8 Digital Techniques for Formant Synthesis

The approximations made of vocal transmission in Section 11.4 can be represented by linear differential equations with constant coefficients. In turn, such equations can be approximated as linear difference equations. The difference equations can be programmed in a digital computer as arithmetic operations upon discrete values of

---

[6] See Eq. (5.4), Chapter 5, for a definition of the short-time phase spectrum.

the variables[7]. As an example, the input and output voltages for the series electrical resonator shown in Fig. 11.9a are related by

$$e_i = LC\frac{d^2e_0}{dt^2} + RC\frac{de_0}{dt} + e_0 \tag{11.40}$$

If the derivatives are approximated by differences between successive values of the dependent variable–sampled at uniform, discrete values of the independent variable–the equation can be written as

$$e_i = e_0 + RC\Delta e_0 + LC\Delta^2 e_0,$$

where Δ is the first backward difference divided by the sampling interval. Explicitly,

$$e_i(t_n) = e_0(t_n) + RC\left[\frac{e_0(t_n) - e_0(t_{n-1})}{(t_n - t_{n-1})}\right] + LC\left[\frac{e_0(t_n) - 2e_0(t_{n-1}) + e_0(t_{n-2})}{(t_n - t_{n-1})(t_{n-1} - t_{n-2})}\right] \tag{11.41}$$

Collecting terms

$$\begin{aligned} e_{in} &= e_{0n}\left[1 + \frac{RC}{D} + \frac{LC}{D^2}\right] - e_{0(n-1)}\left[\frac{RC}{D} + \frac{2LC}{D^2}\right] + e_{0(n-2)}\left[\frac{LC}{D}\right], \\ &= ae_{0n} + be_{0(n-1)} + ce_{0(n-2)} \end{aligned} \tag{11.42}$$

where $D = (t_n - t_{n-1})$ is the sampling interval and $e_{0n} = e_0(t_n)$.

The theory of linear difference equations (Hildebrand [1952]) shows that the unforced homogeneous solution ($e_{in} = 0$) of Eq. (11.42) is a linear combination of exponential terms

$$e_{0n} = K_1\beta_1^n + K_2\beta_2^n, \tag{11.43}$$

where β_1 and β_2 are the roots of the determinantal equation

$$a\beta^2 + b\beta + c = 0,$$

K_1 and K_2 are arbitrary constants, and a, b and c are defined in (11.42). In the present instance the roots will be complex conjugate, and

$$\beta = -\frac{b \pm j\sqrt{4ac - b^2}}{2a} = e^{r_1 \pm jr_2}, \tag{11.44}$$

where

$$e^{r_1} = \sqrt{\frac{c}{a}}$$

and

$$r_2 = \tan^{-1}\frac{\sqrt{4ac - b^2}}{-b}.$$

---

[7] Alternatively, special purpose digital hardware can accomplish the arithmetic operations.

Therefore,

$$e_{0n} = e^{r_1 n}\left(K_1' \cos r_2 n + K_2' \sin r_2 n\right),$$

where K_1' and K_2' are linear combinations of K_1 and K_2, and the response samples are those of a damped sinusoid. Following through the arithmetic gives

$$e^{r_1} = \left[\frac{1}{1 + 2\alpha D + \omega_0^2 D^2}\right]^{\frac{1}{2}},$$

where

$$\alpha = \frac{R}{2L} \quad \text{and} \quad \omega_0^2 = \frac{1}{LC},$$

and

$$r_1 = -\frac{1}{2}\ln\left[1 + 2\alpha D + \omega_0^2 D^2\right]. \tag{11.45}$$

Expanding the logarithm as a series for $\ln(l + x)$, $-1 < x < 1$, and taking the first term yields

$$r_1 \approx -D\left(\alpha + \frac{\omega_0^2 D}{2}\right).$$

For a sufficiently small sampling interval D,

$$\frac{\omega_0^2 D}{2} \ll \alpha$$

and

$$r_1 \approx -\alpha D,$$

and the response samples are damped approximately as $e^{-\alpha n D}$, which is similar to the solution for the continuous equation.

In the same fashion

$$r_2 = \tan^{-1} D\left\{\frac{\left(\frac{1}{LC} - \frac{R^2}{4L}\right)}{1 + \frac{RD}{L} + \frac{R^2 D^2}{4L^2}}\right\}^{\frac{1}{2}} \tag{11.46}$$

$$r_2 = \tan^{-1} D\left\{\frac{(\omega_0^2 - \alpha^2)}{1 + 2\alpha D + \alpha^2 D^2}\right\}^{\frac{1}{2}}$$

$$r_2 = \tan^{-1}\frac{D\omega}{(1 + \alpha D)}.$$

so that for small values of sampling interval

$$r_2 \approx \frac{D\omega}{1 + \alpha D}$$

and for small damping $r_2 \approx D\omega$. The response samples are then approximately those of a damped sinusoid with angular frequency ω, which is the continuous equation solution. One notices, however, that if the sampling is coarse the solution to the difference equation begins to depart substantially from the sampled values of the continuous system. This situation can be improved by more sophisticated approximations to the derivative (which of course require additional computation). The trades which can be made between sampling rate and derivative approximation is a topic area worthy of study.

A different approach permits one to compute exact samples of the continuous impulse response. If, in addition, the sampling rate exceeds twice the bandwidth of the continuous signal, the continuous response can be reconstructed by low-pass filtering. The approach employs the z-transform. Consider the same series RLC formant resonator used in the preceding discussion [see Fig. 11.9a]. Its transfer function, in terms of a Laplace transform, is

$$\frac{e_0(s)}{e_i(s)} = F(s) = \frac{s_1 s_1^*}{(s-s_1)(s-s_1^*)} = \frac{A_1}{(s-s_1)} + \frac{A_1^*}{(s-s_1^*)} \tag{11.47}$$

where where

$$\begin{aligned} s_1 &= -\sigma_1 + j\omega_1 \text{ is the pole frequency,} \\ A_l &= \lim_{s\to s_1} (s-s_1)F(s) \text{ is the complex residue in pole } s_1, \end{aligned}$$

and the asterisk denotes complex conjugate. The inverse transform of $F(s)$ is the impulse response $f(t)$. Sampled values of the latter can be described as impulses with areas equal to the function at the sampling instants, that is,

$$f^\dagger(t) = \sum_{n=0}^{\infty} f(t)\delta(t-nD) \tag{11.48}$$

where $\delta(t)$ is a unit area impulse and $f^\dagger(t)$ is a periodic impulse train with period D representing the sample values $f(nD)$. The transform of $f^\dagger(t)$ is the complex convolution of the transform of its components, or

$$\mathcal{L}\left[f^\dagger(t)\right] = F^\dagger(s) = F(s) * \mathcal{L}\left\{\sum_n \delta(t-nD)\right\}.$$

But

$$\begin{aligned} \mathcal{L}\left\{\sum_n \delta(t-nD)\right\} &= 1 + e^{-sD} + e^{-2sD} + \ldots \\ &= \Delta(s) = \frac{1}{1-e^{-sD}} \end{aligned}$$

which has poles at $s = \pm j2m\pi/D$, $m = 0, 1, \ldots$. The convolution to be computed is

$$F^{\dagger}(s) = \frac{1}{2\pi j} \int_{c-j\infty}^{c+j\infty} F(\lambda)\Delta(s-\lambda)d\lambda. \tag{11.49}$$

Using the residue theorem and recognizing that the circuit is linear and passive so that the poles of $F(s)$ lie in the left half plane, the integral can be evaluated for a contour of integration enclosing only the poles of $F(s)$.

$$F^{\dagger}(s) = \sum_{k!\text{poles of } F(\lambda)} \mathrm{Res}\left[F(\lambda)\Delta(s-\lambda)\right]_{\lambda=\lambda_k},$$

or

$$F^{\dagger}(s) = \sum_k \left[\frac{1}{1-e^{-D(s-\lambda_k)}}\right] \mathrm{Res}\left[F(\lambda)\right]_{\lambda=\lambda_k}. \tag{11.50}$$

Making the substitution $e^{sD} = z$, Eq. (11.50) can be rewritten

$$F(z) = \sum_k \left[\frac{1}{1-e^{\lambda_k D}z^{-1}}\right] \mathrm{Res}\left[F(\lambda)\right]_{\lambda=\lambda_k}. \tag{11.51}$$

For the example at hand (that is, the single formant resonator)

$$\mathrm{Res}\left[F(s)\right]_{s=s_1} = A_1 = \left(\frac{\sigma_1^2+\omega_1^2}{j2\omega_1}\right),$$

and

$$F(z) = \frac{\sigma_1^2+\omega_1^2}{\omega_1}\left\{\frac{e^{-\sigma_1 D}z^{-1}(\sin\omega_1 D)}{1-2e^{-\sigma_1 D}(\cos\omega_1 D)z^{-1}+e^{-2\sigma_1 D}z^{-2}}\right\}. \tag{11.52}$$

Notice also that Eq. (11.49) can be written

$$F^{\dagger}(s) = \frac{1}{2\pi j} \int_{c-j\infty}^{c+j\infty} F(s-\lambda)\Delta(\lambda)d\lambda,$$

and that the poles of $\Delta(\lambda)$ are

$$\lambda = \pm j\frac{2m\pi}{D}, \quad m = 0, 1, 2, \ldots, \infty.$$

If the integration contour is selected to enclose the $j\omega$-axis poles of $\Delta(\lambda)$, then the integral is

$$F^{\dagger}(s) = \frac{1}{D}\sum_{m=-\infty}^{\infty} F\left(s - j\frac{2m\pi}{D}\right), \tag{11.53}$$

because the residue in any pole of $\Delta(\lambda)$ is $1/D$.

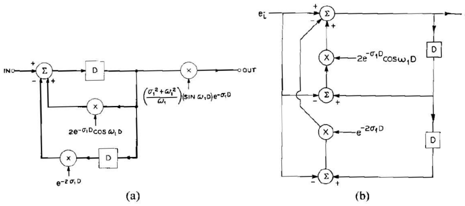


Figure 11.12: Digital operations for simulating a single formant resonance (pole-pair) (a) implementation of the standard z-transform; (b) practical implementation for unitydc gain and minimum multiplication

The system function represented by Eq. (11.50), or by Eq. (11.53), is a transform relating discrete samples of the input and output of the continuous system. Since $z^{-l} = e^{-sD}$ is a delay of one sample interval, D, the digital operations necessary to simulate the sampled response of the single formant resonator, given by Eq. (11.52), involve only delays, multiplications and summations. They are shown in Fig. 11.12a. If the $F(z)$ function in Eq. (11.52) is thought of in terms of the transmission of a common negative feedback amplifier,

$$G = \frac{K}{1+\beta K},$$

the return circuit connections in Fig. 11.12a become apparent.

The resonator of Fig. 11.12a has an impulse response equal to the sampled impulse response of the continuous function of Eq. (11.47). The frequency behavior for the two relations, however, are somewhat different. For example, their dc gains are

$$F(s)\big|_{s\to 0} = 1$$

and

$$F^{\dagger}(s)\big|_{s\to 0} = \frac{1}{D}\sum_{m=-\infty} F\left(-j\frac{2m\pi}{D}\right),$$

respectively. Digital resonators can, however, be specified in terms of their frequency behavior and without direct reference to continuous resonators (GOLD and RADER). Since the formant resonance must correspond to prescribed bandwidth and frequency, and since its de gain must be essentially unity, it is convenient in practice to modify (11.52) to

$$\frac{e_0(z)}{e_i(z)} = F(z) = \frac{1-2e^{-\sigma_1 D}\cos\omega_1 D + e^{-2\sigma_1 D}}{1-2e^{-\sigma_1 D}(\cos\omega_1 D)z^{-1} + e^{-2\sigma_1 D}z^{-2}}. \qquad (11.54)$$

This relation can be programmed for a minimum of two multiplications as shown in Fig. 11.12b. The origin of the configuration of Fig. 11.12b is easily seen by noting the output $e_0(z)$ is given by

$$\begin{aligned} e_0(z) &= (2e^{-\sigma_1 D}\cos\omega_1 D)(z^{-1}e_0 - e_i) \\ &\quad -(e^{-2\sigma_1 D})(z^{-2}e_0 - e_i) + e_i, \end{aligned}$$

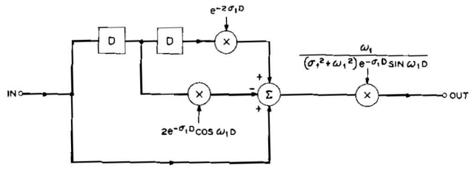


Figure 11.13: Digital operations for simulating a single anti-resonance (zero-pair)

where, as before, z^{-1} is the delay operator e^{-sD}.

The reciprocal of $F(z)$ has zeros where $F(z)$ has poles, so that the sampled-data equivalent of a simple complex conjugate zero is the reciprocal of Eq. 11.52

$$\frac{1}{F(z)} = \left(\frac{\omega_1}{\sigma_1^2+\omega_1^2}\right)\left[\frac{1-2e^{-\sigma_1 D}(\cos\omega_1 D)z^{-1}+e^{-2\sigma_1 D}z^{-2}}{e^{-\sigma_1 D}z^{-1}\sin\omega_1 D}\right]. \tag{11.55}$$

This response is physically unrealizable because the z^{-1} in the denominator implies an output prior to an input. Multiplication by z^{-1} to incur a unit sample delay does not alter the s-plane zero positions and makes the transmission function realizable by the digital operations shown in Fig. 11.13. As in the sampled data pole-pair, the frequency data ω_1 and the bandwidth control σ_1 are supplied to the multipliers. As with the digital resonator, it is practically convenient to have unity gain at zero frequency. The final gain multiplication in Fig. 11.13 can therefore be alternatively made $(l-2e^{-\sigma_1 D}\cos\omega_1 D+e^{-2\sigma_1 D})^{-1}$ to correspond to the reciprocal of the practical resonator shown in Fig. 11.12 b and in Eq. (11.54).

These basic pole and zero operations have been used to simulate a complete formant-vocoder synthesizer on a digital computer. One configuration of the synthesizer is shown in Fig. 11.14 (FLANAGAN, COKER, and BIRD). Voiced sounds are generated by the top branch which contains four variable poles and one variable zero. A fixed pole, not shown in the diagram, is included for high-frequency compensation. For vowels the final pole-zero pair is tuned coincidently so that its combined transmission is unity. Three poles therefore represent vowel spectra, in accordance with the acoustic relations developed in Section 4.7. For voiced nonvowels, such as the nasals, the final pole-zero pair is parted and positioned to represent relations given in Section 4.7.6. In general the pole-zero pair does not critically influence perception, provided the formant data are accurate, but is largely important to obtain realistic overall shape of the synthesized spectrum. Fundamental frequency, F_0, and amplitude of voicing, A_v are also controlled.

The unvoiced sounds are produced by the lower branch composed of one zero and either one or two poles. The amplitude of the noise is controlled by A_n. As Fig. 11.12 and 11.13 indicate, control of frequencies co; and bandwidths an is effected by supplying these values to the multiplying elements in the digital circuits. Image poles, produced at multiples of the sampling frequency [see Eq. (11.53)] make further correction for higher vocal resonances unnecessary. This feature, which must be

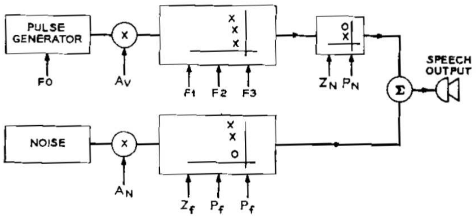


Figure 11.14: Block diagram of a computer-simulated speech synthesizer. (After (Flanagan *et al.* [1970]))

Figure 11.15: Spectrograms of synthetic speech produced by a computer-simulated formant synthesizer and of the original utterance. (After FLANAGAN, COKER and BIRD)

treated explicitly in analog synthesizers (see Section 11.4.1), comes free in the digital representation.

A typical listing of control data–as supplied to the computer on punched cards–is shown in Table 11.1. The data represent approximately 1 sec of synthetic speech. The first column is time in tens of milliseconds; the second, pitch inHz; the next two columns, relative amplitudes of buzz and hiss; and finally, the pole and zero frequencies inHz. Each value entered in the table is held by the circuit until a new value is specified. The control functions are interpolated between specified values in 2.5 msec steps. The sampling rate for the simulation is $l/D = 10KC$. A spectrogram of synthetic speech produced from such data is shown in Fig. 11.15. Also shown is the original speech from which the control functions were derived.

Digitally-simulated formant synthesizers–implemented either by programmed operations in general-purpose computers or as special-purpose digital hardware–have been used in a variety of forms (for example, (Jr. and Gerstman [1961], Flanagan *et al.* [1970], Rabiner [1968b,a])). Analog hardware synthesizers, controlled by digital computers, have over the past had even more extensive use (for example, (Coker and Cummiskey [1965], Holmes [1958], Mattingly and Shearme [1964], Dixon and Maxey [1968], Lee [1969], Ochiai and Kato [1949], Nakata [1961], Fujisaki [1960])). Digital implementations, however, have distinct advantages in stability and accuracy, and current advances in digital circuitry make commitment to full digital operation irresistible.

Table 11.1: Typical listing of control data for the computer-simulated synthesizer of Fig. 11.14

Time	Pitch	A_V	A_N	F_1	F_2	F_3	P_N	Z_N	Z_F	P_F
-20	107		0	170	1290	2190	750	1000	1750	3500
4										
5		100								
7				180	1260	2170	850	950		
8				210	1470	2270	900	900		
9				390	1550	2300				
10				400	1620	2380				
11						1690	2410			
12					1700	2460				
19					1690	2500				
23				410	1510	2430				
24				350	1490	2410				
25				300	1475	2400				
26				250	1490					
28				230	1510					
32				215	1620	2390				
35				210	1700	2330				
36		0	25							
37							610	610		
41									1655	3310
46									1500	2950
47									1400	2800
48			0	320	1420	1800				
51										
52			25						975	1950
54									960	1920
56	120								925	1850
57		100	0							
58	118				1390	1750				
61	112			450	1200	1700				
65	107			600	1140	1710				
70				690	1115	1910				
72				700	1150	2000				
78					1305	2070				

Much of the formant synthesis work over the past several years has made extensive use of interactive laboratory computers (see, for example, various work referenced in (Flanagan *et al.* [1970])). Expecially valuable have been small interactive computers of integrated circuit design. Their ability for high-speed arithmetic and logic operations, and their ability to store sizeable amounts of information (both in primary and secondary memories) has substantially aided work in speech analysis and synthesis (Flanagan [1971]). The interactive computer has become a common laboratory tool, and as digital technology continues to develop, laboratory computers will expand in sophistication and utility.

Formant synthesizers, digitally implemented or controlled, have been used in many studies of speech synthesis-by-rule and in computer synthesis from stored formant data. In synthesis-by-rule, discrete symbols representing each speech phoneme, its duration and pitch are supplied as input. Each specified phoneme calls up from storage a set of formant values appropriate to that phoneme. Transitions of the formant and excitation functions from one phoneme to another are determined by stored rules designed to approximate the constraints of natural speech. The ultimate in synthesis-by-rule is to convert printed language to speech.

Several studies have treated the problem of converting printed English to continuous speech (Teranishi and Umeda [1968], Coker *et al.* [1971], Lee [1969], Allen [1971]). In one of these (Coker *et al.* [1971]) a computer program uses a pronouncing dictionary to convert printed English text into discrete phonemic symbols, each carrying its own modifiers for pitch and duration. The text conversion is accomplished through a programmed syntax analysis and a prosodic feature determination. A dynamic model of the vocal tract (shown previously in Fig. 5.43) responds to the discrete phoneme commands to produce sequences of area changes similar to the real vocal tract. A difference equation solution of the Webster horn equation is periodically made to obtain continuous formant (eigenfrequency) data, and the latter are used to control a digital formant synthesizer to produce the synthetic speech.

A result of the automatic conversion of printed English into discrete control symbols for the synthesizer is shown in Table 11.2. These control symbols actuate articulatory motions in the vocal tract model of Fig. 5.43. The resulting synthetic output, compared with a similar human utterance is shown in Fig. 11.16. Formant motions, word durations and pitch are seen to be realisticaliy similar to natural speech.

In synthesis from stored formant data, libraries of formant-analyzed words, phrases or syllables reside in the machine along with rules for concatenating these elements into connected speech (Schafer and Flanagan [1971]). This approach has the advantage of using naturally-spoken signals to derive the so-called “segmental” information (i.e., the vocal resonances) rather than calculating these data. Additional storage is the price paid.

Input to the system is the English text for the word string to be generated, as illustrated in Fig. 11.17. From the library of words, stored as control functions for a formant synthesizer, the program selects and concatenates the sequence demanded.

Table 11.2: Discrete control symbols for synthesis from printed text. (After (Coker *et al.* [1971]))

English text	Syntax and prosodic rules output
the	4dh 4a
north	6n $4aw 2er 6th
wind	6w *qq5i 4n 4d
and	4aa -n -d
the	-dh 4a
sun	6s *qq5uh 6n
were	4w 4er
arguing	4: $q6ah -r -g -y 4uu 4i 6ng
one	4w &5uh 4n
day	6d *q9ay qq9<
, $,	
when	2h 2w &5eh 4n
a	4a
traveler	4t 4tr *q7aa -v 40 -I 4er
came	4k &4ay 4< 4m
along	4a 41 8aw 4ng
, s	,
wrapped	6r $q8aa 4p 4t
in	4i -n
a	4a
warm	6w $5ah 2er 6m
coat	6k *q20h qq20h 61

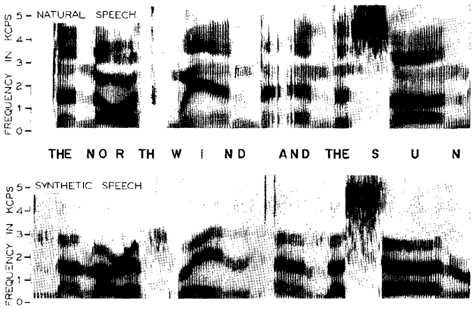


Figure 11.16: Spectrograms comparing natural speech synthesized directly from printed text. (After (Coker *et al.* [1971]))

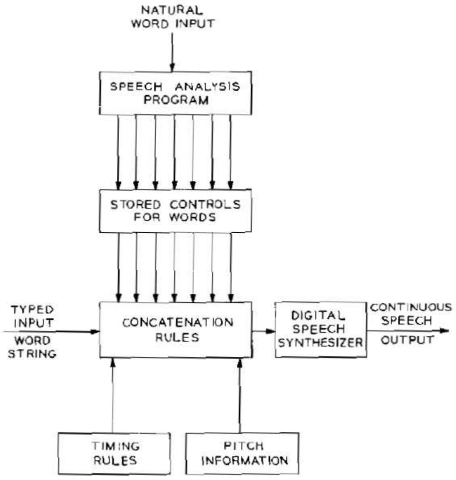


Figure 11.17: Programmed operations for synthesis from stored formant data. (After (Schafer and Flanagan [1971]).)

Formant functions must be interpolated naturally, word durations must be adjusted and pitch variations must be calculated for the connected utterance. The resulting control parameters are supplied to the formant synthesizer for conversion to speech.

Fig. 11.18 illustrates one technique of concatenating formant-analyzed words. At the top is a naturally-spoken sentence. At the bottom is a sentence produced from the same words spoken in isolation, formant analyzed and synthesized, and merely abutted in sequence. The differences are obvious and marked. The center is the result of concatenation of the same isolated words, but where the program imposes formant interpolation, word duration and pitch according to stored rules. The result is much more like the natural signal. In particular one can examine the effects on the /ə/ vowel in " ... away a year ...", seen at about 1000 msec in the top spectrogram. Mere abuttment renders this particularly badly at about 1400 msec in the bottom spectrogram. By rule, however, in the middle spectrogram, the sound is produced relatively well at about 1000 msec. This method of concatenation has been used successfully as an automatic answer back system for speaking seven-digit telephone numbers (Schafer and Flanagan [1971]).

Synthesis-by-rule and concatenation methods both depend critically upon the adequacy of rules for calculating prosodic information; i.e., duration, pitch and intensity variations. This problem represents a whole field of study in itself, and it is the focus of considerable interest in phonetics research.

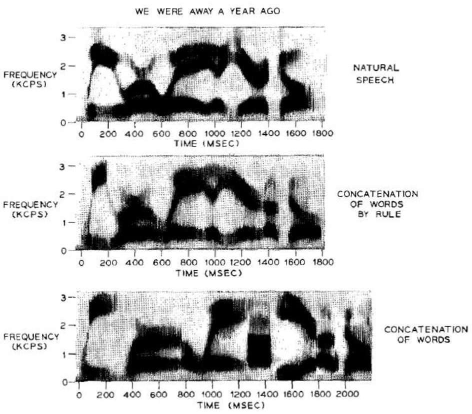


Figure 11.18: Computer synthesis by concatenation of formant coded words. (After (Schafer and Flanagan [1971]).)

11.5 Computer Simulation of the Articulatory System

11.5.1 Reflection-Line Analogs of the Vocal Tract

Formant synthesizers represent the input-output characteristics of the vocal tract. For this reason they are sometimes called "terminal" analogs. In many instances it is desirable to represent the distributed nature of the vocal system. Such representation is particularly important in efforts to model articulatory dynamics of speech (where the primary input data are physiological factors, such as the area function).

Distributed aspects of the system can be treated directly in terms of the wave-equation for one-dimensional wave propagation in a nonuniform pipe (i.e., Webster's horn equation, Eq. (**??**)), or in terms of bilateral transmission-line models of the system (see Section 11.5.2). In the first instance, steady-state solutions can be used to obtain the undamped eigen (formant) frequencies of non-nasal sounds, and transient solutions can be used to compute pressure and volume velocity distributions as functions of time. The Webster equation is

$$\frac{\partial^2 p}{\partial x^2} + \frac{1}{A}\frac{\partial p}{\partial x}\frac{\partial A}{\partial x} = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2}, \tag{11.56}$$

where $p = p(x,t)$ is the sound pressure as a function of distance and time and $A(x)$ is the vocal tract area as a function of distance[8]. For steadystate behavior $p = p(x)e^{j\omega t}$.

---

[8] The volume velocity satisfies an analogous equation

$$A\frac{\partial}{\partial x}\left(\frac{1}{A}\frac{\partial U}{\partial x}\right) = \frac{1}{c^2}\frac{\partial^2 U}{\partial t^2}.$$

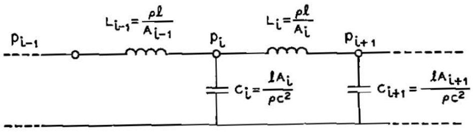


Figure 11.19: Ladder network corresponding to a difference-equation approximation of the Webster wave equation

For convenient numerical solution on a computer, the differential equation can be approximated by a difference equation. A number of possibilities exist for making the approximation. Consider space to be quantized into uniform small intervals $\Delta x = l$. Let a central second difference approximate the second derivatives and a first back difference approximate the first derivative, i.e.,

$$\left.\frac{d^2 f(x)}{dx^2}\right|_{x=x_i} = \left(\frac{f_{i+1} - 2f_i + f_{i-1}}{l^2}\right)$$

and

$$\left.\frac{df(x)}{dx}\right|_{x=x_i} = \left(\frac{f_i - f_{i-1}}{l}\right). \tag{11.57}$$

Then the steady-state pressure at point $x = x_{i+1}$ can be written as the recursion formula

$$p_{i+l} = \left[p_i\left(1 - \frac{\omega^2 l^2}{c^2} + \frac{A_{i-1}}{A_i}\right) - \left(\frac{A_{i-1}}{A_i}\right)p_{i-1}\right]. \tag{11.58}$$

This formulation has been used to calculate the undamped eigenfrequencies of the non-uniform tract (Coker [1968]). Typical boundary conditions for the pressure are $p_{glottis} \neq 0$, $p_{mouth} = 0$. Assuming non-zero pressure at the glottis, the pressures at successive points along the tract are calculated from the recursion formula. By iteration, the value of ω is found that satisfies the mouth boundary. Convergence to the eigenfrequencies is facilitated by observing the number of pressure nodes along the tract which a given value of ω produces. That is, the first eigenfrequency corresponds to a quarter-wave resonance with no pressure nodes along tract; and the second formant to a three-quarter wave resonance with one node. This computation is repeated periodically as the $A(x)$ function changes with time.

It is relevant to note that the difference equation (11.58), so formulated, corresponds to representing the tract by a simple L-section ladder network with the LC elements shown in Fig. 11.19. The node equation relating the pressures p_{i-1}, p_i, p_{i+1} is identical to Eq. (11.58).

Another technique, useful for digital calculation of the transient sound pressure along the vocal tract, is a representation in terms of reflection coefficients (Jr. and Lochbaum [1962a]). This approach depends upon initially approximating the non-uniform pipe by right-circular elements and assuming plane-wave propagation in each section, as discussed in Chapter 4.

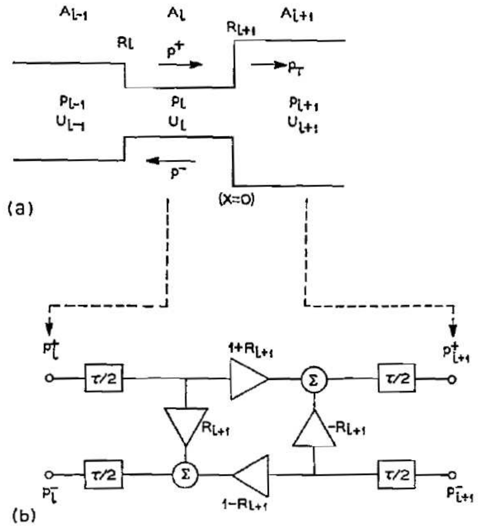


Figure 11.20: Representation of an impedance discontinuity in terms of reflection coefficients

Consider a plane wave moving from the left in the pipe shown in Fig. 11.20a and encountering an impedance discontinuity at $x = 0$. The steady-state pressure and volume velocity in the left tube must satisfy

$$p_i(x) = \left(p^+ e^{-jkx} + p^- e^{jkx}\right)$$

$$U_i(x) = \frac{1}{Z_i}\left(p^+ e^{-jkx} - p^- e^{jkx}\right)$$

where p^+ and p^- are the magnitudes of plane progressive waves moving to the right and the left respectively in the tube section with area A_i, $k = \omega/c$ and Z_i is the characteristic impedance of the left tube. (The pressure and particle velocity in a plane wave are linked by $dp/dx = -j\omega\rho u$.) Since pressure and volume velocity are continuous at the boundary,

$$p_i(0) = p_{i+1}(0) = \left(p^+ + p^-\right)$$

$$U_i(0) = U_{i+1}(0) = \frac{1}{Z_i}\left(p^+ - p^-\right), \tag{11.59}$$

where the subscripts i and $i+1$ correspond to the tube elements A_i and A_{i+1}. If the right-hand tube were infinitely long with characteristic impedance Z_{i+1} a plane wave transmitted and continuing to the right would have magnitude $p_T = (p^+ + p^-)$ and must satisfy

$$\frac{p_T}{U_{i+1}} = Z_{i+1} = \frac{Z_i(p^+ + p^-)}{p^+ - p^-}. \tag{11.60}$$

Then, the left-going wave in the left pipe is

$$p^- = \left(\frac{Z_{i+1} - Z_i}{Z_{i+1} + Z_i}\right) p^+ = R_{i+1} p^+$$

and

$$p_T = \left(p^+ + p^-\right) = \left(1 + R_{i+1}\right) p^+, \tag{11.61}$$

where R_{i+1} is the reflection coefficient at the junction of A_i and A_{i+1}. If the tubes are lossless, their characteristic impedances are real

$$Z_i = \rho c / A_i; \quad Z_{i+1} = \rho c / A_{i+1}$$

and

$$R_{i+1} = \left(\frac{A_i - A_{i+1}}{A_i + A_{i+1}}\right). \tag{11.62}$$

For a plane wave coming originally from the right, instead of the left, the sign of R_{i+1} is changed.

The Eq. (11.61) can therefore be used to represent each junction in a cascade of right-circular elements which approximate the non-uniform tract. The relations for right and left going waves are given in Fig. 11.20b, where the delay τ is the transit time through each section, $\tau = l/c$, and the unilateral amplifier boxes denote multiplication by the indicated parameters. (The $\tau/2$ delays can be lumped into single τ delays, one in the lower branch, one in the upper branch without altering the behavior.)

For any section of the quantized pipe, recursion equations describe the transient values of the $(+)$ and $(-)$ waves. The temporal sampling times correspond to the transit times through the uniform sections. Using i as the spatial index and j as the temporal index, the difference equations are

$$\begin{aligned} p^+_{i,j} &= -R_i p^-_{i,j-1} + p^+_{i-1,j-1}(1 + R_i) \\ p^-_{i,j} &= R_{i+1} p^+_{i,j-1} + p^-_{i-1,j-1}(1 - R_{i+1}) \\ p_{i,j} &= (p^+_{i,j} + p^-_{i,j}), \end{aligned} \tag{11.63}$$

or, more conveniently for digital computation,

$$\begin{aligned} p^+_{i+1,j} &= R_{i+1}(p^+_{i,j-1} - p^-_{i+1,j-1}) + p^+_{i,j-1} \\ p^-_{i,j} &= R_{i+1}(p^+_{i,j-1} - p^-_{i+1,j-1}) + p^-_{i+1,j-1}. \end{aligned} \tag{11.64}$$

The last pipe element of the line terminates in a load that is the radiation impedance of the mouth. Let A_L be the area of the last pipe element and Z_L the terminating radiation load. At the load terminals (the end of pipe A_L), the right-going and left-going pressure waves satisfy

$$\frac{p_L}{U_L} = Z_L = \frac{A_L(p^+_L + p^-_L)}{\rho c(p^+_L - p^-_L)}.$$

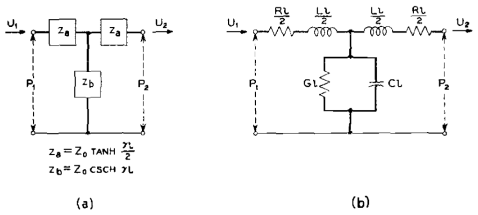


Figure 11.21: T-circuit equivalents for a length l of uniform cylindrical pipe. (a) Exact circuit, (b) first-term approximations to the impedance elements

If Z_L is written in terms of a z-transform, the reflected wave p_L^- can be obtained in terms of weighted and delayed values of p_L^+; that is, a reflection coefficient relation can be set down in which $p_L^- = p_L^+ f(z^{-1})$. The load pressure $(p_L^+ p_L^-)$ produces a mouth volume velocity U_L through Z_L, which, when differentiated, represents the radiated pressure. Formulations such as these have been used in a phoneme-driven vocal-tract synthesizer (Jr. and Lochbaum [1962a]) and in a simulation of articulatory activity (Mermelstein [1969]).

A further useful approach for digital representation of the distributed vocal tract follows the bilateral transmission line developed in Chapter 4. Once the line composed of elemental T or Π sections is set down, transient solutions of pressure and volume velcoity along the line may be obtained from difference equation approximations to the differential equations for the network. Area variations are reflected in network element changes. This approach also permits duplication of absolute acoustic impedances. For this reason it has been used in a vocal-tract synthesizer to study the acoustics of vocal-cord vibration and turbulent sound generation (Cherry [1969]).

11.5.2 Transmission-Line Analogs of the Vocal System

A different method for simulating the vocal transmission is the nonuniform electrical transmission line. The discussion in Chapter 4 indicated how the nonuniform acoustic tubes of the vocal and nasal tracts can be represented by abutting right-circular cylinders (see Fig. 3.35). The approximation to the nonuniform tract is better the more numerous the cylindrical elements.

Each cylindrical section of length l can be represented by its T-equivalent as shown in Fig. 11.21a, where $Z_a = Z_0 \tanh \gamma l/2$ and $Z_b = Z_0 \mathrm{csch} \gamma l$. A practical electrical realization of the individual T-section is obtained by taking the first terms in the series expansions of the hyperbolic quantities. For a hard-walled tube this gives $z_a \approx \frac{1}{2}(R + j\omega L)l$ and $z_b \approx 1/(G + j\omega C)l$ where the R, L, G and C are the per-unit-length acoustic parameters of the tube, as previously discussed. The resulting network is Fig. 11.21b[9].

---

[9] Section 4.7.3 derives a technique for including the effects of a yielding wall.

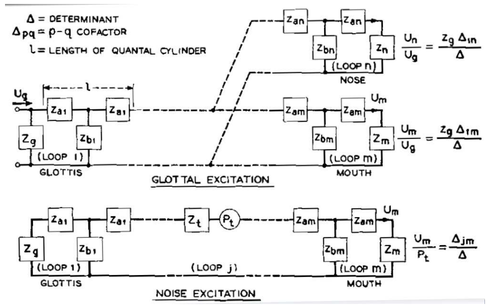


Figure 11.22: Ladder network approximations to the vocal tract. The impedance elements of the network are those shown in Fig. 11.21b

For practical realization, the characteristic impedance of the analogous electrical line may be scaled from the acoustic value by a convenient constant, i.e., $Z_{0e} = kZ_{0a}$, where the superscripts e and a distinguish electrical and acoustical quantities. For low-loss conditions, $Z_{0a} \approx \sqrt{L_a/C_a} = \rho c/A$. Since $L_a = \rho/A$ and $C_a = A/\rho c^2$, a given simulated cross-sectional area is equal $\rho c\sqrt{C_a/L_a}$. The losses R and G require knowledge of the circumference as well as the cross-sectional area of the tract [see Eq. (3.46)]. They can also be introduced into the electrical circuit and their impedances scaled after the fashion just indicated. Given the shape factor, all analogous electrical elements can be determined from the A and l data-pairs, or from area data for a given number of fixed-length cylindrical sections.

A vocal tract representation in terms of equivalent electrical sections forms the ladder networks of Fig. 11.22. The upper circuit is for glottal excitation of the tract by a volume-velocity source U_g and with internal impedance Z_g. The lower circuit is for forward fricative excitation by a pressure source P_t with internal impedance Z_t. Both circuits can be solved–at least in principle–by straightforward matrix methods. If voltage (pressure) equations are written for each circuit loop, beginning at the glottis and ending at the mouth and nose, the number of independent equations is equal the number of loops. The transmissions from glottis to mouth, from glottis to nostril, and from front noise source to mouth are, respectively,

$$\frac{U_m}{U_g} = \frac{Z_g \Delta_{1m}}{\Delta} \tag{11.65}$$

$$\begin{aligned} \frac{U_n}{U_g} &= Z_g \frac{\Delta_{1n}}{\Delta} \\ \frac{U_m}{P_t} &= \frac{\Delta_{jm}}{\Delta}, \end{aligned}$$

where Δ is the impedance determinant (characteristic equation) for the network having impedance members z_{11}, z_{12}, etc., where z_{11} is the self-impedance of loop 1, z_{12}

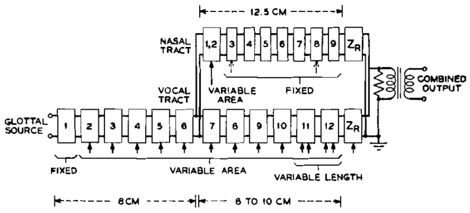


Figure 11.23: Continuously controllable transmission line analog of the vocal system. (After (Rosen [1958], Hecker [1962]))

is the mutual impedance between loops 1 and 2, etc., and Δ_{xy} is the cofactor of the x-th row and y-th column of the determinant Δ. As mentioned earlier, all the transmissions of Eq. (11.65) are minimum phase functions[10].

Several electrical transmission-line synthesizers have been constructed. The first such device consisted of 25 uniform T-sections (Dunn [1950]). Each section represented a tract length of 0.5 cm and a nominal area of 6 cm^2. A variable inductance could be inserted between any two sections to simulate the tongue construction. Another variable inductance at the mouth end of the line represented the lip constriction. Radiation from the mouth was simulated by taking the output voltage across a small series inductance. For voiced sounds, the synthesizer was excited by a high-impedance sawtooth oscillator whose fundamental frequency could be controlled. The source spectrum was adjusted to fall at about -12 dB/octave (recall Fig. 4.13). To simulate unvoiced and whispered sounds, a white noise source was applied at an appropriate point along the line.

At least two other passive line analogs, similar to Dunn's device, have been constructed(Stevens *et al.* [1953], Fant [1960]). These synthesizers incorporate network sections which can be varied independently to simulate the tract geometry in detail. At least one effort has been made to develop a continuously-controllable transmission-line analog. Continuous variation of the network elements by electronic means permits the device to synthesize connected speech (Rosen [1958], Hecker [1962]). This device utilizes saturable-core inductors and electronically-variable capacitors as the line elements. A nasal tract is also provided. The number of network sections and their control points are shown in Fig. 11.23. Control of the synthesizer can be effected either from an electronic data-storage circuit (Rosen [1958]) of from a digital computer (Dennis [1962]).

The transmission-line synthesizer has outstanding potential for directly incorporating the constraints that characterize the vocal mechanism. Success in this direction, however, depends directly upon deriving a realistic model for the area and for the dynamic motions of the vocal tract. Research on one such model has been described

---

[10] The functions are the responses of passive ladder networks. They can have zeros of transmission only for zeros of a shunt element or for poles of a series element. All these poles and zeros must lie in the left half of the complex-frequency plane.

in Section 5.7. Also, the usefulness of a transmission-line synthesizer in a complete analysis-synthesis system depends upon how accurately vocal tract area data, or its equivalent, can be derived automatically from connected speech. Some progress has been made toward analyzing speech signals in articulatory terms from which area and length numbers can be derived (see Section 5.7, Chapter 5).

Besides obvious application in a bandwidth compression system, the transmission-line synthesizer, along with other synthesis devices, has potential use as a computer output device for man-machine communication; as a stimulus generator for psychoacoustic and bioacoustic experimentation; or, as a standard sound generator for speech pathology, therapy or linguistics studies. The possibility of specifying the control functions in articulatory terms makes applications such as the latter particularly attractive.

All transmission-line synthesizers of early design have been implemented as analog network devices. Digital techniques, on the other hand, offer many advantages in stability and accuracy. One of the first digital transmission-line synthesizers was programmed on a computer in terms of the reflection coefficients at the junctions of cylindrical tube-elements (Jr. and Lochbaum [1962a]).

Another computer implementation has duplicated the bilateral properties of the transmission line by a difference-equation equivalent. Because absolute impedance relations are preserved in this formulation, it has been useful in studying the acoustic interaction between the vocal tract and the vocal cords. The same formulation has also been used as a complete synthesizer for voiced and unvoiced sounds (Flanagan and Landgraf [1968], Cherry [1969]).

Further discussion of digital representation of transmission-line synthesizers is given in Section 11.5.

TO DO: This section also needs to discuss the transmission-line and ABCD matrix synthesizers of Sondhi and Schroeter (Sondhi and Schroeter [1987]) and Qiguang Lin (Fant and Lin [1987, 1988], Lin *et al.* [1988], Lin [1990])

11.5.3 Nonlinear Simulations of the Vocal Tract System

This section will describe in more detail the nonlinear models of Pelorson (Pelorson *et al.* [1994]) and especially Huang and Levinson (Huang and Levinson [1999])

11.6 Excitation of Terminal Analog and Articulatory Synthesizers

The preceding sections have discussed simulation of the vocal transmission both from the transfer-function point of view and from the transmission-line approach. Having implemented one or the other for electrical synthesis of speech, the system must

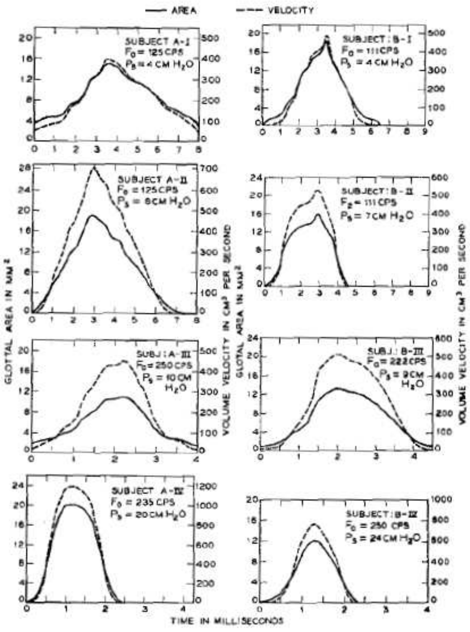


Figure 11.24: Single periods of measured glottal area and calculated volume velocity functions for two men (A and B) phonating the vowel /æ/ under four different conditions of pitch and intensity. F_0 is the fundamental frequency and P_s the sub glottal pressure. The velocity wave is computed according to the technique described in Section 4.4.2. (After (Flanagan [1958]))

be excited from signal sources analogous to those of the vocal tract. This section considers vocal source characteristics that appear relevant in synthesis.

11.6.1 Simulation of the Glottal Wave

The results of Chapter 4 suggested that the vocal cord source is approximately a high-impedance, constant volume-velocity generator. Hence, to a first-order approximation, the vocal tract and glottal source can be assumed not to interact greatly. To the extent that this is true (and we shall subsequently discuss this matter further), the source and system can be analyzed independently, and their characteristics can be simulated individually.

The shape and periodicity of the vocal cord wave can vary considerably. This is partially illustrated by the single periods of glottal area and volume-velocity waves shown in Fig. 11.24. The extent to which variability in period and shape affect speech naturalness and quality is an important research question. In many existing electrical synthesizers, the properties of the vocal cord source are approximated only in a gross form. It is customary to specify the vocal pitch as a smooth, continuous time function and to use a fixed glottal wave shape whose amplitude spectrum falls at about -

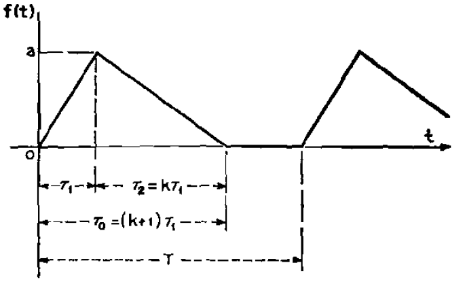


Figure 11.25: Triangular approximation to the glottal wave. The asymmetryfactor is k

12 dB/octave. In many synthesizers the source is produced by repeated impulse excitation of a fixed, spectral-shaping network. Such lack of fidelity in duplicating actual glottal characteristics undoubtedly detracts from speech naturalness and the ability to simulate a given voice.

Spectral Properties of Triangular Waves

Under some conditions of voicing (commonly, mid-range pitch and intensity), the glottal wave is roughly triangular in shape. The spectral properties of triangular waves therefore appear to have relevance to voiced excitation. They have been studied in some detail with a view toward better understanding the relations between waveform and spectrum in real glottal waves(Dunn *et al.* [1962])[11].

Fig. 11.25 shows a triangular approximation to the glottal wave. The opening time is τ_1, the closing time $\tau_2 = k\tau_1$, and the total open time $\tau_0 = (1 + k)\tau_1$. The amplitude of the wave is a and its period T. Its Laplace transform is

$$F(s) = \frac{a}{s^2}\left[\frac{1}{\tau_1} - \left(\frac{1}{\tau_1} + \frac{1}{\tau_2}\right)e^{-s\tau_1} + \frac{1}{\tau_2}e^{-s(\tau_1+\tau_2)}\right]. \tag{11.66}$$

The spectral zeros are the complex values of s which make $F(s) = 0$. Except for the $s = 0$ root, the zeros are the roots of the bracketed expression, or the roots of

$$\left[e^{-(k+1)s\tau_1} - (k + 1)e^{-s\tau_1} + k\right] = 0. \tag{11.67}$$

Because the equation is transcendental it can be solved exactly only for special values of the asymmetry constant, k. In particular, solutions are straightforward for values of k which can be expressed as the ratio of small whole numbers. In less simple cases, the roots can be obtained by numerical solution.

---

[11] It should be emphasized again that the implication here is not that the glottal pulse is a neat triangular wave, but only that this analytical simplification permits tractable and informative calculations. These data are included because they are not avaliable elsewhere.

Let

$$x = e^{-s\tau_1} = e^{-(\sigma+j\omega)\tau_1} = e^{-\sigma\tau_1}(\cos\omega\tau_1 - j\sin\omega\tau_1). \tag{11.68}$$

The (11.67) becomes

$$x^{(k+1)} - (k+1)x + k = 0. \tag{11.69}$$

When k is an integer, (11.69) will yield $(k+1)$ values of x. These can then be put into (11.68), and both $\sigma\tau_1$ and $\omega\tau_1$ found by equating real and imaginary parts in separate equations.

For integers up to $k = 5$, (11.69) can be solved by straightforward algebraic methods. In the case $k = 5$, (11.69) is a sixth degree equation in x, but a double root exists at $x = 1$, and a fourth degree equation is left when these are removed. For higher values of k, roots can be approximated by known methods.

However, k need not be an integer. Suppose only that it is a rational number (and it can always be approximated as such). Then $(k+1)$ is also rational. Let

$$k + 1 = \frac{p}{q} \tag{11.70}$$

where p and q are positive integers, and $p \geq q$, since k cannot be less than zero. Then (11.69) can be written

$$x^{\frac{p}{q}} - \frac{p}{q}x + \frac{p-q}{p} = 0. \tag{11.71}$$

Let $y = x^{1/q}$, so that (11.71) becomes

$$y^p - \frac{p}{q}y^q + \frac{p-q}{p} = 0 \tag{11.72}$$

and by (11.68)

$$y = e^{-\frac{1}{q}\sigma\tau_1}\left(\cos\frac{1}{q}\omega\tau_1 - j\sin\frac{1}{q}\omega\tau_1\right). \tag{11.73}$$

Eq. (11.72) has integer exponents, and can be solved for y. Then (11.73) can be solved for

$$\frac{1}{q}\sigma\tau_1 \quad \text{and} \quad \frac{1}{q}\omega\tau_1$$

which need only to be multiplied by p to get $\sigma\tau_0$ and $\omega\tau_0$.

The preceding methods become awkward when p is larger than 6. The following is more suitable for numerical approximation by digital computer. Equating the real and imaginary parts of (11.67) separately to zero gives the equations

$$e^{-(k+1)\sigma\tau_1}\cos(k+1)\omega\tau_1 - (k+1)e^{-\sigma\tau_1}\cos\omega\tau_1 + k = 0, \tag{11.74}$$

$$e^{-(k+1)\sigma\tau_1}\sin(k+1)\omega\tau_1 - (k+1)e^{-\sigma\tau_1}\sin\omega\tau_1 = 0, \tag{11.75}$$

Both of these equations must be satisfied by the pair of values of $\sigma\tau_1$ and $\omega\tau_1$ which represent a zero. Eq. (11.75) can be solved for $\sigma\tau_1$

$$\sigma\tau_1 = \frac{1}{k} \log \frac{\sin(k+1)\omega\tau_1}{(k+1)\sin\omega\tau_1}. \tag{11.76}$$

A series of values of $\omega\tau_1$ is put into (11.76) and the $\sigma\tau_1$ computed for each. Each pair of values is substituted into (11.74) to find those which satisfy it. The solutions can be approximated as closely as desired by choosing suitably small increments of $\omega\tau_1$, and by interpolation. A modest amount of computation time on a digital computer produces the first half-dozen roots.

Repetition and Symmetry of the Zero Pattern

Let ω be the imaginary part of a zero that (together with its real part σ) simultaneously satisfies (11.74) and (11.75). Also let k be related to integers p and q as in (11.70). Consider another imaginary part ω' such that

$$\omega'\tau_1 = 2q\pi + \omega\tau_1.$$

Then

$$\omega'\tau_0 = (k+l)\omega'\tau_l = \frac{p}{q}\omega'\tau_1 = 2p\pi + (k+1)\omega\tau_1 \tag{11.77}$$

Both sines and cosines of $\omega'\tau_1$ and $(k+l)\omega'\tau_1$ are the same as those of $\omega\tau_1$ and $(k+1)\omega\tau_1$. Hence, with no change in σ, ω' also represents a zero. The pattern of zeros between $\omega\tau_0 = 0$ and $\omega\tau_0 = 2p\pi$ will be repeated exactly in each $2p\pi$ range of $\omega\tau_0$, to infinity, with an unchanged set of σ's.

Again supposing ω is the imaginary part of a zero, let ω' be a frequency such that

$$\omega'\tau_1 = 2q\pi - \omega\tau_1. \tag{11.78}$$

Now the cosines of $\omega'\tau_1$ and $(k+l)\omega'\tau_1$ are the same as those of $\omega\tau_1$ and $(k+1)\omega\tau_1$ while the sines are both of opposite sign. Both (11.74) and (11.75) will still be satisfied, and ω' represents a zero having the same σ as that of ω. In each $2p\pi$ interval of $\omega\tau_0$, the zeros are symmetrically spaced about the center of the interval (an odd multiple of $p\pi$), each symmetrical pair having equal values of σ. There may or may not be a zero at the center of symmetry, depending upon whether p is odd or even.

Zeros of the Reversed Triangle

If $f(t)$ is the triangular wave, then $f(-t)$ is the wave reversed in time, and

$$\mathcal{L}\left[f(t)\right] = F(s)$$

and,

$$\mathcal{L}\left[f(-t)\right] = F(-s) \tag{11.79}$$

Therefore, the zeros of the reversed triangle are the negatives of those for the original triangle. Since the zeros of the original triangle occur in complex conjugate pairs, the reversed triangle has the same zeros as the original triangle, but with the signs of the real parts reversed.

Also, the asymmetry constant for the reversed triangle, is l/k, where k is the asymmetry of the original triangle.

Zeros of the Right Triangle

When $k = 0$, the triangle is right and has a transform

$$F(s) = \frac{a}{s^2 \tau_0} \left[1 - e^{-s\tau_0}(1 + s\tau_0) \right] \tag{11.80}$$

Its zeros occur for

$$(1 + s\tau_0) = e^{s\tau_0}. \tag{11.81}$$

Equating real and imaginary parts,

$$\begin{aligned} 1 + \sigma\tau_0 &= e^{\sigma\tau_0} \cos \omega\tau_0, &(11.82)\\ \omega\tau_0 &= e^{\sigma\tau_0} \sin \omega\tau_0. &(11.83) \end{aligned}$$

[Note the solution $\omega = 0$, $\sigma = 0$ cannot produce a zero because of the s^2 in the denominator of (11.80).] As before, the roots can be approximated numerically with the computer. Note that with σ and ω real, and taking only positive values of ω, $\sin \omega\tau_0$ is positive according to (11.83). Also, since $\omega\tau_0$ is larger than $\sin \omega\tau_0$, $\sigma\tau_0$ must be positive and the real parts of the zeros must be positive, or they must lie in the right half s-plane. Then by (11.82) $\cos \omega\tau_0$ is also positive which means that all zeros must occur for $\omega\tau_0$ in the first quadrant.

For $k = \infty$, the triangle is also right, but reversed in time. Its zeros are therefore the same as those for $k = 0$, but with the signs of the real parts reversed.

Loci of the Complex Zeros

Using the foregoing relations, enough zeros have been calculated to indicate the low-frequency behavior of the triangular wave. A complexfrequency plot of the zero loci–normalized in terms of $\omega\tau_0$ and $\sigma\tau_0$ and with the asymmetry k as the parameter–is shown in Fig. 11.26. In this plot the asymmetry is restricted to the range $0 \leq k \leq 1$. For $k > 1$, these loci would be mirrored in the vertical axis, that is, the signs of σ would be reversed.

For the symmetrical case ($k = 1$), the zeros are double and fall on the $j\omega$-axis at even multiples of 2π; i.e., at 4π, 8π, 12π, etc. They are represented by the small concentric circles at these points. In terms ofHz, the double zeros lie at $2/\tau_0$, $4/\tau_0$, etc., and the amplitude spectrum is $(sin^2 x/x^2)$. As k is made smaller than unity, the

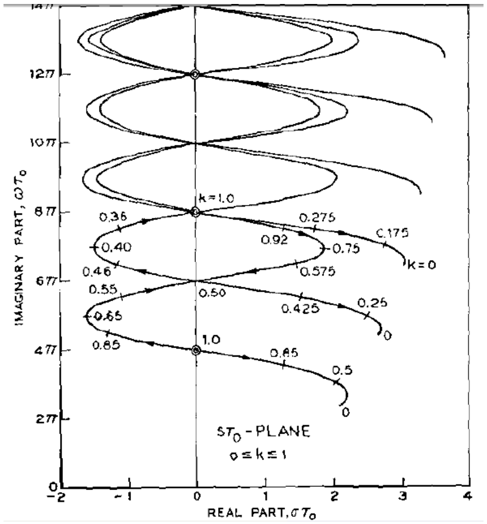


Figure 11.26: Complex frequency loci of the zeros of a triangular pulse. The s-plane is normalized in terms of $\omega\tau_0$ and $\sigma\tau_0$. The asymmetry constant k is the parameter. (After (Dunn *et al.* [1962]))

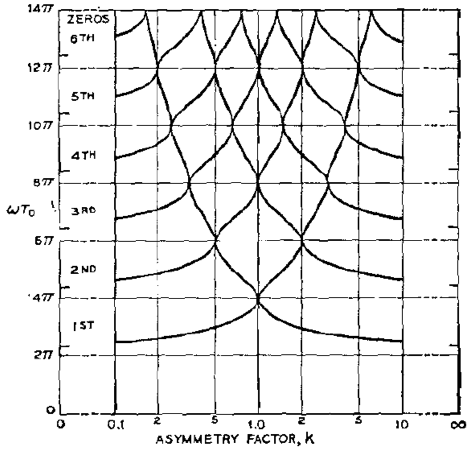


Figure 11.27: Imaginary parts of the complex zeros of a triangular pulse as a function of asymmetry. The imaginary frequency is normalized in terms of $\omega\tau_0$ and the range of asymmetry is $0 \le k \le \infty$. (After (Dunn *et al.* [1962]))

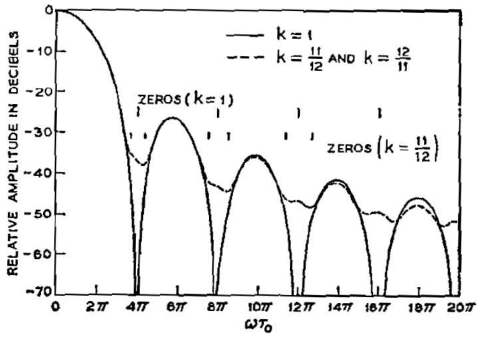


Figure 11.28: Amplitude spectra for two triangular pulses, $k = 1$ and $k = 11/12$. (After (Dunn *et al.* [1962]))

double zeros part-one moving initially into the right half plane and the other into the left. Their paths are plotted.

As the order of the zero increases, the s-plane trajectory also increases in length and complexity for a given change in k. A given reduction in k from unity causes the first zero to move into the right half plane where it remains. The same change in k may cause a higher order zero, say the sixth, to make several excursions between right and left half planes. For the first, second and third zeros, values of k from 1.0 to 0.0 are laid off along the paths. For $k = 0$, the triangle is right, with zero closing time, and all zeros have terminal positions in the right half plane. Note, too, that in the vicinity of the $j\omega$-axis, a relatively small change in symmetry results in a relatively large change in the damping of the zeros.

All imaginary-axis zeros are double and the degree of the zeros never exceeds two. This point is further emphasized in a plot of the loci of the imaginary parts of the zeros as a function of the asymmetry factor k. The pattern is shown in Fig. 11.27. It is plotted for values of k between 0.1 and 10. All points of tangency represent double $j\omega$-axis zeros. The average number of zeros is one per every 2π interval of $\omega\tau_0$. The pattern of imaginary parts is symmetrical about the $k = 1$ value, with the right and left ordinates showing the zeros of the right triangles, i.e., for $k = 0$ and $k = \infty$.

To illustrate the sensitivity of the amplitude spectrum to a specific change in the asymmetry constant, Fig. 11.28 shows amplitude spectra $|F(j\omega)|$ for two values of asymmetry, namely, $k = 1$ and $k = 11/12$ (or $12/11$). For $k = 1$ the zeros are double and are spaced atHz frequencies of $2/\tau_0$, $4/\tau_0$, $6/\tau_0$, etc. The spectrum is sin^2x/x^2 in form. A change in k to $11/12$ (or to $12/11$) causes each double zero to part, one moving into the right half plane and the other into the left. Their $j\omega$-positions are indicated by the ticks on the diagram. The increase in real parts is such as to provide the spectral “fill” indicated by the dotted curve. In this case a relatively small change in symmetry results in a relatively large spectral change.

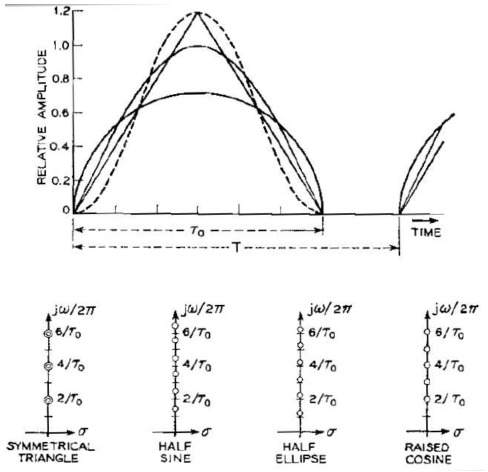


Figure 11.29: Four symmetrical approximations to the glottal pulse and their complex zeros

Other approximations to the Glottal Pulse

The preceding comments have exclusively concerned triangular approximations to the glottal wave. In reality the glottal wave can take on many forms, and it is instructive to consider the zero patterns for other simple approximations. The triangle has three points where slope is discontinuous. What, for example, might be the effect of eliminating one or more of these discontinuities by rounding or smoothing the wave?

There are several symmetrical geometries that might be considered realistic approximations to glottal waves with more rounding. Three, for example, are pulses described respectively by a half (rectified) sine wave, a half ellipse, and a raised cosine. The waveforms are plotted in the top part of Fig. 11.29. The first two have two points of discontinuous slope; the latter has none. They can be described temporally and spectrally as follows.

Half-sine wave

$$f(t) = \begin{cases} a\sin\beta t, & 0 \le t \le \frac{\pi}{\beta}, \quad \beta = \frac{\pi}{\tau_0} \\ 0, & \text{elsewhere} \end{cases} \tag{11.84}$$

$$F(\omega) = \left(\frac{\beta a}{\beta^2 - \omega^2}\right)\left(1 - e^{-j\pi\omega/\beta}\right),$$

where the zeros occur at:

$$\omega = \pm\frac{(2n+1)\pi}{\tau_0} = \pm(2n+1)\beta, \quad n = 1, 2, \ldots\,^{12}$$

---

[12] For all these symmetrical waves, the zeros lie on the $j\omega$-axis.

Half-ellipse

$$f(t) = \begin{cases} \frac{4}{\pi\tau_0}\left[1 - \left(\frac{2t}{\tau_0}\right)^2\right]^{\frac{1}{2}}, & |t| \le \tau_0/2 \\ 0, & \text{elsewhere} \end{cases} \tag{11.85}$$

$$F(\omega) = \frac{2J_1(\omega\tau_0/2)}{\omega\tau_0/2},$$

where, except for $\omega = 0$, the zeros occur at the roots of $J_1(\omega\tau_0/2)$.

Raised Cosine

$$f(t) = \begin{cases} a(1 - \cos\beta t), & 0 \le t \le \frac{2\pi}{\beta}, \quad \beta = \frac{2\pi}{\tau_0} \\ = 0, & \text{elsewhere} \end{cases} \tag{11.86}$$

$$F(\omega) = a\left[\frac{\beta^2}{j\omega(\beta^2 - \omega^2)}\right]\left[1 - e^{-j2\pi\omega/\beta}\right]$$

and the zeros occur at:

$$\omega = \pm n\beta = \pm\frac{2n\pi}{\tau_0}, \quad n = 2, 3, \ldots$$

The complex zeros for these functions are plotted in the lower part of Fig. 11.29. The plots suggest that relatively small changes in rounding and pulse shape can have appreciable influence upon the zero pattern and upon the low-frequency behavior of the glottal spectrum. Although the zeros may shift around, the average number of zeros in a given frequency interval (above a frequency of about $1/\tau_0$) still remains the same for all the waves, namely one per $1/\tau_0$Hz[13].

The Liljencrants-Fant Model

Asymptotic Density of Source Zeros

This average density of zeros also holds at high frequencies. Consider an arbitrary glottal pulse, $f(t)$, which is finite and nonzero in the interval $0 < t < \tau_0$ and zero elsewhere. Since $\int_0^\infty f(t)e^{-st}dt$ must be finite, the function can have no poles. Suppose the second derivative of $f(t)$ is bounded inside the same interval and that the slope is discontinuous at $t = 0$ and $t = \tau_0$ Except at $s = 0$, two differentiations of $f(t)$ leave the zeros the same, and produce impulses of areas $f'(0_+)$ and $f'(\tau_{0-})$ at the leading

---

[13] The spectra given here are for single pulses, that is, continuous spectra given by the Laplace or Fourier transforms of the pulses. For periodically repeated pulses, the spectra are discrete harmonic lines whose amplitudes are given by $(1/T)F(m\Omega_0)$ where $F(m\Omega_0)$ is the Fourier transform of a single pulse evaluated at the harmonic frequencies $m\Omega_0 = m2\pi/T$, $m = 1, 2, 3, \ldots$.

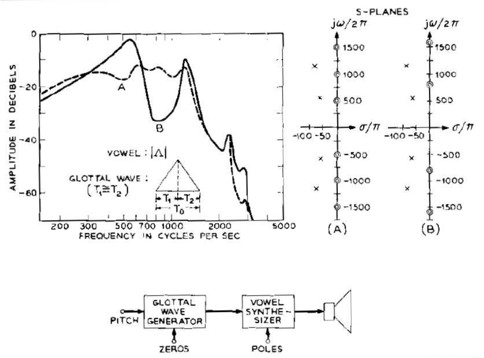


Figure 11.30: Effect of glottal zeros upon the measured spectrum of a synthetic vowel sound. (a) $\tau_0 = 4.0$ msec. (b) $\tau_0 = 2.5$ msec, (After FLANAGAN, 1961b)

and trailing edges of the pulse. The transform of the twice-differentiated pulse is therefore

$$s^2F(s) = \int_0^\infty f''(t)e^{-st}dt = f'(0_+) + f'(\tau_{0-})e^{-s\tau_0} + \int_{0_+}^{\tau_{0-}} f''(t)e^{-st}dt.$$

Since $f''(t)$ is bounded in $0 < t < \infty$, the integral of the third term must be of order $1/s$ or less. At high frequencies it becomes small compared to the first two terms and the transform is approximately

$$s^2F(s) \approx \left[f'(0_+) + f'(\tau_{0-})e^{-s\tau_0}\right],$$

with zeros at

$$s = -\frac{1}{\tau_0}\ln\left|\frac{f'(0_+)}{f'(\tau_{0-})}\right| \pm j\frac{(2n+1)\pi}{\tau_0}, \quad n = 0, 1, \ldots \tag{11.87}$$

At low frequencies, however, the zero positions may be much more irregular, as the previous computations show.

Perceptual Effects of Glottal Zeros

A relevant question concerns the effect of glottal zeros in real speech. Are they perceptually significant? Should they be taken into account in speech analysis techniques such as spectral pattern matching? Are they important for synthesizing natural speech? The complete answers to these questions are not clear and comprehensive subjective testing is needed. It is clear, however, that under particular conditions (which can sometimes be identified in sound spectrograms), a glottal zero may fall proximate to a speech formant and may alter both the spectrum and the percept.

The formant nullifying potential of a glottal zero can easily be demonstrated in synthetic speech. Fig. 11.30 shows a four-resonance vowel synthesizer circuit. The circuit is excited by an approximately symmetrical, triangular glottal wave. The amplitude spectra actually measured with a wave analyzer are shown for two conditions

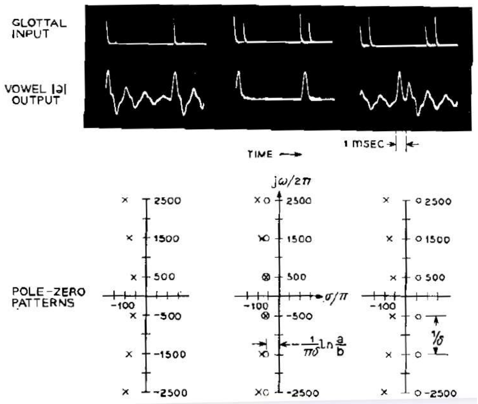


Figure 11.31: Method for manipulating source zeros to influence vowel quality. Left column, no zeros. Middle column, left-half plane zeros. Right column, right-half plane zeros. (After (Flanagan [1961]))

of open time of the glottal wave. The vowel is /ʌ/. In case (A), the open time is chosen to position the first double glottal zero near to the first formant ($\tau_0 \approx 4$ msec). In case (B), the first glottal zero is positioned between the first and second formants ($\tau_0 \approx 2.5$ msec). The relative pole-zero positions are shown for the first two formants in the s-plane diagrams. The first formant peak is clearly suppressed and flattened in the first case[14]. A significant difference in vowel quality is obvious in listening to the two conditions.

If an even more artificial situation is posed, the effect of source zeros can be made still more dramatic. For example, suppose the synthesizer is set for the vowel /ə/ which has nearly uniformly-spaced poles. Suppose also that the excitation is brief, double pulses described by $f(t) = a(t) + b(t - \delta)$, where $a(t)$ and $b(t)$ are impulses with areas a and b, respectively. The frequency transform of $f(t)$ is $F(s) = (a+be^{-s\delta})$ which has zeros at

$$s = \left[-\frac{1}{\delta}\ln\frac{a}{b} \pm j\frac{(2n+1)\pi}{\delta}\right], \quad n = 0, 1, \ldots \tag{11.88}$$

That is, this excitation produces the same zero pattern as the asymptotic high frequency spacing given in Eq. (11.87). By suitable choice of a/b and δ, the source zeros can be placed near the formants. Three different excitation conditions (including a single pulse) are shown in three columns in Fig. 11.31. The input excitation and the resulting synthetic sound waveforms are also shown. In the first case the vowel is clearly heard and identified as /ə/. In the second and third cases, the vowel quality and color are substantially altered. Cases 2 and 3 differ very little perceptually, although the sound waveforms are greatly different. From the perceptual standpoint

---

[14] In neither case does the measured amplitude spectrum go to zero at the frequency of the zeros. The laboratory-generated glottal wave was not precisely symmetrical and its zeros did not lie exactly on the $j\omega$-axis.

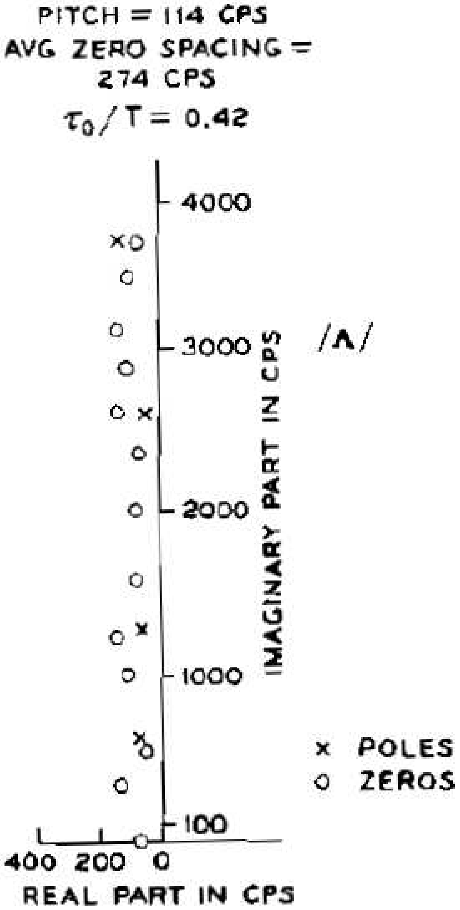


Figure 11.32: Best fitting pole-zero model for the spectrum of a single pitch period of a natural vowel sound. (After (Mathews and Walker [1962]))

there appears to be a relatively narrow vertical strip, centered about the $j\omega$-axis, in which a glottal zero has the potential for substantially influencing the percept[15]. The double pulse excitation provides a simple means for manipulating the zero pattern for subjective testing. Also, to a very crude approximation, it is somewhat similar to the phenomenon of diplophonia (Smith [1958]).

As emphasized earlier in this section, the perceptual importance of glottal wave detail and of source zeros has not been thoroughly established. At least one speech analysis procedure, however, has taken glottal zeros into account to obtain more precise spectral analyses (Mathews and Walker [1962]). A pole-zero model, with an average zero density of one per $1/\tau_0$Hz, is fitted in a weighted-least-square sense to real speech spectra (see Section 5.5.1). A typical pole-zero fit to the spectrum of a

---

[15] Symmetric glottal pulses produce zeros on the $j\omega$-axis, as described in the preceding discussion. In natural speech this region appears to be largely avoided through vocal-cord adjustments.

single pitch period of a natural vowels is shown in Fig. 11.32. The analysis procedure does not discriminate between right and left half-plane zeros, and all zeros are plotted in the left half-plane. An open time of the glottal wave of about 0.4 times the pitch period is suggested by the result.

Whether the precise positions of source zeros are perceptually significant remains a question for additional study. Only their influence on over-all spectral balance and gross shape may be the important factor. The vocal excitation may vary in waveform so rapidly in connected speech that the zero pattern is not stationary long enough to influence the percept. A speaker also might adjust his glottal wave by auditory feedback to minimize unwanted suppression of formant frequencies.

One experiment leads to the view that the glottal wave can be represented by a fixed analytical form, and that period-to-period irregularities in the pitch function can be smoothed out(Rosenberg [1971a]). Natural speech was analyzed pitch-synchronously. Pitch, formant frequencies and an inverse-filter approximation to the glottal wave were determined for each period. The glottal wave shape was "cartoonized" and characterized by fixed, smooth, analytical functions, whose glottis-open times depended only upon pitch period[16]. Using the analyzed pitch and formant data, the speech was synthesized with this artificial characterization of the glottal wave. Listening tests were then conducted.

Subjects preferred asymmetric wave characterizations with one slope discontinuity (corresponding to cord closure) and with opening and closing times equal to 40% and 16% of the pitch period. The subjects were relatively insensitive to variations in the precise shape and open-close times. Very small opening or closing times, and approximately equal opening and closing times were clearly not preferred. The latter, as discussed above, leads to spectral zeros near the jill-axis. The results also demonstrated that elimination of fine temporal detail in the glottal wave shape does not degrade speech quality. These results appear consistent with data on factors found important in formant-synthesized speech (Holmes [1961]).

Another experiment, using the same analysis techniques, determined the amount of averaging of pitch and formant data that is perceptually tolerable in synthetic speech (Rosenberg [1971a]). In the vowel portions of syllables in connected speech, averaging over as much as four to eight pitch periods did not degrade quality. This averaging completely eliminated fine detail (period-to-period fluctuations) in the pitch and formant data. Longer averaging, which modified the underlying pitch and formant trajectories, did definitely impair quality.

Acoustic interaction between the vocal cords and vocal tract contributes some temporal details to the glottal volume flow waveform. This interaction also influences the temporal variation of voice pitch. These experiments suggest that the fine structure, both in wave shape and in pitch-period variation, is not perceptually significant, but that variations in values averaged over several pitch periods are significant.

---

[16] Note that the spectral zeros of such waves vary in frequency position as the fundamental frequency changes. Only for monotone pitch are the spectral zeros constant in position.

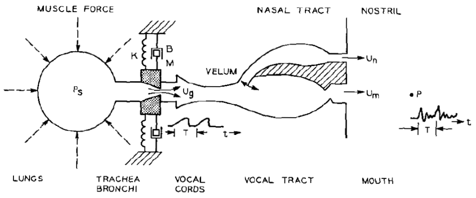


Figure 11.33: Schematic diagram of the human vocal mechanism. (After (Flanagan *et al.* [1970]))

One point should perhaps be emphasized in considering inverse-filter estimates of glottal wave shape. The fundamental hypothesis is that the source and system are linearly separable, and that the acoustic properties of each can be uniquely assigned. The glottal wave is usually obtained from the inverse filter according to some criterion such as minimum ripple. Such criteria are completely acceptable within the frame of a particular analysis model; that is, by specifically defining non-interactive source and system. On the other hand, if the objective is an accurate estimate of the real glottal flow, which in fact may have substantial ripple and detail, then the inverse-filter method can be treacherous. Properties justly belonging to the source might be assigned to the system, and vice versa.

Model for Voiced Excitation

Increased insight into vocal-tract excitation can be obtained from efforts to model the acoustics of human sound generation (Flanagan and Landgraf [1968], Cherry [1957], Ishizaka and Flanagan [1972b], Flanagan [1969]). Such efforts are also directly relevant to speech synthesis by vocal-tract simulation.

Following the analyses of Chapter 4, voiced excitation of the vocal system can be represented as in Fig. 11.33. The lungs are represented by the air reservoir at the left. The force of the rib-cage muscles raises the air in the lungs to subglottal pressure P_s. This pressure expells a flow of air with volume velocity U_g through the glottal orifice and produces a local Bernoulli pressure. The vocal cords are represented as a symmetric mechanical oscillator, composed of mass M, spring K and viscous damping, B. The cord oscillator is actuated by a function of the subglottal pressure and the glottal Bernoulli pressure. The sketched waveform illustrates the pulsive form of the U_g flow during voiced sounds. The vocal tract and nasal tract are shown as tubes whose cross-sectional areas change with distance. The acoustic volume velocities at the mouth and nostrils are U_m and U_n respectively. The sound pressure P in front of the mouth is approximately a linear superposition of the time derivatives $\dot{U}_m$ and $\dot{U}_n$.

Following the transmission-line relations derived in Chapter 4, the acoustic system of Fig. 11.33 can be approximated by the network of Fig. 11.34. The lung volume is represented by a capacity and loss whose sizes depend upon the state of inflation. The

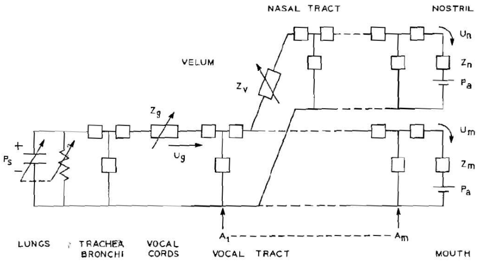


Figure 11.34: Network representation of the vocal system

lungs are connected to the vocal cords by the trachea and bronchi tubes, represented in the figure as a single T-section. The impedance of the vocal cords Z_g is both time-varying and dependent upon the glottal volume velocity U_g. The vocal tract is approximated as a cascade of T-sections in which the element impedances are determined by the cross-sectional areas $A_1 \ldots A_n$. The value of N is determined by the precision to which the area variation is to be represented. The line is terminated in a radiation load at the mouth Z_m, which is taken as the radiation impedance of a circular piston in a plane baffle. U_m is the mouth current and, for simulation of d.c. quantities, a battery P_a represents atmospheric pressure.

The nasal tract is coupled by the variable velar impedance Z_v. The nasal tract is essentially fixed in shape, and the nostril current U_n flows through the radiation impedance Z_n.

This formulation of the vocal system can simulate respiration as well as phonation. The glottis is opened (Z_g is reduced), the rib cage muscles enlarge the lung capacitor (volume), and the atmospheric pressure forces a charge of air through the tract and onto the capacitor. The glottis is then clenched and increased in impedance; the rib cage muscles contract, raising the voltage (pressure) across the lung capacity, and force out a flow of air. Under proper conditions, the vocal-cord oscillator is set into stable vibration, and the network is excited by periodic pulses of volume velocity. The lung pressure, cord parameters, velar coupling, and vocal tract area all vary with time during an utterance. A differenceequation specification of the network, with these variable coefficients, permits calculation of the Nyquist samples of all pressures and volume velocities, including the output sound pressure (FLANAGAN and LANDGRAF).

To simplify computation and to focus attention on the properties of the vocal-cord oscillator, the cords can be represented by a single moveable mass as shown in Fig. 11.35 (it being understood that the normal movement is bilaterally symmetric with the opposing cord-mass experiencing identical displacement). The cords have thickness d and length l. Vertical displacement x, of the mass changes the glottal area A_g, and varies the flow U_g. At rest, the glottal opening has the phonation neutral area A_{g0}.

The mechanical oscillator is forced by a function of the subglottal pressure and

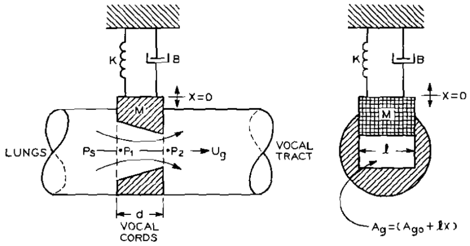


Figure 11.35: Acoustic oscillator model of the vocal cords. (After (Flanagan and Landgraf [1968]))

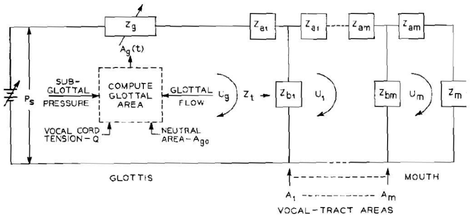


Figure 11.36: Simplified network of the vocal system for voiced sounds. (After (Flanagan and Landgraf [1968]))

the Bernoulli pressure in the orifice. The Bernoulli pressure is dependent upon U_g^2 which, in turn, is conditioned by the nonlinear, time-varying acoustic impedance of the glottal opening. In qualitative terms, the operation is as follows: the cords are set to the neutral or rest area, and the subglottal pressure applied. As the flow builds up, so does the negative Bernoulli pressure. The latter draws the mass down to interrupt the flow. As the flow diminishes, so does the Bernoulli pressure, and the spring acts to retrieve the mass. Under appropriate conditions, stable oscillation results.

The undamped natural frequency of the oscillator is proportional to $(K/M)^{\frac{1}{2}}$. It is convenient to define a vocal-cord tension parameter Q, which scales the natural frequency by multiplying the stiffness and dividing the mass. This is loosely analogous to the physiological tensing of the cords, which stiffens them and reduces their distributed mass. Since the trachea-bronchi impedance is relatively low (compared to that of the glottal orifice), and since the large lung volume is maintained at nearly constant pressure over short durations, a source of constant pressure can approximate the subglottal system. For voiced, non-nasal sounds, this modification to the network is shown in Fig. 11.36.

The acoustic impedance of the glottal orifice is characterized by two loss elements, R_v and R_k, and an inertance, L_g[17]. The values of these impedances depend upon the time-varying glottal area $A_g(t)$. In addition, R_k is dependent upon $|U_g|$. The glottal area is linked to P_s and to U_g through the differential equation that describes the

---

[17] See Section 4.4.2

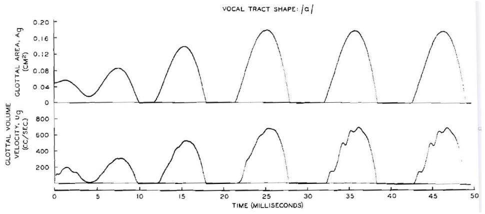


Figure 11.37: Glottal area and acoustic volume velocity functions computed from the vocal-cord model. Voicing is initiated at $t = 0$

vocal-cord motion and its forcing function. The values of the tension parameter Q and of the phonation-neutral area A_{g0} are also introduced into this equation. In other words, the dashed box of Fig. 11.36 represents iterative solutions to the differential equation for the system described in Fig. 11.35.

This continuous system can be represented by $(m+2)$ differential equations, which, in turn, can be approximated by difference equations. These difference equations are programmed for simultaneous solution on a digital computer. The program accepts as input data time-varying samples of the subglottal pressure P_s, the cord tension Q, the neutral area A_{g0} and the vocal tract areas $(A_1 \ldots A_m)$, and it computes sampled values of all volume velocities, including the glottal flow and mouth output. The resulting functions can be digital-to-analog converted and fed to a display scope or loudspeaker. A typical glottal area and volume velocity, plotted by the computer for a vocal-tract shape corresponding to the vowel /ɑ/, is shown in Fig. 11.37. This figure shows the initial 50 msec of voicing.

The top curve is the glottal area result, and the lower curve the glottal flow. The calculation is for a subglottal pressure of 8 cm H_20, a neutral area of 0.05 cm^2 and a tension value that places the cord oscillation in the pitch range of a man. One notices that by about the fourth period a steady state is achieved. One sees, in this case, irregularities in the glottal flow that are caused by acoustic interaction at the first formant frequency of the tract. One also notices that this temporal detail in the volume flow is not noticeably reflected in the mechanical behavior, that is in the area wave.

The behavior of the vocal-cord model over a range of glottal conditions suggests that it duplicates many of the features of human speech. Furthermore, the parameters necessary for complete synthesis of foiced sounds are now reduced to the articulatory quantities: tract areas, $A_1 \ldots A_m$; subglottal pressure, P_s; cord tension, Q; and phonation neutral area A_{g0}. A spectrogram of the audible output for a linear transition in vocal tract shape from the vowel /i/ to the vowel /ɑ/ is shown in Fig. 11.38. The glottal conditions in this case are constant and are: $P_s = 8$ cn H_20, $A_{g0} = 0.05$ cm^2 and $Q = 2.0$. The resulting fundamental frequency of these sounds is not only a function of the glottal parameters, but also of the tract shape; this is, a function of the acoustic loading that the tract presents to the vocal cords. The spectral sections indicate realistic formant and pitch values.

The single-mass model of the cords, because of its simplicity and because it produces many features of human speech, is attractive for use in transmission-line synthesizers. It does not, however, represent physiological details such as phase differences between the upper and lower edges of the real cords. Also, its acoustic interaction with the vocal system is critically dependent upon the relations assumed for intraglottal pressure distribution. (The values determined by VAN DEN BERG were used in the above simulations.) If a more detailed simulation of the physiology and the acoustic interaction is needed, the acoustic oscillator concept can be extended to multiple mass-spring representations of the cord mass and compliance (Flanagan and Landgraf [1968]). A two-mass oscillator, stiffness coupled, has been found to represent with additional accuracy the real-cord behavior (Ishizaka and Matsudaira [1968], Dudgeon [1970], Ishizaka and Flanagan [1972b]). Continuing research aims to use this additional sophistication in synthesis.

11.6.2 Simulation of Unvoiced Excitation

The discussion of Chapter 4 pointed out the uncertainties in our present knowledge of unvoiced sources of excitation. Existing measurements (Heinz [1958]) suggest that the source for voiceless continuants (fricatives) has a relatively flat spectrum in the mid-audio frequency range, and that the source impedance is largely resistive. In electrical synthesis of speech, these sounds are commonly generated by having a broadband random noise source excite the simulated vocal resonances. Stop sounds, on the other hand, are often produced by a transient excitation of the resonators, either with electrical pulses or brief noise bursts. Voiced fricatives, since they are excited by pitch-synchronous noise bursts in the real vocal tract, can be simulated by multiplying the simulated glottal wave with an on-going broadband noise signal.

With slight modification, and with no additional control data, the system of Fig. 11.36 can be arranged to include fricative and stop excitation. Fricative excitation is generated by turbulent air flow at a constriction, and stop excitation is produced by making a complete closure, building up pressure and abruptly releasing it. The stop release is frequently followed by a noise excitation owing to turbulence generated at the constriction after the release.

Experimental measurements indicate that the noise sound pressure generated by turbulence is proportional to the square of the Reynolds number for the flow (see Section 4.5). To the extent that a one-dimensional wave treatment is valid, the noise sound pressure can be taken as proportional to the square of the volume velocity and inversely proportional to the constriction area. Measurements also suggest that the noise source is spatially distributed, but generally can be located at, or immediately downstream of the closure. Its internal impedance is primarily resistive, and it excites the vocal system as a series pressure source. Its spectrum is broadly peaked in the midaudio range and falls off at low and high frequencies (Heinz [1958]).

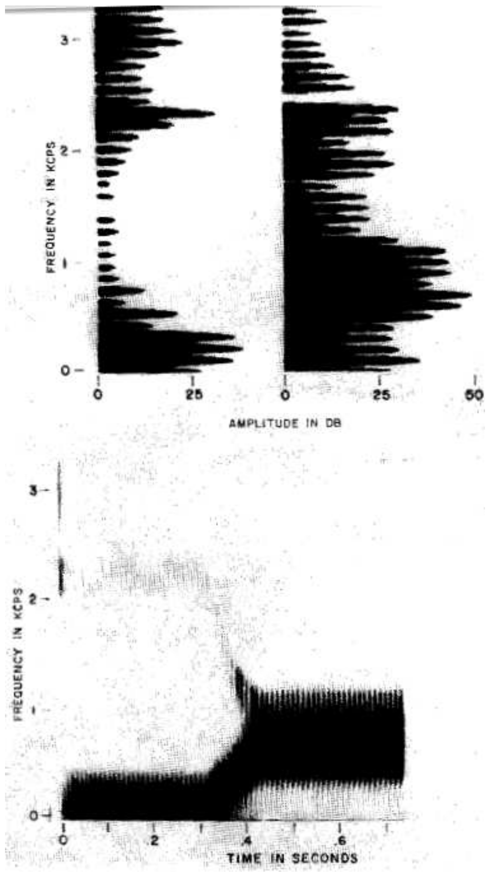


Figure 11.38: Spectrogram of a vowel-vowel transition synthesized from the cord oscillator and vocal tract model. The output corresponds to a linear transition from the vowel /i/ to the vowel /ɑ/. Amplitude sections are shown for the central portion of each vowel

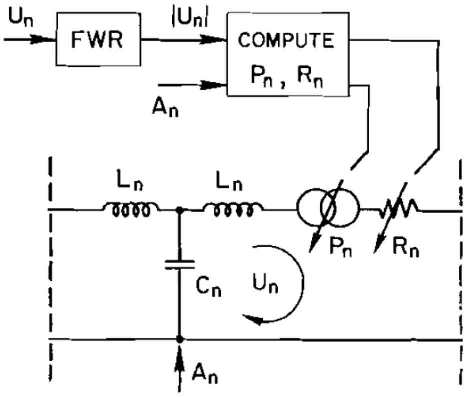


Figure 11.39: Modification of network elements for simulating the properties of turbulent flow in the vocal tract. (After (Cherry [1969])

The transmission-line vocal tract, including the vocal-cord model, can be modified to approximate the nonlinearities ofturbulent flow (FLANAGAN and CHERRY). Fig. 11.39 shows a single section of the transmission line so modified. A series noise source P_n, with internal resistance R_n is introduced into each section of the line. The area of the section is A_n and the volume current circulating in the right branch is U_n. The level of the noise source and the value of its internal resistance are functions of U_n and A_n. The noise source is modulated in amplitude by a function proportional to the squared Reynolds number; namely, U_n^2/A_n. The source resistance is a flow-dependent loss similar to the glottal resistance. To first order, it is proportional to $|U_n|$ and inversely proportional to A_n^2. The diagram indicates that these quantities are used to determine P_n and R_n. In the computer simulation they are calculated on a sample-by-sample basis.

By continually noting the magnitudes of the volume currents in each section, and knowing the corresponding areas, the synthesizer detects conditions suitable to turbulent flow. Noise excitation and loss are therefore introduced automatically at any constriction. Small constrictions and low Reynolds numbers produce inaudible noise. The square-law dependence of P_n upon U_n has the perceptual effect of a noise threshold. (A real threshold switch can be used on the noise source, if desired.) The original control data, namely, vocal-tract shape, subglottal pressure, neutral area and cord tension, in effect, determine the place of the constriction and the loss and noise introduced there.

The P_n source is taken as Gaussian noise, bandpassed between 500 and 4000 Hz. Also, to ensure stability, the volume flow U_n is lowpass filtered to 500 Hz before it modulates the noise source. In other words, the noise is produced by the low-frequency components of U_n including the *dc* flow.

This noise excitation works equally well for voiced and unvoiced sounds. The operation for voiced fricatives includes all features of the formulation, and is a good vehicle for illustration. For example, consider what happens in a vowel when the constriction is made substantially smaller than normal, giving rise to conditions favorable for turbulent flow. Since we have already shown results for the vowel /ɑ/, consider the same vowel with the constriction narrowed. (This configuration is not proposed as a realistic English sound, but merely to illustrate the effect of tightening the vowel constriction.) The situation is shown in Fig. 11.40. All glottal conditions are the same as before, but the constriction is narrowed to less than half the normal vowel constriction (namely, to 0.3 cm^2).

The top trace shows the glottal area, and one notices that it settles to a periodic oscillation in about four periods. The final pitch here is somewhat less than that in Fig. 11.37 because the acoustic load is different. The second trace from the top shows the glottal flow. The glottal flow is about the same in peak value as before and is conditioned primarily by the glottal impedance and not by the tract constriction. At about the third period, noise that has been produced at the constriction by the flow buildup has propagated back to the glottis and influences the U_g flow. Note, too,

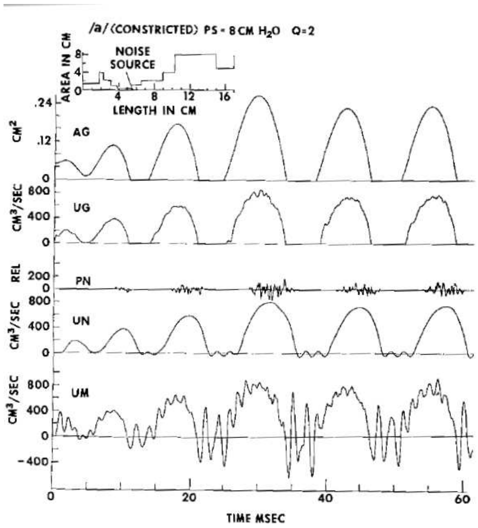


Figure 11.40: Waveforms of vocal functions. The functions are calculated for a voiced fricative articulation corresponding to the constricted vowel /ɑ/. (After (Cherry [1969]))

that noise influence on the mechanical oscillator motion (i.e., the area function) is negligible.

The third trace shows the output of the noise source at the constriction. This output is proportional to the constriction current squared, divided by the constriction area. The fourth trace shows the low-passed constriction current that produces the noise. One sees that the tendency is for the noise to be generated in pitch-synchronous bursts, corresponding to the pulses of glottal volume flow. The result is a combined excitation in which the voicing and noise signals are multiplicatively related, as they are in the human.

The final trace is the volume flow at the mouth, and one can notice noise perturbations in the waveform. Note, too, that the epoch of greatest formant excitation corresponds to the falling phase of the glottal flow. A spectrogram of this audible output is compared with that for a normal /ɑ/ in Fig. 11.41. The normal vowel is shown on the left; the constricted vowel on the right. Note in the constricted, noisy /ɑ/ that: (1) the first formant has been lowered in frequency, (2) the fundamental frequency is slightly lower, and (3) pitch-synchronous noise excitation is clearly evident, particularly at the higher frequencies.

Voiceless sounds are produced in this cord-tract model simply by setting the neutral area of the vocal cords (A_{g0}) to a relatively large value, for example 1 cm^2. As this is done, the Bernoulli pressure in the glottal orifice diminishes, the oscillations of the vocal cords decay, and the cord displacement assumes a steady large value. Control of A_{g0} therefore corresponds to the voiced-voiceless distinction in the model. Measurements on real speech suggest this kind of effect in passing from voiced to

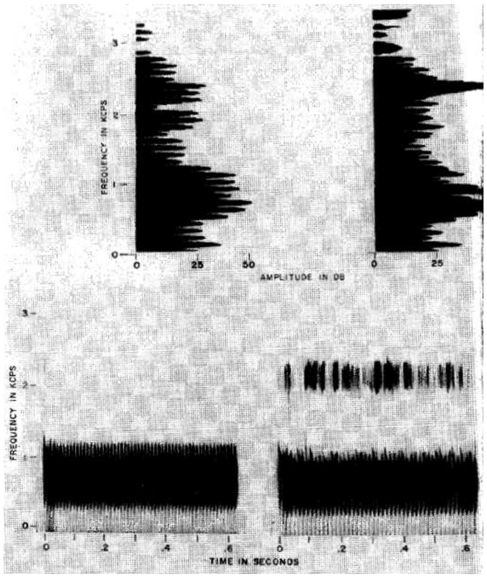


Figure 11.41: Sound spectrograms of the synthesized output for a normal vowel /ɑ/ (left) and the constricted /ɑ/ shown in Fig. 11.40 (right). Amplitude sections are shown for the central portion of each vowel

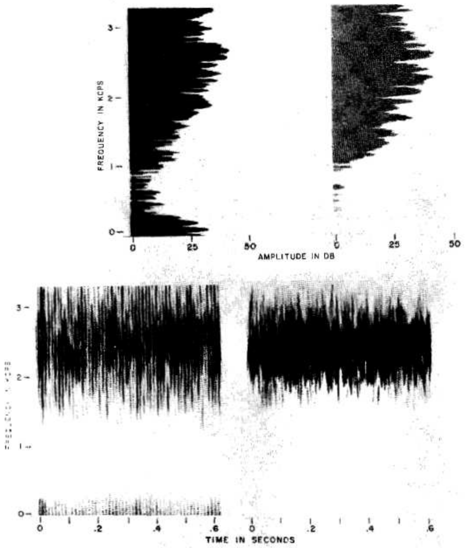


Figure 11.42: Spectrograms for the voiced-voiceless cognates /ʒ/ and /ʃ/. Amplitude sections are shown for the central portion of each sound

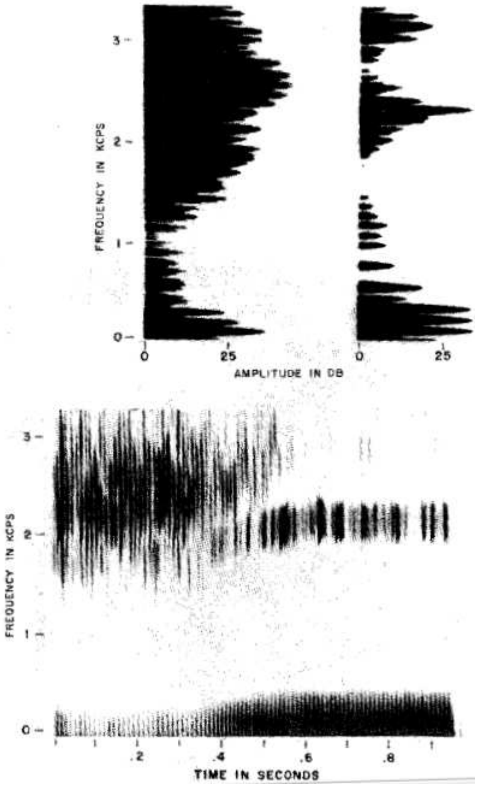


Figure 11.43: Sound spectrogram for the synthesized syllable /ʒi/. Amplitude sections are shown for the central portion of each sound. (After (Cherry [1969]))

voiceless sounds (Sawashima [1968]). Corresponding exactly to this change, spectrograms of the audible output for the voiced-voiceless cognates /ʒ/ and /ʃ/ are compared in Fig. 11.42. The vocal-tract shape is the same for both sounds. One sees a pronounced voice bar in /ʒ/ (left spectrogram) that, of course, is absent in /ʃ/ (right spectrogram). The eigenfrequencies of the two systems are similar but not exactly the same because of the difference in glottal termination. Lower resonances are not strongly evident in the /ʃ/ output, because its transmission function, from point of constriction to mouth, exhibits low-frequency zeros.

The dynamics of continuous synthesis can be illustrated by a consonant-vowel syllable. Fig. 11.43 shows the syllable /ʒi/ synthesized by the system. In this case, the subglottal pressure, the phonation neutral area and cord tension are held constant and the vocal tract area function is changed linearly from the configuration for /ʒ/ to that for /i/. Heavy noise excitation is apparent during the tightly constricted /ʒ/, and the noise diminishes as the articulation shifts to /i/. Also in this case, the high front vowel /i/ is characterized by a relatively tight constriction and a small amount of noise excitation continues in the /i/. This same effect can be seen in human speech.

This model also appears capable of treating sounds such as glottal stops and the glottal aspiration that accompanies /h/. In the former, the tension control can cause an abrupt glottal closure and cessation of voicing. Restoration to a normal tension and quiescent glottal opening permits voicing to again be initiated. In the latter, the flow velocity and area at the glottis can be monitored just as is done along the tract. When conditions suitable for turbulence exist, a noise excitation can be introduced at the glottal location. Note, too, that the central parameters lor the voiceless synthesis are exactly the same as for voiced synthesis; namely, $A_1 \dots A_m$, P_s, Q and A_{g0}. No additional control data are necessary. Place and intensity of voiceless excitation are deduced from these data.

Although crude in its representations of acoustic nonlinearities, this model for voiced and voiceless excitation appears to give realistic results. It is applicable to speech synthesis by vocal tract simulation and it provides a point of departure for further study of sound generation in the human vocal tract.

11.7 Vocal Radiation Factors

Electrical synthesizers usually attempt to account for source characteristics, vocal transmission and mouth-nostril radiation. In a terminal-analog synthesizer, the radiation factor is essentially the functional relation between sound pressure at a point in space and the acoustic volume current passing the radiating port. A transmission-line analog, on the other hand, should be terminated in an impedance actually analogous to the acoustic load on the radiating port. For most speech frequencies, the latter is adequately approximated as the radiation load on a piston in a large baffle (see Section 4.2). The former, for frequencies less than about 4000Hz, is adequately approximated by the relations for a small spherical source (see Section 4.3). That is, the pressure at a point in front of the speaker is proportional to the time derivative of the mouth volume velocity.

To simulate the radiation function in terminal-analog synthesizers, a frequency equalization proportional to frequency (i.e., a 6 dB/oct boost) can be applied to the vocal transmission function. Similarly in the transmission-line analog, the current through the radiation load can be differentiated to represent the output sound pressure (alternatively, the voltage directly across the radiation load can be taken as the pressure). Because the mouth and nostrils are spatially proximate (a fraction of a wavelength apart at the lower speech frequencies), the effect of simultaneous radiation from these two points can be approximated by linearly superposing their volume currents or sound pressures.

11.8 Homework

Problem 11.1

In this problem, you will explore the relationship between the LPC reflection-line synthesis filter and the transmission-line model of the vocal tract.

Recall that the reflection-line filter iteratively removes the redundancy from the speech signal $S(z)$ in order to calculate the forward prediction error, $E^{(i)}(z)$, and the backward prediction error, $B^{(i)}(z)$:

$$E^{(i)}(z) = E^{(i-1)}(z) - k_i z^{-1} B^{(i-1)}(z) \tag{11.89}$$

$$B^{(i)}(z) = z^{-1} B^{(i-1)}(z) - k_i E^{(i-1)}(z) \tag{11.90}$$

$$E^{(0)}(z) = B^{(0)}(z) = S(z) \tag{11.91}$$

a. Suppose you are given as inputs the forward error of order (i) and the backward error of order $(i-1)$, $E^{(i)}(z)$ and $B^{(i-1)}(z)$, and you are asked to synthesize $E^{(i-1)}(z)$ and $B^{(i)}(z)$. By re-arranging the equations above, devise a filter structure to accomplish this. Draw the filter structure.

b. Suppose you are given only the forward error of order (2), $E^{(2)}(z)$, and asked to synthesize the speech signal $S(z)$ and the backward error $B^{(2)}(z)$. Devise a lattice filter structure to accomplish this.

c. Fig. 2 shows a section of a digital concatenated tube model, similar to the model shown in R&S Fig. 3.40, but with a delay on the backward arc rather than the forward arc. Suppose that the left-hand nodes are the (i)th order LPC prediction errors, $A(z) = E^{(i)}(z)$ and $B(z) = B^{(i)}(z)$. Show that the right-hand nodes are $C(z) = \gamma E^{(i-1)}(z)$ and $D(z) = \gamma B^{(i-1)}(z)$, for some constant γ. What is the value of γ?

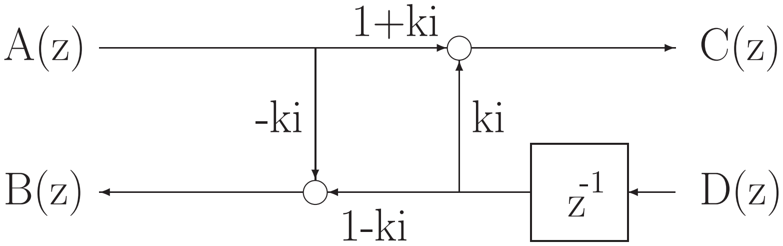


Figure 2: Section of a digital concatenated tube model.

d. Find values of the reflection coefficients r_G, r_1, r_2, and r_L such that the transfer function of the concatenated tube model in Fig. 3.40c of your text is

$$\frac{U_l(z)}{U_g(z)} = z^{-3/2} G \frac{S(z)}{E^{(2)}(z)} \tag{11.92}$$

where $S(z)$ and $E^{(2)}(z)$ are as defined in part (b) of this problem, and G is a constant.

e. In the concatenated tube model of part (d), what is the acoustic impedance at the lips? at the glottis? (you may express one or both impedances in terms of the cross-sectional tube areas, if necessary).

f. The concatenated tube elements in your model of part (d) are lossless elements, and yet the impulse response of the system decays gradually over time, implying that at least one element in the system must be lossy. Where are the losses occurring? In a real vocal tract, where else would losses occur?

Chapter 12

Speech Coding

The discussions in Chapters 4 and 7 considered the basic physics of the mechanisms for speech production and hearing. The topics of Chapters 5, 11, 9, and 10 set forward certain principles relating to the analysis, artificial generation, and perception of speech. The present and final chapter proposes to indicate how the foregoing results, in combination, may be applied to the efficient transmission of speech.

Efficient communication suggests transmission of the minimum information necessary to specify a speech event and to evoke a desired response. Implicit is the notion that the message ensemble contains only the sounds of human speech. No other signals are relevant. The basic problem is to design a system so that it transmits with maximum efficiency only the perceptually significant information of speech.

One approach to the goal is to determine the physical characteristics of speech production, perception and language and to incorporate these characteristics into the transmission system. As such, they represent information that need not be transmitted. Ideally, the characteristics are described by a few independent parameters, and these parameters serve as the information-bearing signals. Transmission systems in which a conscious effort is made to exploit these factors are generally referred to as *analysis-synthesis* systems.

In the ideal analysis-synthesis system, the analysis and synthesis procedures are presumably accurate models of human speech production. To the extent this is true, the resulting signal is coded and synthesized in a distortionless form. Additional economies in transmission can accrue from perceptual and linguistic factors. The pure analysis-synthesis system therefore has the greatest potential for bandsaving, and its analysis and synthesis processings typically require complex operations.

In contrast, other transmission systems aim for modest or little band savings, with terminal apparatus which is simple and inexpensive. Such systems typically exploit fewer properties of speech, hearing and language than do the pure analysis-synthesis systems. Nevertheless, they are of considerable interest and importance, and their potential applications range from mobile radio and scatter links to various commercial wire circuits. Although emphasis in this chapter is given to analysis-

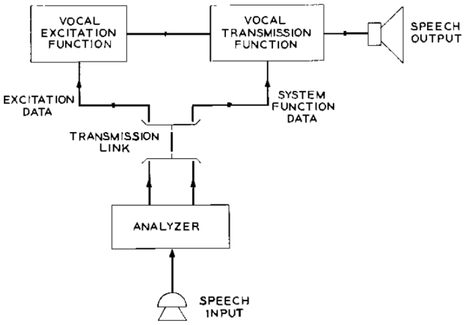


Figure 12.1: Source-system representation of speech production

synthesis techniques, systems of the latter category are also brought in for discussion, especially in the context of digital coding and transmission.

The results of Chapter 4 and 7 showed that speech signals can be described in terms of the properties of the signal-producing mechanism, that is, the vocal tract and its excitation. This characterization suggests important possibilities for efficient encoding of speech. In fact, it forms the common basis for a large class of bandwidth-compression systems. The idea is schemetized in Fig. 12.1. Three operations are involved. First, the automatic analysis of the signal into quantities that describe the vocal excitation and mode structure; second, the multiplexing and transmission of these parameters; and finally, the reconstruction of the original signal from them.

In a parallel manner, the discussion in Chapter 7 suggested that the ear performs a kind of short-time frequency analysis at its periphery. The analysis includes a mechanical filtering, the equivalent of a rectification, and a neural encoding which–apparently at an early stage–involves an integration. In such a process, certain details of the original speech wave are lost and are not perceptually significant. Presumably a transmission system might also discard this information without noticeably influencing the preceived signal. It might thereby effect an economy in requisite channel capacity.

In a similar fashion, other aspects of the signal–for example, the sequential constraints on the sounds of a given language, or the natural pauses in connected speech–might be used to advantage. In short, practically all aspects of speech production, hearing and language have relevance to analysis-synthesis telephony. The following sections propose to discuss complete analysis-synthesis systems, and a number of these factors will be put in evidence.

12.1 Assessment of Speech Perceptual Quality

Deciding on an appropriate measurement of quality is one of the most difficult aspects of speech coder design, and is an area of current research and standardization. Early military speech coders were judged according to only one criterion: intelligibility. With the advent of consumer-grade speech coders, intelligibility is no longer a sufficient condition for speech coder acceptability. Consumers want speech that sounds "natural." A large number of subjective and objective measures have been developed to quantify "naturalness," but it must be stressed that any scalar measurement of "naturalness" is an oversimplification. "Naturalness" is a multivariate quantity, including such factors as the metallic vs. breathy quality of speech, the presence of noise, the color of the noise (narrowband noise tends to be more annoying than wideband noise, but the parameters which predict "annoyance"are not well understood), the presence of unnatural spectral envelope modulations (e.g. flutter noise), the absence of natural spectral envelope modulations, etc.

12.1.1 Psychophysical Measures of Speech Quality (Subjective Tests)

The final judgment of speech coder quality is the judgment made by human listeners: if consumers (and reviewers) like the way the product sounds, then the speech coder is a success. The reaction of consumers can often be predicted to a certain extent by evaluating the reactions of experimental listeners in a controlled psychophysical testing paradigm. Psychophysical tests (often called "subjective tests") vary depending on the quantity being evaluated, and the structure of the test.

Intelligibility

Speech coder intelligibility is evaluated by coding a number of prepared words, asking listeners to write down the words they hear, and calculating the percentage of correct transcriptions (an adjustment for guessing may be subtracted from the score). The Diagnostic Rhyme Test (DRT) and Diagnostic Alliteration Test (DALT) are intelligibility tests which use a controlled vocabulary to test for specific types of intelligibility loss (Voiers [1983, 1991]). Each test consists of 96 pairs of confusable words spoken in isolation. The words in a pair differ in only one distinctive feature, where the distinctive feature dimensions proposed by Voiers are voicing, nasality, sustention, sibilation, graveness, and compactness. In the DRT, the words in a pair differ in only one distinctive feature of the initial consonant, e.g. "jest" and "guest" differ in the sibilation of the initial consonant. In the DALT, words differ in the final consonant, e.g. "oaf" and "oath" differ in the graveness of the final consonant. Listeners hear one of the words in each pair, and are asked to select the word from two written alternatives. Professional testing firms employ trained listeners who are familiar with the speakers and speech tokens in the database, in order to minimize test-retest variability

Intelligibility scores quoted in the speech coding literature often refer to the composite results of a DRT. In a comparison of two Federal standard coders, the LPC 10e algorithm resulted in 90% intelligibility, while the FS-1016 CELP algorithm had 91% intelligibility (Kohler [1997]). An evaluation of waveform interpolative (WI) coding published DRT scores of 87.2% for the WI algorithm, and 87.7% for FS-1016 (Kleijn and Haagen [1995]).

Numerical Measures of Perceptual Quality

Perhaps the most commonly used speech quality measure is the Mean Opinion Score (MOS). A Mean Opinion Score is computed by coding a set of spoken phrases using a variety of coders, presenting all of the coded speech together with undegraded speech in random order, asking listeners to rate the quality of each phrase on a numerical scale, and then averaging the numerical ratings of all phrases coded by a particular coder. The five-point numerical scale is associated with a standard set of descriptive terms: 5=Excellent, 4=Good, 3=Fair, 2=Poor, and 1=Bad. A rating of 4 is supposed to correspond to standard toll-quality speech, quantized at 64 kbps using ITU standard G.711 (ITU-T [1993b]).

Mean opinion scores vary considerably depending on background noise conditions: for example, CVSD performs significantly worse than LPC-based methods in quiet recording conditions, but significantly better under extreme noise conditions (Tardelli and Kreamer [1996]). Gender of the speaker may also affect the relative ranking of coders (Tardelli and Kreamer [1996]). Expert listeners tend to give higher rankings to speech coders with which they are familiar, even when they are not consciously aware of the order in which coders are presented (Tardelli and Kreamer [1996]). Factors such as language and location of the testing laboratory may shift the scores of all coders up or down, but tend not to change the rank order of individual coders (ISO/IEC [1998e]). For all of these reasons, a serious MOS test must evaluate several reference coders in parallel with the coder of interest, and under identical test conditions. If an MOS test is performed carefully, inter-coder differences of approximately 0.15 opinion points may be considered significant. Figure 12.2 is a plot of MOS as a function of bit rate for coders evaluated under quiet listening conditions in five published studies (one study included separately tabulated data from two different testing sites (Tardelli and Kreamer [1996])).

The diagnostic acceptability measure (DAM) is an attempt to control some of the factors which lead to variability in published MOS scores (Voiers [1977]). The DAM employs trained listeners, who rate the quality of standardized test phrases on ten independent perceptual scales, including six scales which rate the speech itself (fluttering, thin, rasping, muffled, interrupted, nasal), and four scales which rate the background noise (hissing, buzzing, babbling, rumbling). Each of these is a 100-point scale, with a range of approximately 30 points between the LPC-10e algorithm (50 points) and clean speech (80 points) (Tardelli and Kreamer [1996]). Scores on the various perceptual scales are combined into a composite quality rating. DAM scores are useful for pointing out specific defects in a speech coding algorithm. If the only desired test outcome is a relative quality ranking of multiple coders, a carefully controlled MOS test in which all coders of interest are tested under the same conditions

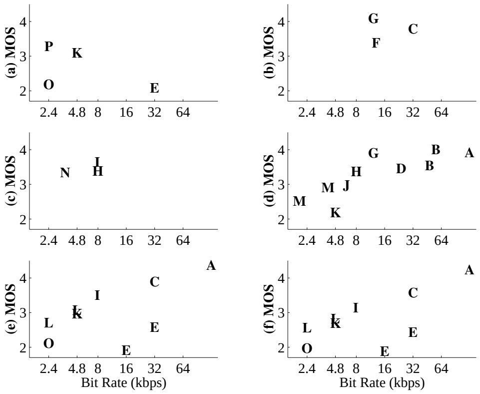


Figure 12.2: Mean opinion scores from five published studies in quiet recording conditions: JARVINEN (Jarvinen *et al.* [1997]), KOHLER (Kohler [1997]), MPEG (ISO/IEC [1998e]), YELDENER (Yeldener [1999]), and the COMSAT and MPC sites from Tardelli et al. (Tardelli and Kreamer [1996]). (A) Unmodified speech, (B) ITU G.722 Subband ADPCM, (C) ITU G.726 ADPCM (D) ISO MPEG-II Layer 3 subband audio coder, (E) DDVPC CVSD, (F) GSM Full-rate RPE-LTP, (G) GSM EFR ACELP, (H) ITU G.729 ACELP, (I) TIA IS-54 VSELP, (J) ITU G.723.1 MPLPC, (K) DDVPC FS-1016 CELP, (L) sinusoidal transform coding, (M) ISO MPEG-IV HVXC, (N) INMARSAT Mini-M AMBE, (O) DDVPC FS-1015 LPC-10e, (P) DDVPC MELP.

may be as reliable as DAM testing (Tardelli and Kreamer [1996]).

Comparative Measures of Perceptual Quality

It is sometimes difficult to evaluate the statistical significance of a reported MOS difference between two coders. A more powerful statistical test can be applied if coders are evaluated in explicit A/B comparisons. In a comparative test, a listener hears the same phrase coded by two different coders, and chooses the one which sounds better. The result of a comparative test is an apparent preference score, and an estimate of the significance of the observed preference: for example, in a recent study, WI coding at 4.0 kbps was preferred to 4 kbps HVXC 63.7% of the time, to 5.3 kbps G.723.1 57.5% of the time (statistically significant differences), and to 6.3 kbps G.723.1 53.9% of the time (not statistically significant) (Gottesman and Gersho [1999]). It should be noted that "statistical significance" in such a test refers only to the probability that the same listeners listening to the same waveforms will show the same preference in a future test.

12.1.2 Objective Measures: Broadband

Psychophysical testing is often inconvenient; it is not possible to run psychophysical tests to evaluate every proposed adjustment to a speech coder. For this reason, a number of algorithms have been proposed which approximate, to a greater or lesser extent, the results of psychophysical testing. The most universal of these measures, though shunned in the speech coding literature, is signal-to-noise ratio (SNR). An easily computed measure with much better perceptual relevance is the segmental SNR (SEGSNR); because of its simplicity, SEGSNR is one of the most widely cited objective measures in the speech coding literature.

Psychological experiments suggest that the ability of listeners to detect the noise in a signal is well predicted by their ability to detect a change in the loudness of the signal. Specifically, listeners can detect the difference between signals $x[n]$ and $\hat{x}[n]$ if there is any period of N samples such that

$$\Delta L < \left| 10 \log_{10} \frac{\sum_{n=0}^{N-1} \hat{x}^2[n]}{\sum_{n=0}^{N-1} x^2[n]} \right| \tag{12.1}$$

That is, listeners detect a difference if the difference in signal power, over any N-sample period, is at least ΔL dB (in fact, equation 12.1 is done separately in each frequency band, but we will ignore frequency dependence for now). Experiments show that there is a signal-dependent tradeoff between the threshold ΔL and the length of the integrating window N. For transient and noisy sounds, $N \approx 5 - 10$ms, and $\Delta L \approx 1$ dB (indeed, the reason that acousticians measure in decibels instead of Bels, centibels, or millibels, is that 1dB is a good approximation to ΔL). For longer, tonal, periodic sounds, $N \approx 20 - 50$ms, and $\Delta L \approx 0.1$dB (one centibel).

Psychologists model signal comparisons using equation 12.1; audio engineers model signal comparisons using SNR. Converting from equation 12.1 to an equation in terms of SNR requires some approximations. First, assume that $x[n]$ and $e[n]$ are uncorrelated, so that

$$10\log_{10}\frac{\sum_{n=0}^{N-1}\hat{x}^2[n]}{\sum_{n=0}^{N-1}x^2[n]} \approx 10\log_{10}\left(1+\frac{\sum_{n=0}^{N-1}e^2[n]}{\sum_{n=0}^{N-1}x^2[n]}\right) \tag{12.2}$$

Second, use the Taylor expansion of $\log(1+z)$ to obtain

$$10\log_{10}\left(1+\frac{\sum_{n=0}^{N-1}e^2[n]}{\sum_{n=0}^{N-1}x^2[n]}\right) \approx \frac{10}{\log(10)}\left(\frac{\sum_{n=0}^{N-1}e^2[n]}{\sum_{n=0}^{N-1}x^2[n]}\right) \tag{12.3}$$

So, from equation 12.1, noise is audible if

$$\frac{\log(10)\Delta L}{10} < \frac{\sum_{n=0}^{N-1}e^2[n]}{\sum_{n=0}^{N-1}x^2[n]} \tag{12.4}$$

Taking the logarithm of both sides, we find that noise is audible if

$$\text{dB SNR}_0^{N-1} < 10 - 10\log_{10}(\log(10)\Delta L) \tag{12.5}$$

For transient or noise-like signals, we find that noise is audible when the SNR in any given 5-10ms frame (within any frequency band) is less than about 5.5dB. For tonal signals, we find that noise is audible when the SNR in any given 20-30ms frame is less than 15.5dB.

The signal to noise ratio of a frame of N speech samples starting at sample number n may be defined as

$$SNR(n) = \frac{\sum_{m=n}^{n+N-1}s^2(m)}{\sum_{m=n}^{n+N-1}e^2(m)} \tag{12.6}$$

High-energy signal components can mask quantization error which is synchronous with the signal component, or separated by at most a few tens of milliseconds. Over longer periods of time, listeners accumulate a general perception of quantization noise, which can be modeled as the average log segmental SNR:

$$SEGSNR = \frac{1}{K}\sum_{k=0}^{K-1}10\log_{10}SNR(kN) \tag{12.7}$$

12.1.3 Objective Measures: Critical Band

High-amplitude signal components tend to mask quantization error components at nearby frequencies and times. A high-amplitude spectral peak in the speech signal is able to mask quantization error components at the same frequency, at higher frequencies, and to a much lesser extent, at lower frequencies. Given a short-time speech spectrum $S(e^{j\omega})$, it is possible to compute a short-time "masking spectrum" $M(e^{j\omega})$ which describes the threshold energy at frequency ω below which noise components are inaudible. The perceptual salience of a noise signal $e(n)$ may be estimated by filtering the noise signal into K different subband signals $e_k(n)$, and computing the ratio between the noise energy and the masking threshold in each subband:

$$NMR(n,k) = \frac{\sum_{m=n}^{n+N-1} e_k^2(m)}{\int_{\omega_k}^{\omega_{k+1}} |M(e^{j\omega})|^2 d\omega} \tag{12.8}$$

where ω_k is the lower edge of band k, and ω_{k+1} is the upper band edge. The band edges must be close enough together that all of the signal components in band k are effective in masking the signal $e_k(n)$. The requirement of effective masking is met if each band is exactly one Bark in width, where the Bark frequency scale is described in many references (Rabiner and Juang [1993], Moore [1997]).

Fletcher has shown that the perceived loudness of a signal may be approximated by adding the cube roots of the signal power in each one-Bark subband, after properly accounting for masking effects (Fletcher [1953]). The total loudness of a quantization noise signal may therefore be approximated as

$$NMR(n) = \sum_{k=0}^{K-1} \left(\frac{\sum_{m=n}^{n+N-1} e_k^2[m]}{\int_{\omega_k}^{\omega_{k+1}} |M(e^{j\omega})|^2 d\omega} \right)^{1/3} \tag{12.9}$$

12.1.4 Automatic Prediction of Subjective Measures

The ITU Perceptual Speech Quality Measure (PSQM) computes the perceptual quality of a speech signal by filtering the input and quantized signals using a Bark-scale filterbank, nonlinearly compressing the amplitudes in each band, and then computing an average sub-band signal to noise ratio (ITU-T [1998a]). The development of algorithms which accurately predict the results of MOS or comparative testing is an area of active current research, and a number of improvements, alternatives, and/or extensions to the PSQM measure have been proposed. An algorithm which has been the focus of considerable research activity is the Bark Spectral Distortion measure (Wang *et al.* [1992], Yang *et al.* [1998], Yang and Yantorno [1999], Novorita [1999]). The ITU has also proposed an extension of the PSQM standard called Perceptual Evaluation of Speech Quality (PESQ) (Rix *et al.* [2000]), which will be released as ITU standard P.862.

12.1.5 Computationally Efficient Measures

Not all types of distortion are equally audible. Many types of speech coders, including LPC-AS coders, use simple models of human perception in order to minimize the audibility of different types of distortion. In LPC-AS coding, two types of perceptual weighting are commonly used. The first type, perceptual weighting of the residual quantization error, is used during the LPC excitation search in order to choose the excitation vector with the least audible quantization error. The second type, adaptive post-filtering, is used to reduce the perceptual importance of any remaining quantization error.

Perceptual Weighting of the Residual Quantization Error

The excitation in an LPC-AS coder is chosen to minimize a perceptually weighted error metric. Usually, the error metric is a function of the time domain waveform error signal

$$e(n) = s(n) - \hat{s}(n) \tag{12.10}$$

Early LPC-AS coders minimized the mean-squared error

$$\sum_n e^2(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |E(\omega)|^2 d\omega \tag{12.11}$$

It turns out that the MSE is minimized if the error spectrum, $E(\omega)$, is white—that is, if the error signal $e(n)$ is an uncorrelated random noise signal, as shown in Figure 12.24.

Not all noises are equally audible. In particular, noise components near peaks of the speech spectrum are hidden by a "masking spectrum" $M(\omega)$, so that a shaped noise spectrum at lower SNR may be less audible than a white noise spectrum at higher SNR. The audibility of noise may be estimated using a noise-to-masker ratio $|E_w|^2$:

$$|E_w|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|E(\omega)|^2}{|M(\omega)|^2} d\omega \tag{12.12}$$

The masking spectrum $M(\omega)$ has peaks and valleys at the same frequencies as the speech spectrum, but the difference in amplitude between peaks and valleys is somewhat smaller than that of the speech spectrum. A variety of algorithms exist for estimating the masking spectrum, ranging from extremely simple to extremely complex (ITU-T [1998a]). One of the simplest model masking spectra which has the properties just described is as follows (Atal and Remde [1982]):

$$M(z) = \frac{|A(z/\gamma_2)|}{|A(z/\gamma_1)|}, \qquad 0 < \gamma_2 < \gamma_1 \leq 1 \tag{12.13}$$

where $1/A(z)$ is an LPC model of the speech spectrum. The poles and zeros of $M(z)$ are at the same frequencies as the poles of $1/A(z)$, but have broader bandwidths. Since the zeros of $M(z)$ have broader bandwidth than its poles, $M(z)$ has peaks where $1/A(z)$ has peaks, but the difference between peak and valley amplitudes is somewhat reduced.

The noise-to-masker ratio may be efficiently computed by filtering the speech signal using a perceptual weighting filter $W(z) = 1/M(z)$. The perceptually weighted input speech signal is

$$S_w(z) = W(z)S(z) \tag{12.14}$$

Likewise, for any particular candidate excitation signal, the perceptually weighted output speech signal is

$$\hat{S}_w(z) = W(z)\hat{S}(z) \tag{12.15}$$

Given $s_w(n)$ and $\hat{s}_w(n)$, the noise-to-masker ratio may be computed as follows:

$$|E_w|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |S_w(\omega) - \hat{S}_w(\omega)|^2 d\omega = \sum_n (s_w^2(n) - \hat{s}_w^2(n)) \tag{12.16}$$

Adaptive Post-Filtering

Despite the use of perceptually weighted error minimization, the synthesized speech coming from an LPC-AS coder may contain audible quantization noise. In order to minimize the perceptual effects of this noise, the last step in the decoding process is often a set of adaptive post-filters (Ramamoorthy and Jayant [1984], Chen and Gersho [1995]). Adaptive post-filtering improves the perceptual quality of noisy speech by giving a small extra emphasis to features of the spectrum which are important for human-to-human communication, including the pitch periodicity (if any) and the peaks in the spectral envelope.

A pitch post-filter (or long-term predictive post-filter) enhances the periodicity of voiced speech by applying either an FIR or IIR comb filter to the output. The time delay and gain of the comb filter may be set equal to the transmitted pitch lag and gain, or they may be recalculated at the decoder using the reconstructed signal $\hat{s}(n)$. The pitch post-filter is only applied if the proposed comb filter gain is above a threshold; if the comb filter gain is below threshold, the speech is considered unvoiced, and no pitch post-filter is used. For improved perceptual quality, the LPC excitation signal may be interpolated to a higher sampling rate in order to allow the use of fractional pitch periods; for example, the post-filter in the ITU G.729 coder uses pitch periods quantized to 1/8 sample.

A short-term predictive post-filter enhances peaks in the spectral envelope. The form of the short-term post-filter is similar to the form of the masking function $M(z)$ introduced in the previous Section: the filter has peaks at the same frequencies as $1/A(z)$, but the peak-to-valley ration is less than that of $A(z)$.

Post-filtering may change the gain and the average spectral tilt of $\hat{s}(n)$. In order to correct these problems, systems which employ post-filtering may pass the final signal through a one-tap FIR pre-emphasis filter, and then modify its gain, prior to sending the reconstructed signal to an A/D converter.

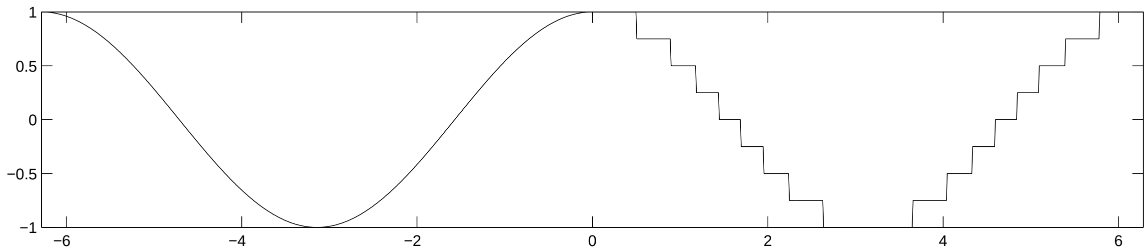


Figure 12.3: Memoryless quantization encodes an audio signal by rounding it to the nearest of a set of fixed quantization levels.

12.2 Quantization

A memoryless quantizer is an audio decoder that converts each speech sample $x[n]$ into a code word $q[n]$, and hence into one synthesized audio sample $\hat{x}[n]$, using a time-invariant mapping called a "codebook." Since the mapping is time-invariant, we can drop the n dependence, and just write that

$$q \rightarrow \hat{x}_q, \quad 0 \leq q \leq Q-1 \tag{12.17}$$

In order to use a memoryless dequantizer, it is necessary that both the quantizer and dequantizer are using the same codebook: that is, both quantizer and dequantizer know the values $\hat{x}_0, \ldots, \hat{x}_{Q-1}$. If both quantizer and dequantizer know the codebook values, then the minimum mean-squared-error quantization rule is given by

$$x \rightarrow q \;\text{ such that }\; (x-\hat{x}_q)^2 = \min_{0\leq i\leq Q-1} (x-\hat{x}_i)^2 \tag{12.18}$$

The quantization process given in equation 12.18 is sometimes referred to as "rounding $x[n]$ to the nearest reconstruction level." In the case of uniform quantization, equation 12.18 is actually most easily implemented using an integer rounding function.

Equation 12.18 can also be written in terms of a set of "thresholds" $T_1, \ldots, T_{Q-1}$:

$$q \rightarrow \begin{cases} 0 & x < T_1 \\ q & T_q \leq x < T_{q+1}, 1 \leq q \leq Q-1 \\ Q-1 & T_{Q-1} \leq x \end{cases} \tag{12.19}$$

where, in order to minimize mean-squared synthesis error, the thresholds must be set so that

$$T_q = \frac{1}{2}\left(\hat{x}_{q-1} + \hat{x}_q\right), \quad 1 \leq q \leq Q-1 \tag{12.20}$$

In order to uniquely transmit the integers $q[n]$, it is necessary to transmit at least B bits per sample, where

$$B = \text{ceil} \log_2 Q \tag{12.21}$$

where the notation means that B is the smallest integer greater than or equal to $\log_2 Q$. Given B bits per sample, there are $Q = 2^B$ codewords available to be used. It might seem that the use of any value of Q less than 2^B is a waste of codewords, but almost every existing audio coding standard uses $Q = 2^B - 1$, for reasons we will discuss later. Note that a data rate of B bits per sample corresponds to a data rate of BF_s bits/second, where F_s is the sampling frequency.

12.2.1 Uniform Quantization

Uniform quantization (also called "linear quantization") is memoryless quantization defined by the following set of codebook values:

$$\begin{aligned}\hat{x}_q &= \hat{x}_0 + q\Delta & (12.22)\\ &= T_q + \frac{\Delta}{2} & (12.23)\\ &= T_{q+1} - \frac{\Delta}{2} & (12.24)\end{aligned}$$

where

$$\Delta = \frac{T_Q - T_0}{Q} \tag{12.25}$$

Notice that equation 12.25 defines two additional thresholds, T_0 and T_Q, that bound the range of effective quantization. T_0 and T_Q are not used for quantization: it is still true that $q = 0$ whenever $x < T_1$, as shown in equation 12.19. Instead, T_0 and T_Q are abstract codebook definition parameters that are only really useful because of their definition in equation 12.25.

Uniform quantization is the most common type of quantization for two reasons. First, it is not necessary to explicitly store the codebook at both coder and decoder. Instead it is sufficient to store any three of the following four quantities: T_0, T_Q, Δ, Q. Indeed, it is not even necessary to explicitly find the minimum-MSE reconstruction level: instead, one simply rounds off $x[n]$ to the nearest reconstruction level:

$$q[n] = \min\left(Q-1, \max\left(0, \text{floor}\left(\frac{x[n]-T_0}{\Delta}\right)\right)\right) \tag{12.26}$$

where floor(x) truncates x to the greatest integer less than or equal to x.

Second, the signal to noise ratio of a uniform quantizer can be computed from theory, and the theoretical SNR is very accurate. There are two types of errors to be concerned about. "Clipping error" occurs when $x < T_0$ or $x > T_Q$. Clipping errors can be very large, so T_0 and T_Q should be chosen such that the probability $P(x < T_0 \text{ or } x > T_Q)$ is as small as possible; usually, this means choosing the range $T_Q - T_0$ to be as large as possible. "Quantization error" occurs when $T_0 \leq x \leq T_Q$. Quantization error is proportional to Δ, which is proportional to $T_Q - T_0$, and therefore the range $T_Q - T_0$ should be chosen to be as small as possible. Obviously, the objectives of minimum clipping and minimum quantization error are in conflict. Balancing these two conflicting goals requires a detailed analysis of the probability density of x. The analysis below will set up the problem formulation, and then suggest a standard solution.

The total probability of clipping may be computed by integrating the probability density function of x over the clipping regions:

$$P_{clip} = \int_{-\infty}^{T_0} p(x)dx + \int_{T_Q}^{\infty} p(x)dx \tag{12.27}$$

P_{clip} can be minimized by choosing T_0 and T_Q so that they cover the range of values for which $p(x)$ is largest.

Given that clipping does not occur, the probability density function of the error is obtained from $p(x)$ by adding up the Q different ways in which an error can occur:

$$p(e|\text{not clipped}) = \sum_{q=0}^{Q-1} p\left(x - \hat{x}_q \;\middle|\; |x - \hat{x}_q| < \frac{\Delta}{2}\right) \tag{12.28}$$

In most audio applications, $p(x)$ is symmetric with respect to its mean, and therefore when the various overlapping PDFs in equation 12.28 are added together, the result is uniformly distributed between $\Delta/2$ and $-\Delta/2$:

$$p(e|\text{not clipped}) = \begin{cases} \frac{1}{\Delta} & -\frac{\Delta}{2} \le e \le \frac{\Delta}{2} \\ 0 & \text{otherwise} \end{cases} \tag{12.29}$$

and the error power is given by

$$E[e^2|\text{not clipped}] = \frac{\Delta^2}{12} \tag{12.30}$$

If T_0 and T_Q are set so that $P_{clip} \approx 0$, then the signal to noise ratio of the quantizer is

$$\text{dB SNR} = 10\log_{10}\frac{12\sigma_x^2}{\Delta^2} \quad \text{if } P_{clip} \approx 0 \tag{12.31}$$

where σ_x^2 is the signal power. It is often useful to write the SNR in terms of the number of quantization levels. The result is:

$$\text{dB SNR} = 10\log_{10}\frac{12\sigma_x^2 Q^2}{(T_Q - T_0)^2} \quad \text{if } P_{clip} \approx 0 \tag{12.32}$$

Remember that the threshold levels T_0 and T_Q must be chosen so that P_{clip} is as small as possible. Often, T_0 and T_Q are chosen so that $T_Q - T_0$ is some multiple of the signal standard deviation σ_x. Define the "safety ratio" to be

$$R = \frac{T_Q - T_0}{\sigma_x\sqrt{12}} \tag{12.33}$$

then the SNR is given by

$$\text{dB SNR} = 10\log_{10}\frac{Q^2}{R^2} \quad \text{if } P_{clip} \approx 0 \tag{12.34}$$

It is also interesting to compute the power spectrum of the error. Consider first the two extreme cases: $Q = 0$, and Q very large. When $Q = 0$, there are zero bits transmitted, so the synthesized signal is $\hat{x}[n] = 0$ always. The error is therefore $e[n] = -x[n]$, and the power spectrum is $R_e(\omega) = R_x(\omega)$. At the other extreme, when Q is large enough, $\hat{x}[n] \neq \hat{x}[n-1]$ for all n. Unless there is an unusual relationship between $p(x)$ and the reconstruction levels $\hat{x}_q$, it will generally be the case that the errors $\hat{x}[n] - x[n]$ and $\hat{x}[n-1] - x[n-1]$ are uncorrelated random variables, thus

$$\begin{aligned} r_e[\tau] &= \sigma_e^2\delta[\tau] \quad \text{if } Q \text{ large enough} && (12.35)\\ R_e(\omega) &= \sigma_e^2 && (12.36)\end{aligned}$$

where $\delta[\tau]$ is the Dirac delta function. In between these two extremes, for values of Q between about $2 \leq Q \leq 16$, the power spectrum of $R_e(\omega)$ is gradually smoothed: detailed spectral features (e.g., harmonics of a periodic signal) disappear even at low values of Q, while broad peaks and valleys in the spectrum disappear if Q is larger.

12.2.2 Zero-Mean Uniform Quantization

Audio signals $x[n]$ are distributed symmetrically around $x[n] = 0$, so almost all audio coders choose reconstruction levels that are also symmetric with respect to 0:

$$\begin{aligned} T_0 &= -T_Q && (12.37)\\ \hat{x}_0 &= -\hat{x}_{Q-1} && (12.38)\end{aligned}$$

With this choice, the "safety ratio" is defined to be

$$R = \frac{T_Q}{\sigma_x\sqrt{3}} \tag{12.39}$$

The safety ratio, in this case, has a very simple interpretation. Since x is zero-mean, the probability of clipping is just

$$P_{clip} = \text{Prob}(|x| > \sqrt{3}R\sigma_x) \tag{12.40}$$

If x is a Gaussian random variable, for example, then given R, it is possible to look up the probability of clipping in the Gaussian cumulative density tables that one can find in the back of any statistics textbook. Recall that the SNR is

$$\text{dB SNR} = 10\log_{10}\frac{Q^2}{R^2} \quad \text{if } P_{clip} \approx 0 \tag{12.41}$$

Decreasing R decreases the no-clipping SNR, but increases the probability of clipping. The tradeoff between clipping and SNR is different for different coders; typical values used in audio coder range from $R \approx 1$ ($P_{clip} \approx 0.05$, dB SNR$\approx 10\log_{10} Q^2$) to $R \approx \sqrt{3}$ ($P_{clip} \approx 0.001$, dB SNR $\approx 10\log_{10} Q^2/3$).

Uniform quantizers used for audio coding always use one of two possible codebooks. The first codebook, called a "mid-riser" quantizer, sets Q equal to a power of two ($Q = 2^B$, where B is the number of bits per sample). The result of this choice is that zero is a threshold value: $T_{Q/2} = 0$. The smallest available reconstruction levels are $\hat{x}_{Q/2-1} = -\Delta/2$ and $\hat{x}_{Q/2} = \Delta/2$. During silent portions of the audio signal, when $x[n] \approx 0$, the reconstructed signal can not be set exactly equal to zero; instead, it randomly switches back and forth between $\Delta/2$ and $-\Delta/2$. Since there is nothing else happening in the signal, this low-level random switching is often audible.

In order to avoid low-level random switching, most audio coders use a "mid-tread" quantizer, defined by an odd number of quantization levels $Q = 2^B - 1$. The result of this choice is that zero is a quantization level ($\hat{x}_{(Q-1)/2} = 0$), and therefore silent portions of the signal can be exactly reconstructed as $\hat{x}[n] = 0$. The disadvantage of this choice is that, in a sense, one reconstruction level is "wasted." In particular, a one-bit mid-tread quantizer is absolutely useless: the number of reconstruction levels is $2^1 - 1 = 1$, and the single available reconstruction level is $\hat{x} = 0$.

The signal to noise ratio of a mid-riser quantizer is

$$\text{dB SNR} = 10\log_{10}\frac{(2^{2B})}{R^2} \tag{12.42}$$

Since $10\log_{10} 4 \approx 6$, the SNR of a mid-riser quantizer is often written as

$$\text{dB SNR} = 6B - 20\log_{10} R \tag{12.43}$$

The signal to noise ratio of a mid-tread quantizer is

$$\text{dB SNR} = 10\log_{10}\frac{(2^B - 1)^2}{R^2} \leq 6B - 20\log_{10} R \tag{12.44}$$

Many audio engineers use the approximation that dB SNR$\approx 6B$ for fast mental computations. Equation 12.43 demonstrates that dB SNR$= 6B$ exactly when a mid-riser quantizer is designed with a safety ratio of $R = 1$. Thus, for example, an 8-bit uniform quantizer has 48dB SNR; a 16-bit uniform quantizer has 96dB SNR.

12.2.3 Companded PCM

Companded PCM is the name given to coders in which the reconstruction levels $\hat{s}_k$ are not uniformly distributed. Such coders may be modeled using a compressive nonlinearity, followed by uniform PCM, followed by an expansive nonlinearity:

$$s(n) \rightarrow \boxed{\text{Compress}} \rightarrow t(n) \rightarrow \boxed{\text{Uniform PCM}} \rightarrow \hat{t}(n) \rightarrow \boxed{\text{Expand}} \rightarrow \hat{s}(n) \tag{12.45}$$

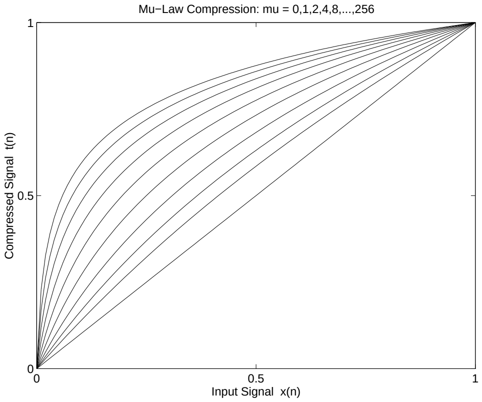


Figure 12.4: μ-law companding function, μ=0,1,2,4,8,. . . ,256.

It can be shown that, if small values of $s(n)$ are more likely than large values, expected error power is minimized by a companding function which results in a higher density of reconstruction levels $\hat{x}_k$ at low signal levels than at high signal levels (Rabiner and Schafer [1978]). A typical example is the μ-law companding function ((ITU-T [1993a]), Figure 12.4), which is given by

$$t(n) = S_{max}\frac{\log\left(1+\mu|s(n)/S_{max}|\right)}{\log(1+\mu)}\text{sign}(s(n)) \qquad (12.46)$$

where μ is typically between 0 and 256 and determines the amount of non-linear compression applied.

12.2.4 Optimum Quantization

12.2.5 Vector Quantization

Equation 12.43 demonstrates that the SNR of an 8-bit linear quantizer with a safety ratio of $R = 1$ is 48dB. Most non-professional listeners, listening over headphones, can easily hear the quantization noise in an 8-bit quantized musical signal. Equation 12.5 demonstrates, on the other hand, that quantization noise is only audible if the SNR in any given frequency band, in any given 20ms frame, is less than 15.5dB. We can infer, therefore, that the only reason that noise is audible in a 48dB SNR signal is that the SNR is not always 48dB. Specifically, the quantization noise always has the same noise power ($\Delta^2/12$), but the signal energy changes dramatically from one frame to the next, and from one frequency band to the next.

Suppose that σ_x^2 is the long-term average signal power, but the the signal power in any particular N-sample frame is

$$\sigma_x^2(f) = \frac{1}{N}\sum_{n=fS}^{fS+N-1}(x[n]-E[x])^2 \qquad (12.47)$$

then the SNR in frame number f is

$$\text{dB SNR}(f) = 6B - 20\log_{10} R + 10\log_{10}\frac{\sigma_x^2(f)}{\sigma_x^2} \tag{12.48}$$

In audio signals, it is quite common for the signal power $\sigma_x^2(f)$ to vary by 60dB from frame to frame; thus, even if the long-term SNR is 48dB, the SNR in any given frame may be as low as -12dB.

A remarkably simple, relatively effective perceptual audio coder may be created by separately coding the maximum amplitude $T_Q(f)$ in any given 10ms "block," then using a 6-bit uniform quantizer (6 bits means approximately 36dB SNR) to code the block of normalized samples $y_f[n]$:

$$\begin{aligned} T_Q(f) &= \max_{fN \le n \le fN+N-1} |x[n]| && (12.49)\\ y_f[n] &= \frac{1}{T_Q(f)}x[n] \quad fN \le n \le fN+N-1 && (12.50) \end{aligned}$$

$y_f[n]$ is usually quantized with uniform quantization on a linear scale, but equation 12.48 suggests that $T_Q(f)$ should be quantized on a logarithmic scale, not a linear scale. Log quantization is achieved by simply using a uniform quantizer to quantize $\log T_Q(f)$, with quantization levels spanning the range between $\log T_{min}$ and $\log T_{max}$. The values of T_{min} and T_{max} vary depending on application. As a common example, consider what happens if $x[n]$ are 16-bit signed integer samples; then the lowest meaningful value of $|x[n]|$ is $T_{min} = 1$, and the highest meaningful value is $T_{max} = 2^{15}$. A simple four-bit quantization of $\log T_Q(f)$ is achieved by transmitting the position of the highest nonzero bit in the binary integer representation of $T_Q(f)$.

The total bit rate of a block uniform quantizer is

$$\text{bits per second} = BF_s + \frac{B_aF_s}{N} \tag{12.51}$$

where B is the number of bits per sample of $y_f[n]$, B_a is the number of bits per amplitude parameter $\log T_Q(f)$, F_s is the sample rate, and F_s/N is the frame rate. The total number of bits/sample is $B + B_a/N$.

Variation of SNR from one frame to the next is measured explicitly by the SEGSNR (segmental SNR). A B-bit uniform quantizer may boast a long-term SNR of $6B$, but SEGSNR is usually much lower. Block quantization explicitly controls SEGSNR by controlling the SNR in each frame, thus long-term SNR and SEGSNR are both approximately equal to $6B$.

12.3 Transform and Sub-Band Coding

One approach to describing a signal with the fewest independent parameters is to approximate the signal, in some sense, by a series of orthogonal functions. The coefficients of the expansion then become the information-bearing quantities. The orthogonal functions chosen for the representation presumably should capitalize upon some known characteristic of the signal.

The orthogonal function approach has been considered for describing both the speech waveform and the amplitude spectrum. A precise waveform description holds relatively small potential for bandwidth reduction–unless information such as phase is discarded and use is made of voiced-unvoiced and pitch tracking measurements. The spectral description , or its time-domain equivalent, promises more. The relationships between short-time spectral analysis and correlation analysis suggest techniques for efficient description of the speech spectrum.

12.3.1 Analytic Rooter

Another technique for frequency division of formant bands of speech is called analytic rooting (SCHROEDER, FLANAGAN and LUNDRY). The processing is done in terms of the analytic signal. This approach avoids characteristic degradations that frequency division methods such as used in the Vobanc introduce.

The analytic signal $\sigma(t)$ of a real, bandlimited signal $s(t)$ is defined as

$$\sigma(t) = s(t) + j\hat{s}(t), \tag{12.52}$$

where $\hat{s}(t)$ is the Hilbert transform of $s(t)$. In polar form the analytic signal is

$$\sigma(t) = a(t)e^{j\Phi(t)},) \tag{12.53}$$

where

$$\begin{aligned} a(t) &= [s^2(t) + \hat{s}^2(t)]^{\frac{1}{2}} \\ \Phi(t) &= \tan^{-l}[\hat{s}(t)/s(t)]. \end{aligned}$$

It follows that

$$s(t) = a(t)\cos[\Phi(t)], \quad \text{and} \quad \hat{s}(t) = a(t)\sin[\Phi(t)]. \tag{12.54}$$

A real signal $s_{1/n}(t)$ corresponding to the n-th root of the analytic signal can be defined as

$$s_{1/n}(t) = \Re[\sigma(t)]^{1/n} \tag{12.55}$$

$$\begin{aligned} &= \Re[s(t) + j\hat{s}(t)]^{1/n} \\ &= [a(t)]^{1/n}\cos[\Phi(t)/n]. \end{aligned}$$

The analytic signal rooting therefore implies division of the instantaneous frequency by a factor n, and taking the nth root of the signal envelope[1]. For the case $n = 2$ the

---

[1] Note that for those cases where perceived pitch is determined by the envelope of the signal waveform, this process leaves the pitch unaltered. This method is therefore attractive for restoring speech distorted by a helium atmosphere, such as breathed by a deep-sea diver.

relations are particularly tractable for computer simulation.

$$s_{\frac{1}{2}}(t) = [a(t)]^{\frac{1}{2}} \cos[\frac{1}{2}\Phi(t)] \tag{12.56}$$

$$= [a(t)]^{\frac{1}{2}}[\frac{1}{2}(1 + \cos\Phi(t))]^{\frac{1}{2}}.$$

Since $a(t)\cos\Phi(t) = s(t)$, one may write (12.56) as

$$s_{\frac{1}{2}}(t) = \left(\frac{1}{2}\right)^{\frac{1}{2}} [a(t) + s(t)]^{\frac{1}{2}} \tag{12.57}$$

Similarly, it can be shown that the Hilbert transform $\hat{s}_{\frac{1}{2}}(t)$ of $s_{\frac{1}{2}}(t)$ is

$$\hat{s}_{\frac{1}{2}}(t) = \left(\frac{1}{2}\right)^{\frac{1}{2}} [a(t) - s(t)]^{\frac{1}{2}} \tag{12.58}$$

Eq. (12.58) also follows from (12.57) by the observation that multiplication of $s(t)$ by -1 is equivalent to a phase shift of π and that, according to (12.56), this corresponds to a phase shift of $\pi/2$ in $s_{\frac{1}{2}}(t)$, i.e., a Hilbert transformation.

Eq. (12.57) is a simple relation which is easy to simulate on a computer and amenable to straight-forward instrumentation–except for one difficulty: the sign of the square root and therefore of $s_{\frac{1}{2}}(t)$, according to (12.57), is indeterminate.

The proper sign can be recovered by changing the sign of the square root in (12.57) every time the phase $\Phi(t)$ of the original signal $s(t)$ goes through 2π (or an integer multiple of 2π). According to (12.54) this is the case when $\hat{s}(t) = 0$, while $s(t) < 0$.

A remaining phase ambiguity of π in $s_{\frac{1}{2}}(t)$ is unavoidable and is a direct consequence of the 2π phase ambiguity in the original signal $s(t)$. This phase ambiguity has no practical consequence.

The inverse operation of analytic-signal rooting is given by

$$s_n(t) = \Re\left[s(t) + j\hat{s}(t)\right]^n . \tag{12.59}$$

By writing

$$s_n(t) = [a(t)]^n \cos\left[n\Phi(t)\right], \tag{12.60}$$

and by comparing (12.60) with (12.55), the inverse relationship is evident.

For $n = 2$, (12.59) yields

$$s_2(t) = \Re\left[s^2(t) + 2js(t)\hat{s}(t) - \hat{s}^2(t)\right], \tag{12.61}$$

or

$$s_2(t) = s^2(t) - \hat{s}^2(t).$$

If process (12.61) is applied to $s_{\frac{1}{2}}(t)$, the original signal $s(t)$ is recovered. This can be verified by substituting $s_{\frac{1}{2}}(t)$ and $\hat{s}_{\frac{1}{2}}(t)$ from (12.57) and (12.58) into (12.61):

$$s_2(t) = \frac{1}{2}\left\{[a(t) + s(t)] - [a(t) - s(t)]\right\},$$

or

$$s_2(t) = s(t). \tag{12.62}$$

The Hilbert transform of the original signal can be recovered by multiplying $s_{\frac{1}{2}}(t)$ and $\hat{s}_{\frac{1}{2}}(t)$:

$$2s_{\frac{1}{2}}(t) \cdot \hat{s}_{\frac{1}{2}}(t) = \left\{[a(t) + s(t)][a(t) - s(t)]\right\}^{\frac{1}{2}} \tag{12.63}$$

$$\begin{aligned} &= \left\{a^2(t) - s^2(t)\right\}^{\frac{1}{2}} \\ &= s(t). \end{aligned}$$

For a signal whose bandwidth is narrow compared to its center frequency, the original signal can be approximately recovered by squaring $s_{\frac{1}{2}}(t)$ and subsequent bandpass filtering. From (12.57),

$$2s^2_{\frac{1}{2}}(t) = a(t) + s(t). \tag{12.64}$$

If the spectrum of $a(t)$ does not overlap that of $s(t)$, which is approximately true for narrowband signals, then $s(t)$ can be recovered by bandpass filtering.

A complete transmission system based upon the foregoing principles has been simulated on a digital computer. In the simulation, the speech spectrum is first divided into four contiguous passbands, each nominally containing no more than one formant. Each bandpass signal is then analytically rooted, band-limited, and recovered in accordance with the previous explanation.

To accomplish square rooting of the signal, and a band reduction of 2-to-1, a typical channel in the flow diagram for the simulation program is shown in Fig. 12.5. The bandpass filter BPF1 separates a spectral segment which nominally contains no more than one formant. The Hilbert transform of this signal is formed by a transversal filter HT1. Since the Hilbert transform filter ideally has a response which is neither time-limited nor band-limited, an approximation is made to the transform which is valid over the frequency range of interest and which is truncated in time.

In a parallel path, the bandpass signal $s(t)$ is delayed by an amount DEL1 equal to one-half the duration of the impulse response of the Hilbert filter. It, too, is squared and $(s^2 + \hat{s}^2)$ is formed by ADD1. The square root of this result yields $a(t)$ in accordance with (12.53), and the addition of the delayed $s(t)$ in ADD2 gives $[a(t) + s(t)]$. Multiplication by $\frac{1}{2}$ and the subsequent square rooting form $s_{\frac{1}{2}}(t)$, according to (12.57).

Selection of the sign of $s_{\frac{1}{2}}(t)$ is accomplished by the following logical decisions in SWITCH. The algebraic sign of $s_{\frac{1}{2}}(t)$ is changed whenever $\hat{s}(t)$ goes through zero

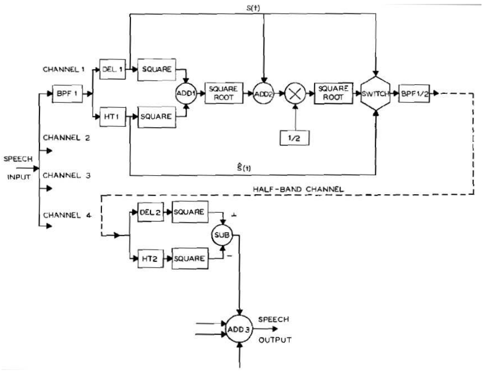


Figure 12.5: Diagram for computer simulation of the analytic rooter. (After (Flanagan and Lundry [1967]))

Table 12.1: Eighth-order Butterworth filter cutoff frequencies in Hz

	BPF1	BPF$\frac{1}{2}$	Formants nominally in passband
Channel 1	238-714	119-357	F1
Channel 2	714-1428	357-714	Fl or F2
Channel 3	1428-2142	714-1071	F2 or F 3
Channel 4	2142-2856	1071-1428	F3

while $s(t) < 0$. The signal $s_{\frac{1}{2}}(t)$, so signed, is then applied to BPF$\frac{1}{2}$, having cutoff frequencies, and hence bandwidth, equal to one-half the values for BPF1.

Analytic squaring of this band-limited version of $s_{\frac{1}{2}}(t)$ is accomplished in accordance with (12.61). The Hilbert transform is produced by HT2, which is similar to HT1 except that the duration of the impulse response of the former is twice that of the latter. Subtracting $\hat{s}^2(t)$ from $s^2(t)$ recovers an approximation to the original bandpassed signal $s(t)$.

The programmed operations in all four channels are identical except that the bandpass filters, Hilbert transform filters, and delays are chosen in accordance with the desired passband characteristics. In the computer implementation, eighth-order Butterworth filters with cutoff frequencies listed in Table 12.1 are used for the bandpass filters.

The Hilbert filters are realized from a band-limited and time-limited approximation to the Hilbert transform. Ideally, the impulse response (inverse) of the Hilbert transform is $h(t) = 1/\pi t$, and the magnitude of the transform is unity at all frequencies. Truncating the spectrum of the transform at frequency ω_c produces an impulse response $\tilde{h}(t) = (\cos \omega_c t - 1)/\pi t$, which although band-limited is not time-limited. The function $\tilde{h}(t)$ is asymmetric and its even Nyquist samples are identically zero. Odd Nyquist samples have the value $2/\pi nT$, where n is the sample number and T is the Nyquist interval. The response $\tilde{h}(t)$ can be truncated (limited) in time at a

Table 12.2: Impulse response durations for the Hilbert filters

	τ in ms	
	HTl	HT2
Channel 1	5.0	10.0
Channel 2	2.5	5.0
Channel 3	1.3	2.5
Channe1 4	0.9	1.7

sufficiently long duration so that over the frequency range of interest the transform remains acceptable.

For programming ease, the transform is realized by an asymmetric transversal filter whose even (Nyquist) coefficients are zero and whose odd coefficients are $2/\pi nT$, weighted with a Hamming window of duration T. Specifically,

$$\tilde{h}(nT) = \frac{1}{\pi nT}\left\{0.54 - 0.46\cos\left(\frac{2\pi(nT + \tau/2)}{\tau}\right)\right\}, \tag{12.65}$$

where $n = l, 2, 3, \ldots, \tau/2T$ represents values for one-half the coefficients of the asymmetrical filter. The simulation is for a 10-kHz bandwidth (ω_c) and $T = 0.5 \times 10^{-4}$ second. The values of the Hamming window used for each of the four bands are given in Table 12.2.

A typical result from the system, with the BPF$\frac{1}{2}$ filters included in the transmission path, is shown by the spectrograms in Fig. 12.6. The upper spectrogram shows an input sentence to the system. The lower spectrogram shows the signal recovered from the half-bandwdith transmission. As the spectrograms show, original formant structure and pitch information is preserved relatively well in the recovered signal. The result is a transmission of respectable quality over a channel bandwidth equal to one-half that of the original signal.

At least one practical hardware implementation of the analyticrooter, using solid-state circuitry, has been constructed and tested (SASS and MACKIE).

12.3.2 Transform Coding: Error Analysis

12.3.3 Expansion of the Speech Waveform

A general method has been described in the literature for representing signal waveforms by orthogonalized, exponential functions (Huggins [1957], Kautz [1954]). The method has been applied to the analysis of single pitch periods of voiced sounds (Dolansky [1960]). If $f(t)$ is a single pitch period, then the approximation

$$f(t) \approx \sum_m c_m g_m(t) \tag{12.66}$$

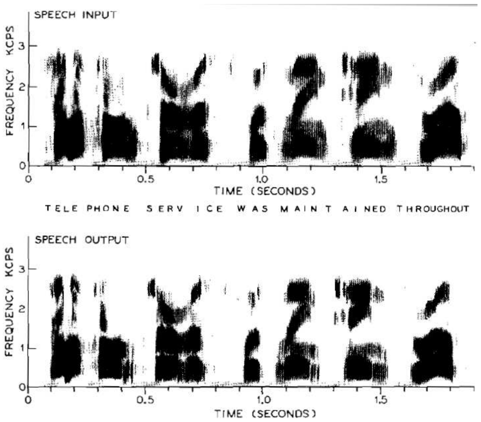


Figure 12.6: Sound spectrograms of speech analyzed and synthesized by the analytic rooter. The transmission bandwidth is one-half the original signal bandwidth. (After (Flanagan and Lundry [1967]))

is made, where the $g_m(t)$ are the set of orthogonalized, exponential functions. Their Laplace transforms of odd and even orders are given by

$$G_{2n-1}(s) = \sqrt{2\alpha_n}\frac{s+|s_n|}{(s-s_n)(s-s_n^*)}\prod_{j=1}^{n-1}\frac{(s+s_j)(s+s_j^*)}{(s-s_j)(s-s_j^*)} \tag{12.67}$$

$$G_{2n}(s) = \sqrt{2\alpha_n}\frac{s-|s_n|}{(s-s_n)(s-s_n^*)}\prod_{j=1}^{n-1}\frac{(s+s_j)(s+s_j^*)}{(s-s_j)(s-s_j^*)}$$

where

$$s_n = (-\alpha_n + j\beta_n)\,.$$

The inverse transforms of Eq. (12.67) are

$$g_{2n-1}(t) = \sum_{k=1}^{n}\frac{1}{\beta_k}\left|\mathcal{K}_{2n-1}(s_k)\right|e^{-\alpha_k t}\sin\left[\beta_k t - \theta_{2n-1}(s_k)\right] \tag{12.68}$$

$$g_{2n}(t) = \sum_{k=1}^{n}\frac{1}{\beta_k}\left|\mathcal{K}_{2n}(s_k)\right|e^{-\alpha_k t}\sin\left[\beta_k t - \theta_{2n}(s_k)\right]$$

where

$$\mathcal{K}_m(s_k) = \left\{G_m(s)\left[(s+\alpha_k^2)+\beta_k^2\right]\right\}_{s=s_k}$$

and

$$\theta_m(s_k) = \frac{\Re\mathcal{K}_m(s_k)}{|\mathcal{K}_m(s_k)|}$$

The first two $g_m(t)$'s, therefore, are simple damped sinusoids which differ in amplitude and phase. The product-series components of $G_m(s)$ are seen to be all-pass functions. An n of 7 (or an m of 14) is considered to be adequate for the speech wave approximation (Dolansky [1960]). The critical frequencies s_n are fixed and are chosen to span the voice frequency range, typically in intervals of a few hundred Hz[2].

Assuming $f(t)$ is zero for $t < 0$, and since

$$\begin{aligned}\int_0^\infty g_p(t)g_q(t)dt &= 1; \quad p = q \\ &= 0; \quad p \neq q\end{aligned}$$

the k-th coefficient of the orthonormal series is given by

$$c_k = \int_0^\infty f(t)g_k(t)dt. \tag{12.69}$$

One straightforward, but impractical, means for measuring the coefficients is apparent. Suppose the signal $f(t)$ is filtered with a realizable filter whose impulse response is $g_k(t)$, the result is

$$O(t) = \int_0^\infty g_k(\tau)f(t-\tau)d\tau. \tag{12.70}$$

If, however, the time-reversed signal $f(-t)$ is filtered, the result is

$$O(t) = \int_0^\infty g_k(\tau)f(t+\tau)d\tau. \tag{12.71}$$

The value $O(0)$, that is, the result at the instant when the time reversed $f(t)$ ends, is the value of c_k. This measurement, performed for all the $g_m(t)$'s, provides the requisite coefficients.

A perhaps more practicable, real-time application of the orthogonal function for speech waveform transmission is shown by the system in Fig. 8.l3a (Manley [1962]). For voiced sounds the input speech is led to a pitch extractor which generates an impulse train at the fundamental frequency. These impulses produce the set of orthogonal functions $g_m(t)$ by exciting realizable networks having the functions as their impulse responses. Approximations to the coefficients of the series (12.66) are obtained by calculating.

$$c_k = \int_0^T g_k(t)f(t)dt, \tag{12.72}$$

where T is a given pitch period. The calculation is carried out by the multipliers, the reset integrators and the sample-and-hold elements shown in the diagram. The pitch pulses reset the integrators and trigger the sampling circuits to read and store the value of the integral at the end of period T. Before multiplexing and transmission, the pitch pulse frequency is converted into an analog signal by a frequency meter,

---

[2] A relevant question might inquire as to the potential of this technique if the s_n could be derived in an adaptive way; that is, if the s_n could be varied to match the signal.

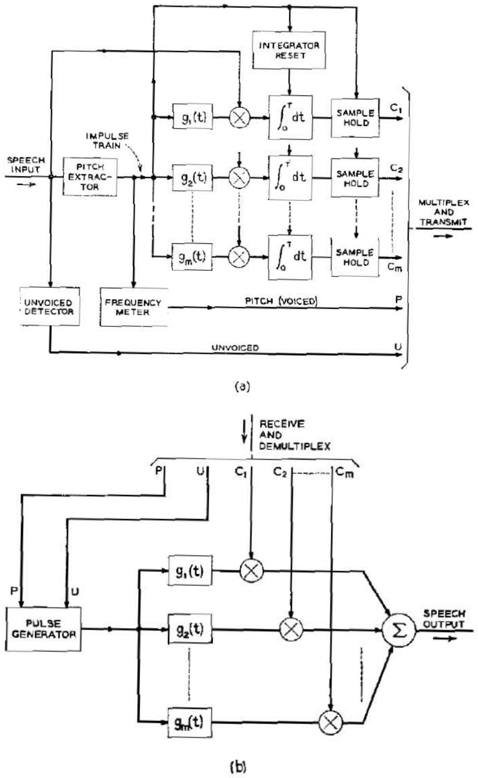


Figure 12.7: System for transmitting speech waveforms in terms of orthogonal functions. (After (Manley [1962])) (a) Analyzer. (b) Synthesizer

and the time-varying coefficients $c_1(t), c_2(t) \ldots c_m(t)$ are further smoothed by low-pass filtering.

At the receiver, in Fig. 12.7b, the signal is reconstructed, pitch period by pitch period, according to Eq. (12.66). A pitch-modulated pulse generator excites an identical set of $g_m(t)$ networks and their outputs are respectively modulated by the $c_m(t)$ coefficients. The sum is an approximation to the original voiced sound.

The processing of unvoiced, aperiodic sounds is slightly different. Ideally they are treated as if their waveforms constituted one pitch period. The onset of an unvoiced sound is detected and, if the unvoiced sound is relatively brief, as in a stop, only one pitch pulse is generated in the transmitter and in the receiver. The unvoiced indication is signalled to the receiver by the $u(t)$ parameter. If the unvoiced sound is sustained (for example, a fricative), the pulse generators are made to continue generating pulses with periods long enough that the periodicity is not significant perceptually.

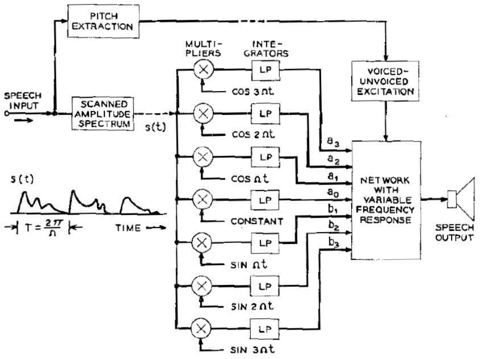


Figure 12.8: Method for describing and synthesizing the short-time speech spectrum in terms of Fourier coefficients. (After (Pirogov [1959a]))

12.3.4 Expansion of the Short-Time Amplitude Spectrum

At least one orthogonal-function description of the short-time amplitude spectrum has been proposed as a bandsaving means for coding speech (Pirogov [1959a,b]). The approach is particularized to a Fourier series description where, in effect, a spectrum of the amplitude spectrum is obtained. The technique is illustrated in Fig. 12.8.

A short-time amplitude spectrum is produced as a time function by scanning at a frequency $1/T$. The operation can be implemented as in the formant extractor described in Section 5.5, or in the manner of the "scan vocoder" discussed in Section 12.4.1, or even with a single scanning filter. The frequency $1/T$ would normally range from 25 to 50 Hz, depending upon the requirements imposed on the quality of transmission. As in the "scan vocoder," the spectral description set) is transmitted over a restricted bandwidth channel. A bandwidth between 75 and 250 Hz is reported to be adequate. Excitation information, that is, pitch and voiced-unvoiced indications, must also be transmitted. As in the conventional vocoder, a bandwidth of 25 to 50 Hz is expected to be adequate for these data. Synchronizing information about the scanning must also be made known to the receiver.

At the receiver, a Fourier series description of the amplitude spectrum is computed, namely,

$$s(t) = \frac{a_0}{2} + \sum_{n=1}^{N} \left[a_n \cos n\Omega t + b_n \sin n\Omega t\right], \tag{12.73}$$

where, as usual, the coefficients are

$$\begin{aligned} a_n &= \frac{2}{T}\int_0^T s(t)\cos n\Omega t dt \\ b_n &= \frac{2}{T}\int_0^T s(t)\sin n\Omega t dt, \end{aligned}$$

and $\Omega = 2\pi/T$. Practically, the Fourier coefficients are obtained by multiplying

$s(t)$ by the outputs of several harmonic oscillators each synchronized to the scanning frequency Ω. An $N = 3$ to 5 is claimed to provide an adequate spectral description (Pirogov [1959a]).

The coefficients vary relatively slowly with time. They are used to control an electrical network so that its frequency response is approximately the same as the measured spectral envelope of the speech signal. The network is then excited in a manner similar to the conventional vocoder, that is, either by periodic pulses or by noise. The reconstructed speech is the output of the variable network shown in Fig. 12.8.

The operation of the controllable network is based upon the fact that $s(t)$ is actually a spectral amplitude $S(\omega)$, $0 \le \omega \le \omega_{max}$. Hence, $\Omega = 2\pi/T = 2\pi/\omega_{max}$, so that Eq. (12.73) can be rewritten as

$$S(\omega) = \frac{a_0}{2} + \sum_{n=1}^{N} a_n \cos \frac{2\pi n \omega}{\omega_{max}} + b_n \sin \frac{2\pi n \omega}{\omega_{max}} \tag{12.74}$$

If the excitation amplitude spectrum is $G(\omega)$, then the output of the variable network should be $S(\omega) \cdot G(\omega)$. Assuming the excitation spectrum is flat and of unity amplitude, a given sine component ω_1 in the excitation spectrum should produce an output time function

$$f_1(t) = \frac{a_0}{2} \sin \omega_1 t + \sin \omega_1 t \sum_{n=1}^{N} a_n \cos \frac{2\pi n \omega_1}{\omega_{max}} + \sin \omega_1 t \sum_{n=1}^{N} b_n \sin \frac{2\pi n \omega_1}{\omega_{max}}. \tag{12.75}$$

Expanding the second and third terms as sums and differences of angles gives

$$\begin{aligned} 2f_1(t) = a_1 \sin \omega_1 t + \sum_{n=1}^{N} a_n \Bigg[\sin \left(\omega_1 t - \frac{2\pi n \omega_1}{\omega_{max}} \right) \\ + \sin \left(\omega_1 t + \frac{2\pi n \omega_1}{\omega_{max}} \right) \Bigg] \\ + \sum_{n=1}^{N} b_n \left[\cos \left(\omega_1 t - \frac{2\pi n \omega_1}{\omega_{max}} \right) - \cos \left(\omega_1 t + \frac{2\pi n \omega_1}{\omega_{max}} \right) \right]. \end{aligned} \tag{12.76}$$

The second terms of the arguments, i.e., $\frac{2\pi n \omega_1}{\omega_{max}}$, correspon to time advances and delays of

$$n\tau = n \cdot \frac{2\pi}{\omega_{max}}$$

The time function can therefore be constructed by the circuit shown in Fig. 12.9. The cosine terms of Eq. (12.76) are obtained by Hilbert transforming a difference of sine terms (i.e., by incurring a broadband π phase shift). Although (12.76) is particularized for a given spectral component of excitation, namely ω_1, the process is the same for all other components. It is reported that with a spectral description of $N = 4$ or 5, the synthesized speech quality is natural enough to satisfy the requirements of ordinary voice channels.

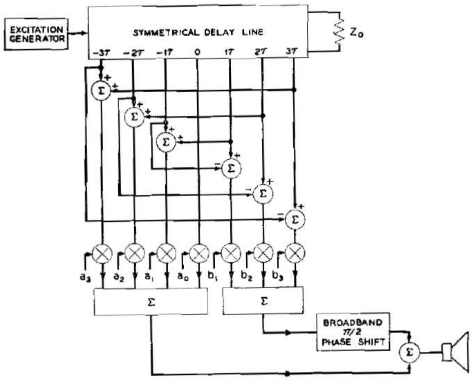


Figure 12.9: Techniques for realizing the variable electrical network of Fig. 12.8

12.3.5 Expansion of the Short-Time Autocorrelation Function

For an on-going time function $f(t)$, the discussion of Chapter 5 derived the relation between the short-time autocorrelation function (defined for positive delays)

$$\phi(\tau,t)=\int_{-\infty}^{t} f(\lambda)f(\lambda-\tau)k(t-\lambda)d\lambda, \quad \tau\geq 0 \tag{12.77}$$

and the measurable short-time amplitude spectrum

$$F(\omega,t)=\int_{-\infty}^{t} f(\lambda)h(t-\lambda)e^{-j\omega\lambda}d\lambda. \tag{12.78}$$

For the specific weighting function

$$k(t)=h^2(t)=2\sigma e^{-2\sigma t},$$

the short-time correlation and spectrum are linked by the weighted Fourier cosine transform

$$\begin{aligned}|F(\omega,t)|^2 &= \int_{-\infty}^{\infty} e^{-\sigma|\tau|}\phi(\tau,t)\cos\omega\tau d\tau \qquad (12.79)\\ &= \frac{1}{2\pi}|H(\omega)|^2 * \Phi(\omega,t),\end{aligned}$$

where $H(\omega)$ and $\Phi(\omega,t)$ are the Fourier transforms of $h(t)$ and $\phi(\tau,t)$, respectively. The transform-pair (12.79) implies that $\phi(\tau,t)$ is an even function of τ.

The preceding section described a technique for representing the short-time amplitude spectrum $|F(\omega,t)|$ in terms of an orthogonal function expansion. Since the correlation function and power spectrum are uniquely linked, it might be expected that a related orthonormal expansion can be written for the correlation function.

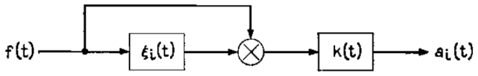


Figure 12.10: Expansion coefficients for the short-time auto-correlation function

This expansion leads to an alternative time-domain representation of the signal. In particular, Laguerre functions have been found a convenient expansion for this description (Lee [1960], Manley [1962], Kulya [1962a,b]).

Suppose the short-time correlation function of $f(t)$ for positive delays is expanded in terms of a realizable function set $\{\xi_i(t)\}$, orthonormal on the internal $0 \leq \tau \leq \infty$ and zero for $\tau < 0$. Then

$$\phi(+\tau, t) = \sum_{i=0}^{\infty} a_i(t)\xi_i(\tau), \quad \tau \geq 0. \tag{12.80}$$

Because of the orthogonal properties

$$a_i(t) = \int_0^{\infty} \phi(+\tau, t)\xi_i(\tau)d\tau \tag{12.81}$$

$$= \int_0^{\infty} \xi_i(\tau)d\tau \int_{-\infty}^{t} f(\lambda)f(\lambda-\tau)k(t-\tau)d\lambda.$$

Changing the order of integration and substituting $\gamma = (\lambda - \tau)$ gives

$$a_i(t) = \int_{-\infty}^{t} f(\lambda)k(t-\lambda)d\lambda \int_{-\infty}^{\lambda} f(\gamma)\xi_i(\lambda-\gamma)d\gamma. \tag{12.82}$$

The coefficients $a_i(t)$ are therefore obtained by first filtering $f(t)$ with a network whose impulse response is $\xi_i(t)$, multiplying the result by $f(t)$ and then filtering the product with a network whose impulse response is $k(t)$. The operations are illustrated in Fig. 12.10.

The $a_i(t)$ coefficients obtained from (12.82) describe $\phi(\tau, t)$ for positive delays $(\tau \geq 0)$. If, as defined and as discussed in Chapter 5, $\phi(\tau, t)$ is an even function of τ, the correlation for negative delay may be written

$$\phi(-\tau, t) = \sum_{i=0}^{\infty} a_i(t)\xi_i(-\tau), \quad \tau < 0, \tag{12.83}$$

and the correlation function for all T is

$$\phi(\tau, t) = \phi(+\tau, t) + \phi(-\tau, t) \tag{12.84}$$

$$= \sum_{i=0}^{\infty} a_i(t)\left[\xi_i(\tau) + \xi_i(-\tau)\right]$$

The Fourier transform of $\phi(\tau, t)$ is the power spectrum

$$\Phi(\omega, t) = \sum_{i=0}^{\infty} a_i(t) \int_{-\infty}^{\infty} \left[\xi_i(\tau) + \xi_i(-\tau)\right] e^{-j\omega\tau} d\tau \tag{12.85}$$

$$= \sum_{i=0}^{\infty} a_i(t) \left\{\Xi_i(\omega) + \Xi_i^*(\omega)\right\}$$

where $\Xi_i(\omega)$ is the Fourier transform of $\xi_i(T)$.

The spectrum $\Phi(\omega, t)$ is related to the measurable power spectrum of Eq. (12.79) such that

$$|F(\omega, t)|^2 = \sum_{i=0}^{\infty} a_i(t) \left\{\Xi_i'(\omega) + \Xi_i'^*(\omega)\right\}, \tag{12.86}$$

where $\Xi_i'(\omega)$ is the Fourier transform of $[e^{-\sigma|\tau|}\xi_i(\tau)]$.

Writing $\Xi_i(\omega)$ in terms of its magnitude and phase,

$$\Xi_i(\omega) = \alpha_i(\omega) e^{-j\beta_i(\omega)}. \tag{12.87}$$

Then

$$\Phi(\omega, t) = \sum_{i=0}^{\infty} a_i(t)\alpha_i(\omega) \left[e^{-j\beta_i(\omega)} + e^{+j\beta_i(\omega)}\right] \tag{12.88}$$

$$= 2\sum_{i=0}^{\infty} a_i(t)\alpha_i(\omega) \cos\beta_i(\omega)$$

Thus the coefficients $a_i(t)$ of an orthonormal expansion of the auto correlation function [Eq. (12.80)] are also the coefficients of a Fourier series expansion of the power spectrum.

So far, the orthogonal filter functions $\xi_i(t)$ have not been particularized. They have only been assumed to be physically realizable impulse responses. One simple set of orthonormal filters–and one that leads to a familiar result–is an ideal delay line with radian bandwidth B and with delay taps spaced at the Nyquist interval $1/2B$. The frequency response at the i-th tap is

$$\Xi_i(\omega) = \begin{cases} e^{-j\left(\frac{i\omega}{2B}\right)}, & 0 \le \omega \le B \\ e^{j\left(\frac{i\omega}{2B}\right)}, & -B \le \omega \le 0 \\ 0, & \text{elsewhere} \end{cases} \tag{12.89}$$

The impulse response at the i-th tap is therefore

$$\xi_i(t) = \frac{B}{\pi} \frac{\sin\left(Bt - \frac{i}{2}\right)}{\left(Bt - \frac{i}{2}\right)} \tag{12.90}$$

As prescribed by Eq. (12.89), the amplitude response is $\alpha_i(\omega) = 1$, and the phase response is

$$\beta_i(\omega) = \left(\frac{i\omega}{2B}\right)$$

The power spectrum expansion of Eq. (12.88) is therefore the Fourier series

$$\Phi(\omega, t) = 2\sum_i a_i(t)\cos\left(\frac{i\omega}{2B}\right). \tag{12.91}$$

The $a_i(t)$, on the other hand, which are computed according to the operations of Fig. 12.10, are simply values of the short-time autocorrelation function $\phi(\tau, t)$ for $\tau = (i\omega/2B)$. These coefficients could be supplied directly to the left side of the synthesizer of Fig. 12.9 and used to generate the spectrum $\Phi(\omega, t)$. In this case, one has a correlation-vocoder synthesizer as described in Section 12.4.

Ideal broadband delay lines are neither physically wieldy nor particularly easy to construct. It is consequently of interest to consider other orthonormal function sets which might be useful in representing the short-time autocorrelation function or the power spectrum. Preferably, the functions should be realizable with simple lumped-element networks. The choice of Laguerre functions has advantages in this connection (Lee [1960]).

Such an orthogonal set is

$$\{\xi_i(t)\} = \{l_i(t)\},$$

where the $l_i(t)$ are described by

$$l_i(t) = (2\lambda)^{\frac{1}{2}} e^{-\lambda t} \sum_{n=0}^{i} \frac{(-1)^n (2\lambda t)^{i-n} (i!/n!)}{[(i-n)!]^2}. \tag{12.92}$$

Its frequency transform is

$$L_i(\omega) = \frac{(2\lambda)^{\frac{1}{2}}}{2\pi} \cdot \frac{(\lambda - j\omega)^i}{(\lambda + j\omega)^{i+1}} \tag{12.93}$$

$$\begin{aligned} &= (-1)^i \frac{1}{\pi(2\lambda)^{\frac{1}{2}}}\left(\frac{\lambda}{j\omega + \lambda}\right)\left(\frac{j\omega - \lambda}{j\omega + \lambda}\right)^i \\ &= A_i[u(\omega)][v(\omega)]^i. \end{aligned}$$

The function (12.93) can be realized by cascading RC circuits of the type shown in Fig. 12.11, together with an amplification A_i.

If (12.93) is put in the form

$$L_i(\omega) = \alpha_i(\omega) e^{j\beta_i(\omega)}, \tag{12.94}$$

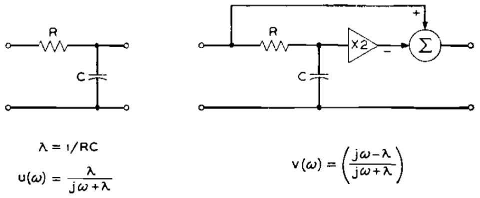


Figure 12.11: Realization of Laguerre functions by RC networks [see Eq. (12.93)]

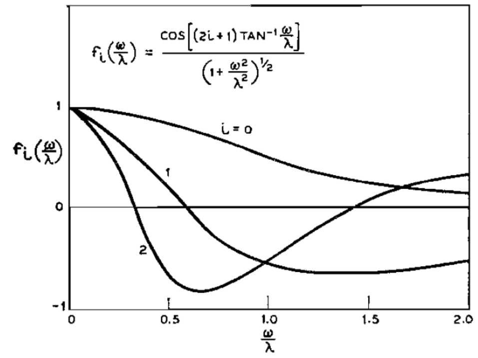


Figure 12.12: Plot of the final factor in Eq. (12.97) showing how the positive frequency range is spanned by the first several Laguerre functions. (After (Manley [1962]))

then

$$L_i(\omega) = \frac{(2\lambda)^{\frac{1}{2}}}{2\pi} \cdot \frac{1}{(\omega^2+\lambda^2)^{\frac{1}{2}}} \cdot e^{j\left[(2i+1)\tan^{-1}\frac{\omega}{\lambda}\right]}. \tag{12.95}$$

Further,

$$[L_i(\omega) + L_i^*(\omega)] = \frac{(2\lambda)^{\frac{1}{2}}}{\pi} \cdot \frac{1}{(\omega^2+\lambda^2)^{\frac{1}{2}}} \cdot \cos\left[(2i+1)\tan^{-1}\frac{\omega}{\lambda}\right]. \tag{12.96}$$

The spectrum $\Phi(\omega, t)$ according to (12.85) and (12.88) is

$$\begin{aligned}\Phi(\omega,t) &= 2\sum_{i=0}^{\infty} a_i(t)\alpha_i(\omega)\cos\beta_i(\omega) \\ &= \left(\frac{2}{\pi^2\lambda}\right)^{\frac{1}{2}} \sum_i a_i(t) \frac{\cos\left[(2i+1)\tan^{-1}\frac{\omega}{\lambda}\right]}{\left(1+\frac{\omega^2}{\lambda^2}\right)^{\frac{1}{2}}} \qquad (12.97)\end{aligned}$$

To show how the positive frequency domain is spanned by the Laguerre functions, the first several terms of the final factor in (12.97) are plotted in Fig. 12.12 (Manley [1962]). The functions are seen to have the desirable feature that they attenuate with increasing frequency, as does the speech spectrum.

A transmission system based upon these relations can be constructed. The assumption is that the spectrum-squaring operation, that is, the synthesis of a signal having a spectrum $\Phi(\omega, t)$, is perceptually acceptable. (See Sections 12.4.1 and 12.4 for other comments on spectrum squaring.) Such a signal is

$$\phi(\tau, t) = \sum_{i=0}^{\infty} a_i(t) \left[l_i(\tau) + l_i(-\tau) \right],$$

having the spectrum

$$\Phi(\omega, t) = \sum_{i=0}^{\infty} a_i(t) \left[L_i(\omega) + L_i^*(\omega) \right]. \tag{12.98}$$

The correlation $\phi(\tau, t)$ is an even function of τ and is produced from $l_i(\tau)$, $\tau \geq 0$. But with the circuits of Fig. 12.11, it is not possible to generate $l_i(-\tau)$. However, the ear is relatively insensitive to modest phase differences, and it suffices perceptually to generate a spectrum whose modulus is the same as $\Phi(\omega, t)$. Such a spectrum can be obtained from the odd function $[l_{m-i}(\tau) + l_{m+i+1}(\tau)]$ (Kulya [1963]). The corresponding spectrum is then

$$\Phi'(\omega, t) = \sum_{i=0}^{\infty} a_i(t) \left[L_{m-i}(\omega) + L_{m+i+l}(\omega) \right],$$

where, from Eq. (12.95),

$$\begin{aligned} &\left[L_{m-i}(\omega) + L_{m+i+1}(\omega) \right] \\ &= \frac{(2\lambda)^{\frac{1}{2}}}{\pi} \frac{1}{(\omega^2 + \lambda^2)^{\frac{1}{2}}} \left[e^{j2(m+1) \tan^{-1} \frac{\omega}{\lambda}} \right] \cos \left[(2i+1) \tan^{-1} \frac{\omega}{\lambda} \right]. \end{aligned} \tag{12.99}$$

Except for the phase angle $\left[e^{j2(m+1) \tan^{-1} \frac{\omega}{\lambda}} \right]$, Eq. (12.99) is identical to Eq. (12.96). The complete transmission system is therefore the circuit shown in Fig. 12.13. In Fig. 12.13a, the Laguerre expansion coefficients are developed according to Eq. (12.98) and after the fashion of Fig. 12.10. A pitch signal $p(t)$ is also extracted. The coefficients and pitch data are multiplexed and transmitted to the synthesizer in Fig. 12.13b. As in the vocoder, the synthesizer excitation is either wide-band noise or pitchmodulated pulses. By resorting to the odd function $[l_{m-i}(\tau) + l_{m+i+1}(\tau)]$, the synthesizer imposes the spectrum $\Phi'(\omega, t)$ upon the broadband excitation. Similar results can be obtained from an orthonormal expansion of the correlation function in terms of Tschebyscheff polynomials (KULYA).

12.4 Correlation Vocoders

The channel vocoder demonstrates that speech intelligibility, to a large extent, is carried in the shape of the short-time amplitude spectrum. Any equivalent specification

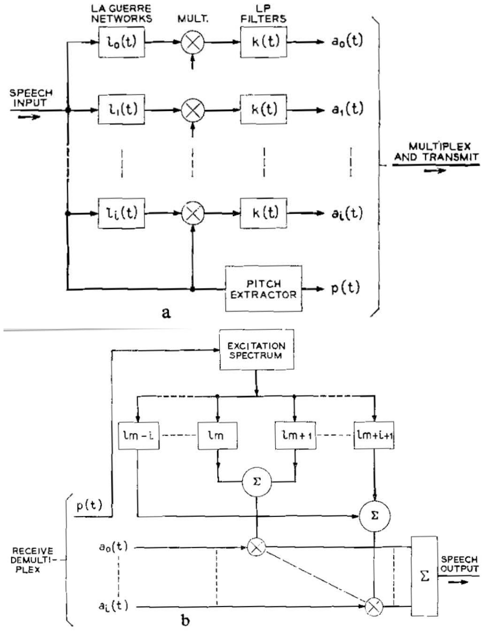


Figure 12.13: A Laguerre function vocoder. (a) Analyzer. (b) Synthesizer. (After (Kulya [1963]))

of the spectral shape would be expected to convey the same information. One equivalent description of the squared amplitude spectrum is the autocorrelation function. The correlation function can be obtained strictly from time-domain operations, and a spectral transformation is not required. Time-domain processing therefore offers simplicities in implementation. The relations linking these quantities have been discussed in detail in Section 5.1, Chapter 5. A short-time autocorrelation specification of the speech signal might therefore be expected to be a time-domain equivalent of the channel vocoder.

In Chapter 5, a short-time autocorrelation function of the function $f(t)$ was defined for the delay parameter, τ, as

$$\phi(\tau, t) = \int_{-\infty}^{t} f(\lambda) f(\lambda + \tau) k(t - \lambda) d\lambda, \tag{12.100}$$

where $k(t) = 0$ for $t < 0$ and is a weighting function or time apperture [usually the impulse response of a physically realizable low-pass filter, see Eq. (5.15)]. Under the special condition $k(t) = 2\alpha e^{-2\alpha t} = h^2(t)$, $\phi(\tau, t)$ can be related to the measurable short-time power spectrum

$$\Psi(\omega, t) = |F(\omega, t)|^2,$$

where

$$F(\omega, t) = \int_{-\infty}^{t} f(\lambda) h(t - \lambda) e^{-j\omega\lambda} d\lambda. \tag{12.101}$$

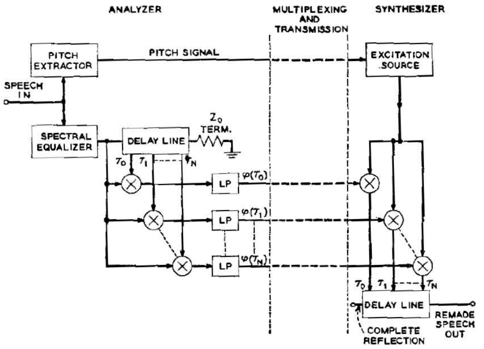


Figure 12.14: Autocorrelation vocoder. (After (Schroeder [1959, 1962]))

In fact, it was shown that

$$\phi(\tau,t) = \frac{e^{\alpha|\tau|}}{2\pi}\int_{-\infty}^{\infty} \Psi(\omega,t)e^{j\omega\tau}d\tau; \tag{12.102}$$

and

$$\Psi(\omega,t) = \int_{-\infty}^{\infty} e^{-\alpha|\tau|}\phi(\tau,t)e^{-j\omega\tau}d\tau. \tag{12.103}$$

In this case the measurable short-time power spectrum–which is essentially the quantity dealt with in the channel vocoder (or rather the square root of it)–is the Fourier transform of the product of the weighting $e^{-\alpha|\tau|}$ and the short-time autocorrelation function $\phi(\tau,t)$. The spectral information might therefore be specified in terms of the correlation. Several transmission methods for utilizing this relation have been examined (Huggins [1954], Schroeder [1959, 1962], Kock [1956, 1959, 1962], Biddulph [1954]).

One method for applying the principle is shown in Fig. 12.14. In the top branch of the circuit, the input speech is submitted to pitch extraction. This information is derived and employed in a manner identical to the channel vocoder. In the lower branch of the circuit, the input signal is put through a spectral equalizer which, in effect, takes the square root of the input signal spectrum. The basis for this operation is that the ultimate processed signal is going to be a correlation function whose Fourier transform is the power spectrum (or, the squared amplitude spectrum) of the input signal. Although spectrum-squared speech is generally quite intelligible, it has an unnatural intensity or stress variation. Since the spectrum squaring is inherent in the process, it is taken into account at the outset.

After spectral square-rooting, the short-time autocorrelation function of the signal is computed for specified delays. This is done by multiplying the appropriate output of a properly terminated delay line with the original input, and low-pass filtering the product (in this case with a 20-Hz low-pass filter). The impulse response of the low-pass filter is the $k(t)$ as given in Eq. (12.100). Since the autocorrelation function is bandlimited to the same frequency range as the signal itself, the correlation function is completely specified by sampling at the Nyquist interval (i.e., $1/2BW$). For a 3000 Hz signal, therefore, a delay interval $\Delta\tau = 0.167$ msec is sufficient. The greatest delay to which the function needs to be specified, practically, is on the order of 3 msec (Schroeder [1962]). Thus a total of 18 delay channels–each using about 20 Hz bandwidth–are required. The total bandwidth is therefore 360 Hz and is about the same as required by the channel vocoder.

At the synthesizer, voiced sounds are produced by generating a periodic waveform in which the individual pitch period is the correlation function described by the values existing on the $n\tau$-channels at that instant. The waveform is generated by letting the pitch pulses of the excitation "sample" the individual τ-channels, The sampling is accomplished by multiplying the excitation and each channel signal. The samples are assembled in the correct order by a delay line, and are low-pass filtered to yield the continuous correlation function. Since the correlation function is even, the synthesized wave is made symmetrical about the τ_0 sample. This can be done practically with the delay line correctly terminated at its output end, but unterminated and completely reflecting at the far end, as shown in Fig. 12.14. Low-pass filtering of the samples emerging from the line recovers the continuous signal.

Because a finite delay is used in the analysis, the measured correlation function is truncated, and discontinuities will generally exist in the synthesized waveform. This leads to a noticeable distortion. The distortion can be reduced by weighting the high-delay correlation values so that they have less influence in the synthesized wave. The discontinuities are thereby smoothed, and the processed speech obtained approaches that from channel vocoders of the same bandwidth compression[3].

12.4.1 Channel Vocoders

Analysis-synthesis telephony came of age, so to speak, with Dudley's invention of the Vocoder (Dudley [1939]). In recent years, the name Vocoder (for Voice Coder) has become largely a generic term, commonly applied to analysis-synthesis systems in which the excitation and system functions are treated separately (see Fig. 12.1). The original Vocoder–now referred to as a spectrum channel vocoder-has probably been described in the literature more times than any other single system. Nevertheless, for the sake of completeness, as a convenient point of departure, and because it set forth, at such an early time, an important philosophy in voice transmission, a brief description of the idea will be repeated once more.

Following the coding scheme illustrated in Fig. 12.1, the Vocoder incorporates one important constraint of speech production and one of perception. It recognizes that the vocal excitation can be a broadspectrum, quasi-harmonic sound (voiced), or a

---

[3] This truncation distortion in synthesis can be avoided if the correlation data are used to control a recursive filter. See the technique devised for the Maximum Likelihood Vocoder (Itakura and Saito [1968]) in Section 12.6.2.

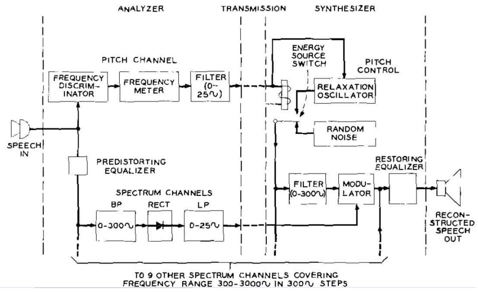


Figure 12.15: Block diagram of the original spectrum channel vocoder. (After (Dudley [1939]))

broad-spectrum, random signal (unvoiced). It also recognizes that perception, to a large degree, is dependent upon preservation of the shape of the short-time amplitude spectrum. A block diagram of an early Vocoder is shown in Fig. 12.15 (DUDLEY, 1939b).

The excitation information is measured by the top branch of the circuit. A frequency discriminator and meter measure the fundamental component of the quasi-periodic voiced sounds. Values of the fundamental frequency and its temporal variations are represented by a proportional electrical voltage from the meter. This "pitch" signal is smoothed by a 25 Hz low-pass filter. Unvoiced sounds normally have insufficient power in the fundamental frequency range to operate the frequency meter. Nonzero outputs of the pitch meter therefore indicate voicing as well as the value of the pitch.

Ten spectrum channels in the lower part of the circuit measure the short-time amplitude spectrum at ten discrete frequencies. Each channel includes a band-pass-filter (300 Hz wide originally), a rectifier and a low-pass filter (25 Hz). The measured spectrum is therefore precisely that described in Section 5.1, Chapter V. The predistorting equalizer pre-emphasizes the signal to produce nearly equal average powers in the spectrum-analyzing filters. The spectrum-defining channel signals consequently have about the same amplitude ranges and signal-to-noise ratios for transmission. The eleven 25-Hz wide signals occupy a total bandwidth of less than 300 Hz and must be multiplexed in frequency or time for transmission.

At the receiver, the speech spectrum is reconstructed from the transmitted data. Excitation, either from a pitch-modulated, constant average power pulse-generator, or from a broadband noise generator, is applied to an identical set of band-pass filters. The outputs from the filters are amplitude modulated by the spectrum-defining signals. A short-time spectrum, approximating that measured at the transmitter, is recreated.

With proper design the synthesized speech can be made surprisingly intelligible. An example of speech transmitted by a 15-channel vocoder is shown by the spectrograms in Fig. 12.16. Important features such as formant structure and voiced-

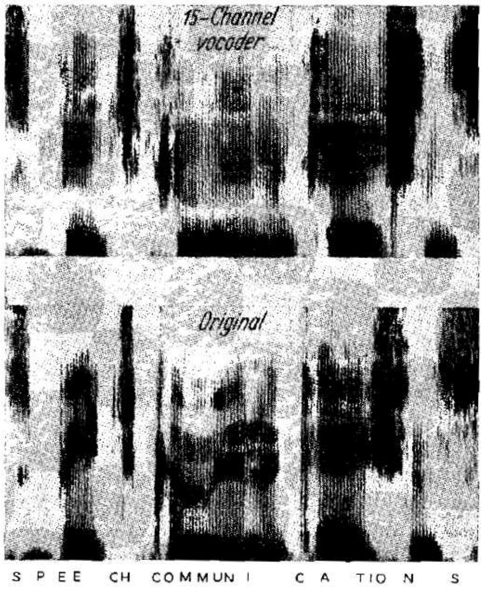


Figure 12.16: Spectrogram of speech transmitted by a 15-channel vocoder

unvoiced excitation are relatively well preserved.

12.4.2 Design Variations in Channel Vocoders

Since the original development of the Vocoder many different versions and variations have been constructed. Number and spacing of the analyzing filters along the frequency scale, their bandwidths, degree of overlap, and selectivity are all matters of design variation. Similarly, many different pitch extraction and voiced-unvoiced detection circuits have been examined, as well as the characteristics of the rectifier and low-pass filter. The number of channels used has ranged from as few as eight to as many as 100, and the filter characteristics have ranged from broad, steep, flat-topped responses to narrow, simple-tuned circuits. Space does not permit a detailed discussion of all these investigations. However, typical analog hardware implementations include those of (Miller [1953], E. E. David [1956], Vilbig and Haase [1956a,b], Slaymaker [1960], Shearme [1962], Shearme *et al.* [1962], Cooper [1957], Werner and Danielsson [1958], L. A. Yaggi [1962], Steele and Cassel [1963b,a]). Digital implementations have been nearly equally varied, and include (Golden [1963], Freudberg *et al.* [1967], Rader [1967]). In particular, Fast Fourier Transform techniques have been found advantageous in digital implementations.

Although intelligibility may be high, practical realizations of conventional channel vocoders generally exhibit a perceptible degradation of speech naturalness and quality. The synthetic speech possesses a machine-like quality which is characteristic of the device. Several factors seem to be responsible. One is the coding of excitation data. Voiced-unvoiced discriminations often are made with noticeable errors.

Table 12.3: Consonant intelligibility for a vocoder. Percent of initial consonants heard correctly in syllables (togatoms). (After (Halsey and Swaffield [1948]))

b–90%	l –97%	r –100%	w –100%
f–74	m–85	s–94	sh –100
h–100	n–99	t–91	th–43
k –85	p–77	y–96	none–70

Relevant structure in the pitch signal may not be preserved, and, under certain conditions, octave errors may be made in the automatic pitch extraction. Voiced sounds are synthesized from a pulse source whose waveform and phase spectrum do not reflect certain details and changes of the real vocal cord wave. The spectral analysis also has a granularity, or lack of resolution, imposed by the number, bandwidth and spacing of the analyzing filters. A given speech formant, for example, might be synthesized with too great a bandwidth. Further, the large dynamic range of the amplitude spectrum may not be covered adequately by practical rectifiers and amplifiers.

The basic channel vocoder design can be improved in several ways. The important excitation problems can be obviated to a large extent by the voice-excitation technique to be discussed in a following section. Also sophisticated pitch extraction methods, such as the cepstrum method described in Section 5.6, Chapter 5, provide more precise pitch and voiced-unvoiced data. The spectral representation problems can be lessened by careful filter design, or by the use of digital techniques such as the Fast Fourier Transform.

12.4.3 Vocoder Performance

Although voice quality and naturalness normally suffer in transmission by vocoder, the intelligibility of the synthesized speech can be maintained relatively high, often with a vocoder having as few as ten channels. For a high-quality microphone input and a fundamental component pitch extractor, typical syllable intelligibility scores for a ten-channel (250 to 2950 Hz) vocoder are on the order of 83 to 85 percent (Halsey and Swaffield [1948]). Typical intelligibility scores for initial consonants range over the values shown in Table 12.3.

Weak fricatives such as θ are not produced well in this system. The 30 percent error indicated for no initial consonant (i.e., for syllables beginning with vowels) indicates imprecision in the voiced-unvoiced switching. Such syllables were heard as beginning with consonants when in fact they did not. Even so, the consonant intelligibilities are reasonably good.

Comparable performances can also be obtained when the vocoder signals are time-sampled (scanned), quantized and encoded in terms of binary pulses. An early model 10-channel vocoder, arranged for digital transmission, gave typical consonant intelligibility scores shown in Table 12.4. The data rates are exclusive of pitch information.

Table 12.4: Vocoder consonant intelligibility as a function of digital data rate. (After (E. E. David [1956]))

	Number of quantizing levels			
	6	5	4	3
Binary pulse rate (bits/sec)	1300	1160	1000	788
Consonant intelligibility (%)	82	79	79	69

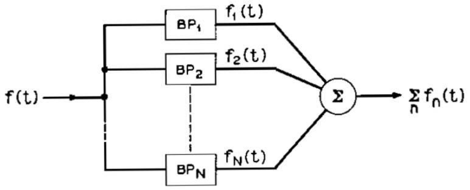


Figure 12.17: Filtering of a speech signal by contiguous band-pass filters

Four different quantizing levels were used (E. E. David [1956])

More elaborate designs provide somewhat higher intelligibilities. For example, a consonant intelligibility of approximately 90 per cent is typical of a 16-channel vocoder whose channel signals are sampled 30 sec^{-1} and quantized to three bits (i.e., 1440 bits/sec) (E. E. David [1956]).

12.4.4 Phase Vocoder

A final frequency division-multiplication method makes use of the short-time phase derivative spectrum of the signal to accomplish the band saving. The method permits non-integer divisions as well as integer values. It can be applied either to single voice harmonics or to wider subbands which can include single formants. It also permits a flexible means for time compressing or expanding the speech signal. The method is called Phase Vocoder (Golden [1966]).

If a speech signal $f(t)$ is passed through a parallel bank of contiguous band-pass filters and then recombined, the signal is not substantially degraded. The operation is illustrated in Fig. 12.17, where $BP_1 \ldots BP_N$ represent the contiguous filters. The filters are assumed to have relatively flat amplitude and linear phase characteristics in their pass bands. The output of the n-th filter is $f_n(t)$, and the original signal is approximated as

$$f(t) \approx \sum_{n=1}^{N} f_n(t). \tag{12.104}$$

Let the impulse response of the n-th filter be

$$g_n(t) = h(t) \cos \omega_n t, \tag{12.105}$$

where the envelope function $h(t)$ is normally the impulse response of a physically-realizable low-pass filter. Then the output of the n-th filter is the convolution of $f(t)$ with $g_n(t)$,

$$\begin{aligned} f_n(t) &= \int_{-\infty}^{t} f(\lambda)h(t-\lambda)\cos\left[\omega_n(t-\lambda)\right]d\lambda \qquad (12.106) \\ &= \Re\left[\exp(j\omega_n t)\int_{-\infty}^{t} f(\lambda)h(t-\lambda)\exp(-j\omega_n\lambda)d\lambda\right]. \end{aligned}$$

The latter integral is a short-time Fourier transform of the input signal $f(t)$, evaluated at radian frequency ω_n. It is the Fourier transform of that part of $f(t)$ which is "viewed" through the sliding time aperture $h(t)$. If we denote the complex value of this transform as $F(\omega_n, t)$, its magnitude is the short-time amplitude spectrum $|F(\omega_n, t)|$, and its angle is the short-time phase spectrum $\phi(\omega_n, t)$. Then

$$f_n(t) = \Re\left[\exp(j\omega_n t)F(\omega_n, t)\right]$$

or

$$f_n(t) = |F(\omega_n, t)|\cos\left[\omega_n t + \phi(\omega_n, t)\right]. \qquad (12.107)$$

Each $f_n(t)$ may, therefore, be described as the simultaneous amplitude and phase modulation of a carrier ($\cos\omega_n t$) by the short-time amplitude and phase spectra of $f(t)$, both evaluated at frequency ω_n.

Experience with channel vocoders shows that the magnitude functions $|F(\omega_n, t)|$ may be band-limited to around 20 to 30 Hz without substantial loss of perceptually-significant detail. The phase functions $\phi(\omega_n, t)$, however, are generally not bounded; hence they are unsuitable as transmission parameters. Their time derivatives $\dot{\phi}(\omega_n, t)$, on the other hand, are more well-behaved, and may be band-limited and used to advantage in transmission. To within an additive constant, the phase functions can be recovered from the integrated (accumulated) values of the derivatives. One practical approximation to $f_n(t)$ is, therefore,

$$\tilde{f}_n(t) = |F(\omega_n, t)|\cos\left[\omega_n t + \tilde{\phi}(\omega_n, t)\right], \qquad (12.108)$$

where

$$\tilde{\phi}(\omega_n, t) = \int_0^t \dot{\phi}(\omega_n, t)dt.$$

The expectation is that loss of the additive phase constant will not be unduly deleterious.

Reconstruction of the original signal is accomplished by summing the outputs of n oscillators modulated in phase and amplitude. The oscillators are set to the nominal frequencies ω_n, and they are simultaneously phase and amplitude modulated from band-limited versions of $\dot{\phi}(\omega_n, t)$ and $|F(\omega_n, t)|$. The synthesis operations are diagrammed in Fig. 12.18.

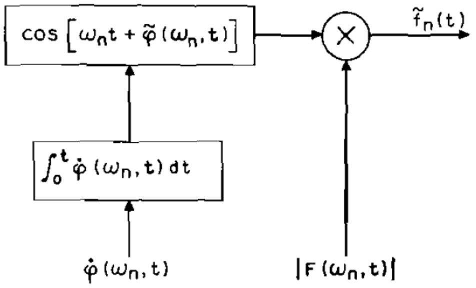


Figure 12.18: Speech synthesis from short-time amplitude and phase-derivative spectra. (After (Golden [1966]))

These analysis-synthesis operations may be viewed in an intuitively appealing way. The conventional channel vocoder separates vocal excitation and spectral envelope functions. The spectral envelope functions of the conventional vocoder are the same as those described here by $|F(\omega, t)|$. The excitation information, however, is contained in a signal which specifies voice pitch and voiced-unvoiced (buzz-hiss) excitation. In the phase vocoder, when the number of channels is reasonably large, information about excitation is conveyed primarily by the $\dot{\phi}(\omega_n, t)$ signals. At the other extreme, with a small number of broad analyzing channels, the amplitude signals contain more information about the excitation, while the $\dot{\phi}$ phase signals tend to contain more information about the spectral shape. Qualitatively, therefore, the number of channels determines the relative amounts of excitation and spectral information carried by the amplitude and phase signals. If good quality and natural transmission are requisites, the indications are that the $\dot{\phi}(\omega, t)$ signals require about the same channel capacity as the spectrumenvelope information. This impression seems not unreasonable in view of experience with voice quality in vocoders.

A complete phase vocoder analyzer and synthesizer has been simulated on a digital computer. In the analyzer, the amplitude and phase spectra are computed by forming the real and imaginary parts of the complex spectrum

$$F(\omega_n, t) = a(\omega_n, t) - b(\omega_n, t),$$

where

$$a(\omega_n, t) = \int_{-\infty}^{t} f(\lambda) h(t-\lambda) \cos \omega_n \lambda d\lambda$$

and

$$b(\omega_n, t) = \int_{-\infty}^{t} f(\lambda) h(t-\lambda) \sin \omega_n \lambda d\lambda \qquad (12.109)$$

Then,

$$|F(\omega_n, t)|^2 = (a^2 + b^2)^{\frac{1}{2}}$$

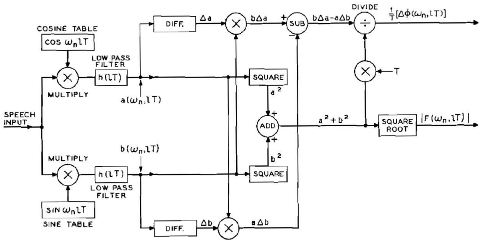


Figure 12.19: Programmed analysis operations for the phase vocoder. (After (Golden [1966]))

and

$$\dot{\phi}(\omega_n, t) = \left(\frac{\dot{a}b - \dot{b}a}{a^2 + b^2} \right)$$

The computer, of course, deals with sampled-data equivalents of these quantities. Transforming the real and imaginary parts of (12.109) into discrete form for programming yields

$$a(\omega_n, mT) = T \sum_{l=0}^{m} f(lT) \left[\cos \omega_n lT\right] h(mT - lT) \tag{12.110}$$

$$b(\omega_n, mT) = T \sum_{l=0}^{m} f(lT) \left[\sin \omega_n lT\right] h(mT - lT),$$

where T is the sampling interval. In the simulation, $T = 10^{-4}$ sec. From these equations, the difference values are computed as

$$\Delta a = a\left[\omega_n, (m+1)T\right] - a\left[\omega_n, mT\right] \tag{12.111}$$

and

$$\Delta b = b\left[\omega_n, (m+1)T\right] - b\left[\omega_n, mT\right]$$

The magnitude function and phase derivative, in discrete form, are computed (12.110) and (12.111) as

$$\left|F\left[\omega_n, mT\right]\right| = \left(a^2 + b^2\right)^{\frac{1}{2}} \tag{12.112}$$

$$\frac{\Delta\phi}{T}\left[\omega_n, mT\right] = \frac{1}{T} \frac{(b\Delta a - a\Delta b)}{a^2 + b^2}.$$

Fig. 12.19 shows a block diagram of a single analyzer channel as realized in the program. This block of coding is required for each channel.

In the simulation, a sixth-order Bessel filter is used for the $h(lT)$ window. The simulation uses 30 channels ($N = 30$) and $\omega_n = 2\pi n(100)$ rad/sec. The equivalent pass bands of the analyzing filters overlap at their 6 dB down points, and a total spectrum range of 50 to 3050 Hz is analyzed.

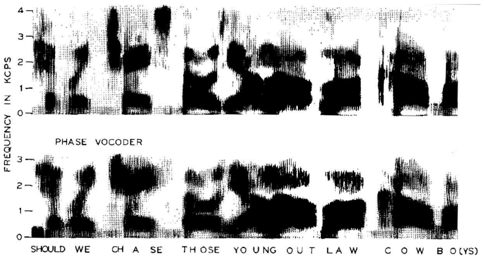


Figure 12.20: Speech transmitted by the phase vocoder. The transmission bandwidth is one-half the original signal bandwidth. Male speaker: "Should we chase those young outlaw cowboys." (After (Golden [1966]))

Programmed low-pass filtering is applied to the amplitude and phase difference signals as defined by Fig. 12.19. Simulation of the whole system is completed by the synthesis operations for each channel performed according to

$$f_n(mT) = |F(\omega_n, mT)| \cos \left[\omega_n mT + T \sum_{l=0}^{m} \frac{\Delta\phi(\omega_n, lT)}{T} \right]. \tag{12.113}$$

Adding the outputs of the n individual channels, according to (12.104), produces the synthesized speech signal.

As part of the simulation, identical (programmed) low-pass filters were applied to the $|F(\omega_n, lT)|$ and $(l/T)[\Delta\phi(\omega_n, lT)]$ signals delivered by the coding block shown in Fig. 12.19. These low-pass filters are similar to the $h(lT)$ filters except they are fourth-order Bessel designs. The cut-off frequency is 25 Hz, and the response is -7.6 dB down at this frequency. This filtering is applied to the amplitude and phase signals of all 30 channels. The total bandwidth occupancy of the system is therefore 1500 Hz, or a band reduction of 2:1.

After band-limitation, the phase and amplitude signals are used to synthesize an output according to (12.113). The result of processing a complete sentence through the programmed system is shown by the sound spectrograms in Fig. 12.20[4]. Since the signal band covered by the analysis and synthesis is 50 to 3050 Hz, the phase-vocoded result is seen to cut off at 3050 Hz. In this example, the system is connected in a "back-to-back" configuration, and the band-limited channel signals are not multiplexed.

Comparison of original and synthesized spectrograms reveals that formant details are well preserved and pitch and voiced-unvoiced features are retained to perceptually significant accuracy. The quality of the resulting signal considerably surpasses that usually associated with conventional channel vocoders.

A frequency-divided signal may be synthesized by division of the $[\omega_n t + \int \dot{\phi}_n dt]$ quantities by some number q. This frequency-divided synthetic signal may be essentially restored to its original spectral position by a time speed-up of q. Such a

---

[4] The input speech signal is band limited to 4000 Hz. It is sampled at 10000 Hz and quantized to 12 bits. It is called into the program from a digital recording prepared previously.

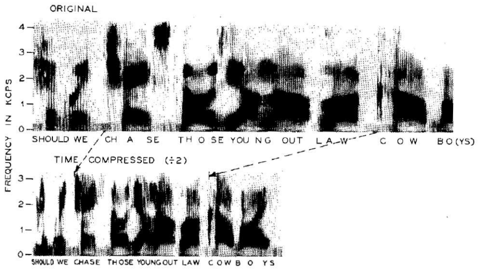


Figure 12.21: Phase vocoder time compression by a factor of 2. Male speaker

speed-up can be accomplished by recording at one speed and replaying q-times faster. The result is that the time scale is compressed and the message, although spectrally correct, lasts $1/q$-th as long as the original. An example of a 2:1 frequency division and time speed-up is shown by the sound spectrograms in Fig. 12.21.

Time-scale expansion of the synthesized signal is likewise possible by the frequency multiplication $q[\omega_n t + \int \dot{\phi}_n dt]$; that is, by recording the frequency-multiplied synthetic signal and then replaying it at a speed q-times slower. An example of time-expanded speech is shown by the spectrograms in Fig. 12.22.

An attractive feature of the phase vocoder is that the operations for expansion and compression of the time and frequency scales can be realized by simple scaling of the phase-derivative spectrum. Since the frequency division and multiplication factors can be non-integers, and can be varied with time, the phase vocoder provides an attractive tool for studying non-uniform alterations of the time scale (Hanauer and Schroeder [1966]).

A number of multiplexing methods may be used for transmission. Conventional space-frequency and time-division methods are obvious techniques. A “self multiplexing” method is also possible in which, say, a two-to-one frequency-divided synthetic signal is transmitted over an analog channel of t the original signal bandwidth. Re-analysis, frequency expansion and synthesis at the receiver recovers the signal[5]. Further, at least one digital implementation of the phase vocoder has been made. The phase and amplitude functions were sampled, quantized and framed for digital transmission at digital rates of 9600 bits/sec and 7200 bits/sec. These transmission rates were compared in listening tests to the same signal coded as log-PCM. The results showed the digital phase vocoder to provide a signal quality comparable to log-PCM at bit rates two to three times higher (Carlson [1968]).

---

[5] The greatest number q by which the ω_n and $\dot{\phi}_n$’s may be divided is determined by how distinct the side-bands about each ω_n/q remain, and by how well each $\dot{\phi}_n/q$ and $|F_n|$ may be retrieved from them. Practically, the greatest number appears to be about 2 or 3 if transmission of acceptable quality is to be realized.

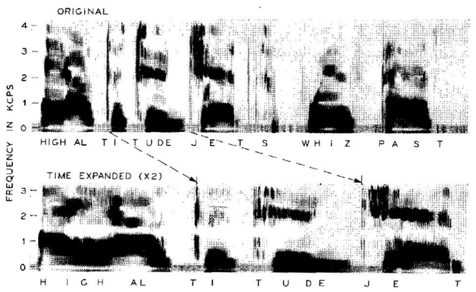


Figure 12.22: Phase vocoder time expansion by a factor of 2. Female speaker

12.4.5 Linear Transformation of Channel Signals

A related approach attempts to discover the dependence among the channel signals and to eliminate this redundancy in a smaller number of signals (Kramer and Mathews [1956]). For n-channel signals, a set of m signals, where $m \leq n$, are formed which are a linear combination of the original n. The coefficients of the linear transformation constitute an $(m \times n)$ matrix of constants. The transformation matrix is realized practically with an $(m \times n)$ array of fixed resistors. Decoding of the m signals to retrieve an approximation to the original n is also accomplished by a linear transformation, namely, the transpose of the $(m \times n)$ matrix. The coefficients of the transformation are obtained to minimize the mean square difference between the original n signals and the reconstructed n signals.

The technique was applied to the spectrum signals of a 16-channel vocoder (i.e., $n = 16$). For a reduction to $m = 6$, it was reported that the output was almost completely understandable, although quality was substantially less than that of the 16-channel vocoder. For $m = 10$, the quality was judged to be bettcr than existing, conventional 10-channel vocoders. In the latter condition, the additional saving in channel capacity is estimated to be in the ratio of 3 to 2.

Another related study used a Hadamard matrix transformation to reduce the redundancy among the channel signals of a 16-channel vocoder (Crowther and Rader [1966]). The Hadamard transformation produces unit-weight linear combinations of the channel signals. It therefore requires no multiplications, but only additions and subtractions. This technique, implemented digitally in a computer, was applied to two different 16-channel vocoders. The results showed that the quality provided by the vocoders when digitized for 4000 bits/sec could be retained in the Hadamard transformation for a data rate as low as J 650 bits/sec. The Hadamard transformation is therefore suggested as a simple, useful means for improving the quality of low bit-rate vocoders (Crowther and Rader [1966]).

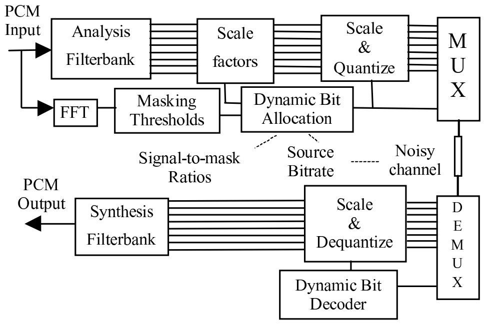


Figure 12.23: Structure of a perceptual subband speech coder (Tang *et al.* [1997]).

12.4.6 Sub-Band Coder

In subband coding, an analysis filterbank is first used to filter the signal into a number of frequency bands and then bits are allocated to each band by a certain criterion. Because of the difficulty in obtaining high-quality speech at low-bit rates using subband coding schemes, these techniques have been mostly used for wideband medium-to-high bit rate speech coders and for audio coding.

An audio signal may only have energy in a small number of frequency bands; all other frequency bands may have very low energy. It is usually possible to use fewer bits, for any given perceptual quality level, by filtering the signal into K sub-bands, downsampling each sub-band by a factor of K, and then allocating bits among the bands in order to minimize mean-squared error. For example, G.722 is a standard in which ADPCM speech coding occurs within two subbands, and bit allocation is set to achieve 7 kHz audio coding at rates of 64 kbps or less.

In (Cox *et al.* [1988, 1991], Gould *et al.* [1993]) subband coding is proposed as a flexible scheme for robust speech coding. A speech production model is not used, ensuring robustness to speech in the presence of background noise, and to non-speech sources. High-quality compression can be achieved by incorporating masking properties of the human auditory system (Jayant *et al.* [1993], Tang *et al.* [1997]). In particular, (Tang *et al.* [1997]) presents a scheme for robust, high-quality, scalable, and embedded speech coding. Figure 12.23 illustrates the basic structure of the coder. Dynamic bit allocation and prioritization and embedded quantization are used to optimize the perceptual quality of the embedded bitstream, resulting in little performance degradation relative to a non-embedded implementation. A subband spectral analysis technique was developed that substantially reduces the complexity of computing the perceptual model.

The encoded bitstream is embedded, allowing the coder output to be scalable from high quality at higher bit rates, to lower quality at lower rates, supporting a wide range of service and resource utilization. The lower bit-rate representation is obtained simply through truncation of the higher bit-rate representation. Since source-rate adaptation is performed through truncation of the encoded stream, interaction with the source coder is not required, making the coder ideally suited for rate adaptive communication systems.

Almost every sub-band coder involves the following steps:

1. Filter $x[n]$ into sub-band signals $x_k[n]$. $x_k[n]$, the kth sub-band signal, is given by
$$x_k[n] = h_k[n] * x[n], \quad X_k(\omega) = H_k(\omega)X(\omega) \tag{12.114}$$
The sub-band filters are almost always created by modulating a prototype lowpass filter $h[n]$ in such a way that the resulting filters uniformly cover the frequency range from 0 to π:
$$h_k[n] = 2h[n]\cos\left(\frac{\pi(2k+1)}{2K}n + \phi_k\right) \tag{12.115}$$
$$H_k(\omega) = e^{-j\phi_k}H\left(\omega + \frac{\pi(2k+1)}{2K}\right) + e^{j\phi_k}H\left(\omega - \frac{\pi(2k+1)}{2K}\right) \tag{12.116}$$
If the prototype lowpass filter $h[n]$ has a cutoff frequency of $\omega_c = \pi/2K$, then the sub-band filter $h_k[n]$ will have a passband of $[\frac{\pi k}{K}, \frac{\pi(k+1)}{K}]$. The filters cover the entire spectrum from $-\pi$ to π, and they have no intentional overlap.

2. The signals $x_k[n]$ are downsampled by a factor of K, creating signals $d_k[n]$:
$$d_k[n] = x_k[Kn] \tag{12.117}$$
$$D_k(\omega) = \frac{1}{K}\sum_{i=0}^{K-1} X_k\left(\frac{\omega - 2\pi i}{K}\right) \tag{12.118}$$
$d_k[n]$ are a set of K different signals, each sampled at a sampling rate of F_s/K samples per second, for a total of F_s samples/second. Sub-band coding provides a lower bit rate for the same audio quality if and only if $d_k[n]$ can be quantized with fewer average bits/sample than $x[n]$.

3. In frame number f, the coder transmits integer code words representing the clipping threshold parameters $T_Q(f,k)$, the bit rates $B(f,k)$, and the samples of signal $d_k[n]$. The decoder uses these parameters to synthesize $\hat{d}_k[n]$.

4. $\hat{d}_k[n]$ is upsampled to create:
$$v_k[n] = \begin{cases} \hat{d}_k[n/K] & n = \text{ multiple of } K \\ 0 & \text{otherwise} \end{cases} \tag{12.119}$$
$$V_k(\omega) = \hat{D}_k(K\omega) \tag{12.120}$$

5. $V_k(\omega)$ is filtered using a sub-band filter $G_k(\omega)$ in order to get rid of all of the aliasing information outside of the target sub-band of $[\frac{\pi k}{K}, \frac{\pi(k+1)}{K}]$. The filtered signal should be as nearly identical as possible to $x_k[n]$, thus we write it as $\hat{x}_k[n]$:

$$\hat{x}_k[n] = g_k[n] * v_k[n], \quad \hat{X}_k(\omega) = G_k(\omega)V_k(\omega) \tag{12.121}$$

   $G_k(\omega)$ must have a passband of $[\frac{\pi k}{K}, \frac{\pi(k+1)}{K}]$, identical to the passband of $H_k(\omega)$. One common choice is

$$g_k[n] = Kh_k[-n] \tag{12.122}$$

$$G_k(\omega) = KH_k^*(\omega) = e^{j\phi_k}H^*\left(\omega + \frac{\pi(2k+1)}{2K}\right) + e^{-j\phi_k}H^*\left(\omega - \frac{\pi(2k+1)}{2K}\right) \tag{12.123}$$

6. Finally, all of the reconstructed sub-band signals are added together to produce

$$\hat{x}[n] = \sum_{k=0}^{K-1} \hat{x}_k[n] \tag{12.124}$$

Avoiding Aliasing: Ideal Bandpass Filters

The goal of minimum-MSE audio coding is to reproduce $x[n]$ as accurately as possible, that is, to minimize $E[(\hat{x}[n] - x[n])^2]$. A sub-band coder introduces two types of error into the synthesized signal. First, error may be introduced by quantization. Second, even if there is no quantization ($\hat{d}_k[n] = d_k[n]$), error may be introduced by aliasing.

Downsampling a signal results in aliasing (equation 12.118). Aliasing can be controlled, but only if $H_k(\omega)$ and $G_k(\omega)$ are very carefully designed in order to minimize aliasing.

Suppose that there is no quantization, i.e., $\hat{d}_k[n] = d_k[n]$. By combining all of the steps in the algorithm, it is possible to write the output in terms of the input:

$$\hat{X}_k(\omega) = \frac{G_k(\omega)}{K} \sum_{i=0}^{K-1} H_k\left(\omega - \frac{2\pi i}{K}\right) X\left(\omega - \frac{2\pi i}{K}\right) \tag{12.125}$$

$$\hat{X}(\omega) = \sum_{k=1}^{K-1} \hat{X}_k(\omega) \tag{12.126}$$

"Aliasing" is any difference between $\hat{X}(\omega)$ and $X(\omega)$ under the condition that $\hat{d}_k[n] = d_k[n]$. The goal of filter design, for a sub-band audio coder, is to design $H_k(\omega)$ and $G_k(\omega)$ so that the signal $\hat{X}(\omega)$ in equation 12.126 is equal to $X(\omega)$.

In order to better understand the aliasing problem, consider an ideal solution:

$$|H_k(\omega)| = \begin{cases} 1 & \frac{\pi k}{K} \le |\omega| < \frac{\pi(k+1)}{K} \\ 0 & \text{otherwise} \end{cases} \tag{12.127}$$

$$G_k(\omega) = KH_k^*(\omega) \tag{12.128}$$

The filters in equations 12.127 and 12.128 result in the following simplification of equation 12.125:

$$\hat{X}_k(\omega) = |H_k(\omega)|^2 X(\omega) \tag{12.129}$$

By the definition of $|H(\omega)|$, therefore, $\hat{X}(\omega) = X(\omega)$. Unfortunately, the filters in equation 12.127 can not be used in any practical system, because the impulse response $h_k[n]$ is infinite in length.

Pseudo Quadrature Mirror Filters (PQMF)

It turns out that, even though though it's impossible to get $\hat{x}_k[n] = x_k[n]$ using non-ideal filters, it is nevertheless possible to design $H_k(\omega)$ so that $\hat{x}[n] = x[n]$. Consider the general formula for $\hat{X}(\omega)$, assuming no quantization ($\hat{d}_k[n] = d_k[n]$), and assuming that $G_k(\omega) = KH_k^*(\omega)$:

$$\hat{X}(\omega) = \sum_{k=0}^{K-1} \hat{X}_k(\omega) \tag{12.130}$$

$$\hat{X}(\omega) = \sum_{k=1}^{K-1} H_k^*(\omega) \sum_{i=0}^{K-1} H_k\left(\omega - \frac{2\pi i}{K}\right) X\left(\omega - \frac{2\pi i}{K}\right) \tag{12.131}$$

Clearly, the only way that we can set $\hat{X}(\omega) = X(\omega)$ is by somehow getting rid of the aliased copies $X(\omega - 2\pi i/K)$; the terms with $i \neq 0$ must not be allowed to "leak through" to the output. There are two ways to get rid of the aliased copies: (1) use an ideal bandpass filter, or (2) use the aliased terms from one band to cancel out the aliased terms from the neighboring band. Filters designed so that the aliasing from one band cancels the aliasing from the neighboring band are called "pseudo quadrature mirror filters (PQMF)." This section describes how to build PQMF.

Suppose that $H_k(\omega)$ and $G_k(\omega)$ both have "transition bands" of width roughly $\pi/2K$ on either side of their passbands, and that $H_k(\omega) \approx 0$ outside of the transition band. If the passband of $H_k(\omega)$ is $[\frac{\pi k}{K}, \frac{\pi(k+1)}{K}]$, then $H_k(\omega)$ should be very close to zero outside the frequency range $[\frac{\pi(2k-1)}{2K}, \frac{\pi(2k+3)}{K}]$.

Now, under this condition, consider equation 12.131 at frequencies $\omega \approx \frac{\pi k_0}{K}$, for some particular k_0. Because of the stop-band condition on $G_k(\omega)$, only two bands contribute aliasing in the vicinity of $\frac{\pi k_0}{K}$: band number $k_0 - 1$, and band number k_0. Furthermore, because of the stop-band condition on $H_k(\omega)$, only two of the aliasing terms have energy near $\omega = \frac{pik_0}{K}$: the $i = 0$ term, and the $i = k_0$ term. Near $\omega \approx \frac{2\pi k_0}{K}$, therefore, equation 12.131 reduces to a four-term sum:

$$\hat{X}(\omega) = \left(|H_{k_0-1}(\omega)|^2 + |H_{k_0}(\omega)|^2\right) X(\omega) + \tag{12.132}$$

$$\left(H_{k_0-1}^*(\omega) H_{k_0-1}\left(\omega - \frac{2\pi k_0}{K}\right) + H_{k_0}^*(\omega) H_{k_0}\left(\omega - \frac{2\pi k_0}{K}\right)\right) X\left(\omega - \frac{2\pi k_0}{K}\right) \tag{12.133}$$

$$\text{for } \omega \approx \frac{\pi k_0}{K}, \quad \text{or more specifically,} \quad \frac{\pi(2k_0-1)}{2K} < \omega < \frac{\pi(2k_0+1)}{2K} \tag{12.134}$$

Suppose now that $H_k(\omega)$ is created by modulating a prototype lowpass filter, as shown in equation 12.116, reproduced here:

$$H_k(\omega) = e^{-j\phi_k} H\left(\omega + \frac{\pi(2k+1)}{2K}\right) + e^{j\phi_k} H\left(\omega - \frac{\pi(2k+1)}{2K}\right) \tag{12.135}$$

Suppose also that $H(\omega)$, the lowpass filter, is real-valued, meaning that $h[n]$ is a non-causal even function. Then the two aliased terms in equation 12.133 are

$$H^*_{k_0-1}(\omega) H_{k_0-1}\left(\omega - \frac{2\pi k_0}{K}\right) = e^{-2j\phi_{k_0-1}} H\left(\omega - \frac{\pi(2k_0-1)}{2K}\right) H\left(\omega - \frac{\pi(2k_0+1)}{2K}\right) \tag{12.136}$$

and

$$H^*_{k_0}(\omega) H_{k_0}\left(\omega - \frac{2\pi k_0}{K}\right) = e^{-2j\phi_{k_0}} H\left(\omega - \frac{\pi(2k_0+1)}{2K}\right) H\left(\omega - \frac{\pi(2k_0-1)}{2K}\right) \tag{12.137}$$

The only difference between these two terms is the phase term. These two aliasing terms cancel if

$$H^*_{k_0-1}(\omega) H_{k_0-1}\left(\omega - \frac{2\pi k_0}{K}\right) = -H^*_{k_0}(\omega) H_{k_0}\left(\omega - \frac{2\pi k_0}{K}\right) = \tag{12.138}$$

which is accomplished if

$$e^{-2j\phi_{k_0}} = -e^{-2j\phi_{k_0-1}} \tag{12.139}$$

Therefore the aliasing terms completely cancel out if the phase difference between neighboring bands is an odd multiple of $\frac{\pi}{2}$, i.e.

$$\phi_k - \phi_{k-1} = \frac{(2r-1)\pi}{2} \quad \text{for integer } r \tag{12.140}$$

Equation 12.140 says that the modulating sinusoids used to construct neighboring bands should be 90 degrees out of phase. Two sinusoids that are 90 degrees out of phase are said to be “in quadrature” with respect to one another. Except for the quadrature term, filters $H_{k-1}(\omega)$ and $H_k(\omega)$ are mirror images of one another reflected around the frequency $\frac{\pi k}{K}$, thus they are called “pseudo-quadrature mirror filters” (PQMF).

A useful choice for the phases is

$$\phi_k = \frac{(2k+1)r\pi}{4} \tag{12.141}$$

for some odd-valued integer r.

Implementation Using the FFT

There are two methods for implementing a sub-band filtering algorithm using the FFT: one method is more intuitively meaningful, and one is more computationally efficient. Let's start with the intuitively meaningful method.

The sub-band signals $x_k[n]$ are computed as

$$x_k[n] = \sum_{p=-\infty}^{\infty} x[n-p]h_k[p] \tag{12.142}$$

Substituting in the value of $h_k[m]$ gives

$$x_k[n] = 2\sum_{p=-\infty}^{\infty} x[n-p]h[p]\cos\left(\frac{\pi(2k+1)}{2K}p + \phi_k\right) \tag{12.143}$$

Remember that we only need the values of $x_k[n]$ at integer multiples of K, thus equation 12.143 only needs to be computed at samples number $n = fK$ for integer values of f. Let us also define a "window function" $w[n] = h[-n]$, and substitute $m = -p$, then we get that

$$x_k[fK] = 2\sum_{m=-\infty}^{\infty} x[fK+m]w[m]\cos\left(-\frac{\pi(2k+1)}{2K}m + \phi_k\right) \tag{12.144}$$

Remember that the short-time Fourier transform (STFT) was defined as

$$X_f(\omega) = \sum_{m=-\infty}^{\infty} x[fK+m]w[m]e^{-j\omega m} \tag{12.145}$$

Comparing equations 12.144 and 12.145, we see that

$$x_k[fK] = 2\Re\left\{e^{j\phi_k}X_f\left(\frac{2\pi(2k+1)}{4K}\right)\right\} \tag{12.146}$$

In other words, $x_k[fK]$ are the real parts of the odd-numbered frequency samples of the $4K$-point STFT, computed using a window of $w[n] = h[-n]$, and phase-shifted by ϕ_k. All of the details of this transformation should seem perfectly reasonable:

- Only the odd-numbered frequency samples are kept. A $4K$-point DFT creates frequency samples at $\omega = 0, \frac{\pi}{2K}, \frac{\pi}{K}, \frac{3\pi}{2K}, \frac{2\pi}{K}, \ldots$. Of these frequency samples, the even-valued samples correspond to band edges ($0, \frac{\pi}{K}$, etc.) while the odd-valued samples correspond to the center frequencies of the sub-bands ($\frac{\pi}{2K}, \frac{2\pi}{2K}$, etc.). It seems intuitively reasonable that the PQMF computation should keep the STFT samples located at the centers of the band-pass filters, and throw away the STFT samples located at the edges of the band-pass filters.

- Only the real part of the STFT is retained because the filters are created by modulating $h[n]$ using cosines rather than using complex exponentials. For the same reason, it is only necessary to keep the frequencies between $0 \leq \omega \leq \pi$ ($0 \leq k \leq K-1$): the STFT is conjugate-symmetric, so $\Re(X(\omega)) = \Re(X(2\pi - \omega))$.
- The window is $w[n] = h[-n]$ because every filter is a modulated copy of $h[n]$.
- The phase shift ϕ_k equals the phase of the modulating cosine.

Equation 12.146 is the most intuitively meaningful way to compute $x_k[fK]$, but it is not the most computationally efficient method. The most computationally efficient method is to pre-modulate the windowed signal using a complex exponential modulator $e^{-\frac{j\pi m}{2K}}$. The complex-valued time domain signal $x[fK+m]w[m]e^{-\frac{j\pi m}{2K}}$ is then Fourier transformed using a length-$2K$ FFT:

$$x_k[fK] = 2\Re\left\{ e^{j\phi_k} \sum_{m=-\infty}^{\infty} x[fK+m]w[m]e^{-\frac{j\pi m}{2K}} e^{-\frac{j2\pi km}{2K}} \right\} \tag{12.147}$$

Just as an FFT is the most efficient method for computing the sub-band filtered signals, an FFT is also the most efficient method for synthesizing the output signal. Consider:

$$\hat{x}[n] = \sum_{k=0}^{K-1} \hat{x}_k[n] = \sum_{k=0}^{K-1} \sum_{m=-\infty}^{\infty} v_k[m] g_k[n-m] \tag{12.148}$$

If there is no quantization, then $v_k[fK] = x_k[fK]$ for integer values of f, and $v[n] = 0$ otherwise. Substituting in $g_k[m] = Kh_k[-m]$, we find that equation 12.148 can be expanded to

$$\hat{x}[n] = 2K \sum_{f=-\infty}^{\infty} h[fK-n] \sum_{k=0}^{K-1} x_k[fK] \cos\left(\frac{\pi(2k+1)(fK-n)}{2K} + \phi_k \right) \tag{12.149}$$

There are a number of ways that equation 12.149 can be rewritten in terms of an FFT. Here is one: define a short-time phase-shifted spectrum $Y_f\left(\frac{2\pi k}{2K}\right)$ as

$$Y_f\left(\frac{2\pi k}{2K}\right) = \begin{cases} e^{j\phi_k} x_k[fK] & 0 \leq k \leq K-1 \\ 0 & \text{otherwise} \end{cases} \tag{12.150}$$

Second, define the inverse STFT of $Y_f(\omega)$ to be

$$y_f[m] = \frac{1}{K} \sum_{k=0}^{2K-1} Y_f\left(\frac{2\pi k}{2K}\right) e^{\frac{j2\pi km}{2K}} \tag{12.151}$$

Third, define the short-time signal $\hat{x}_f[m]$ to be the windowed real part of the modulated inverse:

$$\hat{x}_f[m] = 2K^2 h[m] \Re\left\{ e^{\frac{j\pi m}{2K}} y_f[m] \right\} \tag{12.152}$$

Finally, create $\hat{x}[n]$ by flipping, shifting, and adding the short-time signals:

$$\hat{x}[n] = \sum_{f=-\infty}^{\infty} \hat{x}_f[fK - n] \tag{12.153}$$

We have not yet discussed the length of the window $w[n]$. Recall that $w[n] = h[-n]$, therefore the length of the window must equal the length of the prototype filter $h[n]$. Equations 12.147 and 12.152 use length-$2K$ Fourier transforms. In general, it may be difficult to achieve a narrow transition band, as required by the PQMF theory, using a prototype filter $h[n]$ as short as only $2K$ samples. The MPEG-I and MPEG-II coding standards use $K = 32$ filters, with a prototype filter length of 512 samples. Fortunately, the length of the window really doesn't effect the computational complexity of equations 12.147 and 12.152 at all. In order to use equation 12.147 using a very long window, the long windowed signal should be modulated by $e^{-\frac{j\pi m}{2K}}$, and then time-aliased down to a signal of length $2K$ prior to applying the FFT. Similarly, in order to use equation 12.152, a $2K$-sample inverse FFT operation is used to compute an inverse signal of length $2K$; the resulting signal is then repeated periodically, modulated using the complex exponential $e^{\frac{j\pi m}{2K}}$, and its real part is windowed using $h[m]$.

Malvar has shown that a PQMF filter bank with a window length of $2K$ samples and phase terms of $\phi_k = -\frac{3\pi}{2}(k + \frac{1}{2})$ cancels out not just the aliasing between neighboring pairs of sub-bands, but also the aliasing between every pair of sub-bands.

Sub-Band Quantization

If audio is quantized in sub-bands, the output signal is the sum of the reconstructed signals in all different bands

$$\hat{x}[n] = \sum_{k=1}^{K} \hat{x}_k[n] \tag{12.154}$$

The true signal is the sum of all of the (unknown) true sub-band signals, i.e.

$$x[n] = \sum_{k=1}^{K} x_k[n] \tag{12.155}$$

Therefore the error is just the sum of the sub-band errors

$$e[n] = \sum_{k=1}^{K} e_k[n] \tag{12.156}$$

In the case of ideal bandpass filters, the sub-band signals occupy completely independent frequency bands, thus they are completely uncorrelated. In this case the total error power is just the sum of the sub-band error powers, i.e.

$$\mathrm{SNR} = \frac{\sum_{k=0}^{K-1}\sum_{n=0}^{N-1} x_k^2[n]}{\sum_{k=0}^{K-1}\sum_{n=0}^{N-1} e_k^2[n]} \tag{12.157}$$

It is possible to dynamically code the clipping threshold, T_Q, and transmitting it as “side information” once per frame. In a sub-band coder, the “side information” usually includes two types of information: a clipping threshold for each band, $T_Q(k)$, and a different number of bits used to code the audio in each band, $B(k)$, for $0 \le k \le K-1$. Suppose that the clipping threshold is always set to $T_Q(k) = \sqrt{3}R\sigma_x(k)$, where R is some fixed constant “safety ratio” chosen in advance by the system designer, and $\sigma_x^2(k)$ is the power of the sub-band signal $x_k[n]$. Then, in this case, the total SNR of a frame of audio is

$$\mathrm{SNR} = \frac{\sum_{k=0}^{K-1} T_Q^2(k)}{R^2\sum_{k=0}^{K-1} T_Q^2(k)4^{-B(k)}} \tag{12.158}$$

The “minimum mean-squared error” bit allocation strategy is the strategy that maximizes equation 12.158, subject to a fixed-bit-rate constraint which says that the average number of bits per sample of the audio signal must be B_{ave}:

$$\sum_{k=0}^{K-1} B(k) = KB_{ave} \tag{12.159}$$

Using Lagrange multipliers to minimize equation 12.158 subject to the constraint in equation 12.159 yields the following minimum-MSE bit allocation strategy:

$$B(k) = KB_{ave}\left(\frac{\log(T_Q^2(k))}{\sum_{i=0}^{K-1}\log(T_Q^2(i))}\right) \tag{12.160}$$

Equation 12.160 says that the number of bits allocated to any given channel should be proportional to the channel’s log power, $\log T_Q^2(k)$. The base of the logarithm doesn’t matter, as long as the numerator and denominator are computed using the same base. Notice that equation 12.160 is equivalent to the claim that the number of quantization levels, $Q(f,k)$, should be proportional to $T_Q(f,k)$. Equivalently, the quantization step size $\Delta(f,k)$, and the error variance $\Delta^2/12$, should be made independent of the channel number. This is an intuitively reasonable result: total MSE is minimized when every channel has the same MSE.

In sub-band coding, the coder consists of the following steps.

1. The audio signal $x[n]$ is filtered through bandpass filter $H_k(\omega)$ to create sub-band signal $x_k[n]$, $0 \le k \le K-1$.

2. $x_k[n]$ is critically downsampled by K in order to create $v_k[n]$.

3. Once per frame, the maximum value of $v_k[n]$ is computed. The clipping threshold $T_Q(k)$ must be chosen from a codebook of possible values, known to both the transmitter and receiver. Typically, $T_Q(k)$ is set to the value, from this codebook, that is closest to the maximum amplitude of $v_k[n]$. Once all of the clipping thresholds have been chosen, the number of bits in each band is determined using Eq. 12.160, or using an algorithm such as water-filling bit allocation (next section). There are K different clipping thresholds per frame, and K different bit allocations; these numbers are transmitted as side information.

4. $v_k[n]$ is quantized, using clipping threshold $T_Q(k)$ and $B(k)$ bits. The code word $q_k[n]$ is transmitted.

The decoder performs the same steps in reverse, i.e.,

1. $\hat{v}_k[n]$ is computed using a dequantizer with maximum amplitude $T_Q(k)$, and with $B(k)$ bits.

2. $\hat{v}_k[n]$ is upsampled by a factor of K, then filtered by $H_k^*(\omega)$ in order to create $\hat{x}_k[n]$:
$$\hat{x}_k[n] = h_k[-n] * \hat{v}_k[n/K] \tag{12.161}$$

3. The sub-band signals are added together to create $\hat{x}[n]$:
$$\hat{x}[n] = \sum_{k=0}^{K-1} \hat{x}_k[n] \tag{12.162}$$

Water-Filling Bit Allocation

Equation 12.160 is impossible to implement in practice, because it specifies a non-integer number of bits per channel. Two methods are commonly used to allocate an integer number of bits to each channel. The first method computes an initial bit allocation by truncating equation 12.160; the extra bits are allocated arbitrarily to any channel until the total is KB_{ave}. This algorithm may be expressed as:

$$B(k) \geq \text{floor}\left(KB_{ave} \left(\frac{\log(T_Q^2(f,k))}{\sum_{i=0}^{K-1} \log(T_Q^2(i))} \right) \right) \tag{12.163}$$

The second common method, called the "water-filling" bit allocation method, achieves the result specified by equation 12.163, but with a slightly more principled distribution of the extra bits. Water-filling bit allocation is based on the observation that noise power, in decibels, is equal to signal power minus $6B$:

$$10 \log_{10} \frac{\Delta^2(k)}{12} = 10 \log_{10} \frac{T_Q^2(k)}{3} - 6B(k) \tag{12.164}$$

Noise power may therefore be minimized using the following algorithm:

1. **Initialize:** Set the initial "water level" in each bin to

$$W(k) = 10\log_{10}\frac{T_Q^2(k)}{3} \tag{12.165}$$

2. **Iterate:** For $b = 1, \ldots, KB_{ave}$,
   - (a) Find the band k^* with maximum $W(k)$.
   - (b) Increment $B(k^*)$.
   - (c) Subtract 6dB from $W(k^*)$.

The final values of $W(k)$ are estimates of the noise power in each channel. The resulting final bit allocation is guaranteed to satisfy equation 12.163.

The rationale behind the water-filling bit allocation strategy is that $W(k)$ is an estimate of the noise power, in decibels, in band k. Notice that when $B(k) = 0$, $W(k)$ is initialized to equal the signal power in band k! A "zero bit quantizer" always creates the signal $\hat{x}[n] = 0$. The error is $e[n] = \hat{x}[n] - x[n] = -x[n]$, thus the quantization noise spectrum of a zero-bit quantizer is the same as the spectrum of the input signal, $x[n]$.

The "water-filling bit allocation" algorithm starts out with zero bits per sub-band, thus the quantization noise power in each band is equal to the signal power in that band. The bits are allocated one at a time. Each bit is allocated to the band with the highest remaining quantization noise power. After each bit is allocated to a band, the quantization noise power in that band is reduced by 6dB. Then the next bit is allocated, and so on, until all of the available bits have been allocated.

Example 12.4.1 Water-Filling Bit Allocation

Consider the following signal:

$$x[n] = 3\cos(0.1\pi n) + \cos(0.6\pi n) \tag{12.166}$$

The autocorrelation and power spectrum of $x[n]$ are given by

$$r_x[n] = \frac{9}{2}\cos(0.1\pi n) + \frac{1}{2}\cos(0.6\pi n) \tag{12.167}$$

$$\begin{aligned} R_x(\omega) &= \frac{9}{2}\pi\left[\delta(\omega - 0.1\pi) + \delta(\omega + 0.1\pi)\right] && (12.168) \\ &\quad + \frac{1}{2}\pi\left[\delta(\omega - 0.6\pi) + \delta(\omega + 0.6\pi)\right] && (12.169) \end{aligned}$$

Suppose that $x[n]$ is filtered into four bands. The signal intensities of the four band-pass filtered signals are

$$I_{x1} = \frac{1}{\pi}\int_0^\pi R_{x1}(\omega)d\omega = \frac{9}{2} \quad 10\log_{10}(\frac{9}{2}) \approx 6dB \tag{12.170}$$

$$I_{x2} = \frac{1}{\pi}\int_0^{\pi} R_{x2}(\omega)d\omega = 0 \quad 10\log_{10}(0) = -\infty dB \tag{12.171}$$

$$I_{x3} = \frac{1}{\pi}\int_0^{\pi} R_{x3}(\omega)d\omega = \frac{1}{2} \quad 10\log_{10}(\frac{1}{2}) \approx -3dB \tag{12.172}$$

At zero bits/sample, $e_m[n] = x_m[n]$, so noise intensity is equal to the signal intensity in each band, $\sigma^2_{e_m} = \sigma^2_{x_m}$.

Suppose that we want to create a coder with an average of 1 bit/sample. The sub-band coder has 4 bands, therefore we have a total of 4 bits to allocate among the 4 bands.

With the intensities given in the previous example, we find that bit #1 is allocated to band 1. The residual noise in band 1 after allocation of this bit is 0dB.

Bit #2 is also allocated to band 1. The remaining noise intensity in band 1, after getting two bits allocated to it, is -6dB.

Band #3 now has the highest residual quantization noise, so band #3 gets bit #3. The quantization noise in this band is reduced by 6dB, to -9dB.

Band #1 now has the highest residual quantization noise (-6dB), so the fourth bit is allocated to band #1. The residual quantization noise in band #1 is now reduced to -12dB.

The resulting bit allocation uses 3 bits to code samples from band #1 (an 8-level quantizer), and 1 bit to code samples from band #3 (a two-level quantizer). Bands #2 and #4 have no bits at all, so they are simply not transmitted; the reconstructed signal in these bands is set to 0.

Signal to Mask Ratio

Thus far, we have been allocating each bit to the sub-band with the highest residual quantization noise energy,

$$I_{xm} = \frac{1}{\pi}\int_{\frac{(m-1)\pi}{32}}^{\frac{m\pi}{32}} R_x(\omega)d\omega = I_{em}|_{0 \text{ Bits}} \tag{12.173}$$

Another option would be to use a psychophysical model to estimate the perceptual noise "audibility" or "loudness" in each band, and to allocate each bit to the band with the highest residual noise audibility:

$$A_{xm} = \frac{1}{\pi}\int_{\frac{(m-1)\pi}{32}}^{\frac{m\pi}{32}} \frac{R_x(\omega)}{M_x(\omega)}d\omega = A_{em}|_{0 \text{ Bits}} \tag{12.174}$$

where $M_x(\omega)$ is the "masking spectrum," an estimate of the level of noise at freqeuncy ω that gets masked by the signal in the same band.

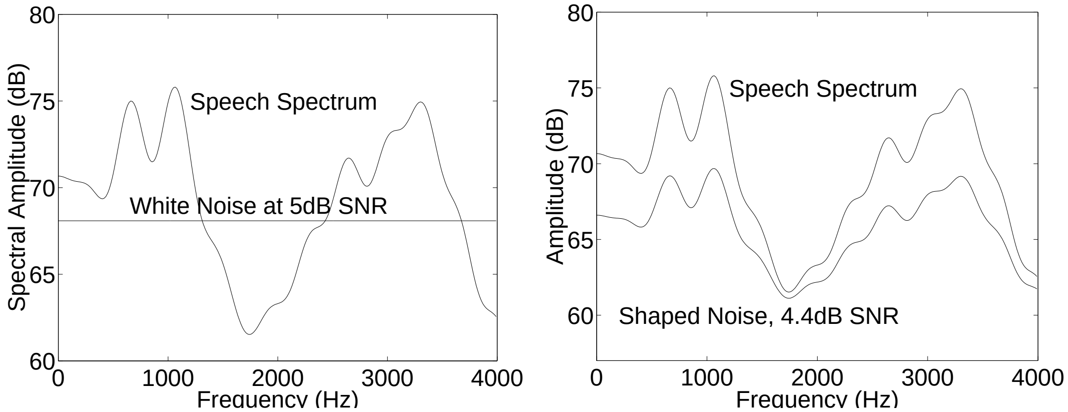


Figure 12.24: White noise at 5dB SNR may be audible, because the noise is louder than the signal in some frequency bands. If the quantization noise is spectrally shaped, with a shape similar to the shape of the speech spectrum, then it may be possible to completely mask the quantization noise so that it is inaudible even at less than 5dB SNR.

The masking spectrum is defined so that noise is audible if and only if

$$\frac{R_e(\omega)}{M_x(\omega)} \geq 1 \quad \text{“Noise To Mask Ratio''} \tag{12.175}$$

There are three types of masking:

1. Masking of one tone by another tone at the same frequency and same time. This type of masking has no name; we can call it “Simultaneous Same-Frequency Masking.”

2. Masking of one tone by another tone at the same time, but at a slightly different frequency. This is called “Simultaneous Masking.”

3. Masking of sounds at one time by sounds at another time is called “Non-simultaneous Masking.”

Example 12.4.2 Simultaneous Same-Frequency Masking

$$x[n] = A\cos(0.1\pi n) \tag{12.176}$$

$$e[n] = B\sin(0.1\pi n) \tag{12.177}$$

$$\hat{x}[n] = x[n] + e[n] \tag{12.178}$$

The signals $x[n]$ and $\hat{x}[n]$ are indistinguishable if the loudness levels of these two tones differ by less than the intensity JND (just-noticeable-difference). The intensity JND is about 1dB, so $x[n]$ and $\hat{x}[n]$ are indistinguishable if

$$10\log_{10}\left(\frac{A^2+B^2}{A^2}\right) < 1dB \tag{12.179}$$

Expanding Eq. 12.179, we find that the noise is masked if

$$\Longleftrightarrow R_e(\omega) < \frac{1}{4}R_x(\omega) \tag{12.180}$$

From the example above, we see that the masking spectrum is given by:

$$M_x(\omega) \approx \frac{1}{4}R_x(\omega) \tag{12.181}$$

$$e[n] \text{ Inaudible If } \frac{\mathrm{R_e}(\omega)}{\mathrm{M_x}(\omega)} < 1 \quad \text{For All } \Omega \tag{12.182}$$

Simultaneous Masking occurs when a tone at one frequency (e.g., $\omega = 0.1\pi$) masks noise at nearby frequencies. Simultaneous masking is possible because of the spread of energy on the basilar membrane: a tone at ω excites every point on the basilar membrane whose characteristic frequency is within one ERB of ω, therefore noise centered at those other frequencies tends to be masked by the strong tone at ω.

Remember that the sound power encoded on a particular auditory nerve fiber is computed by filtering the input power spectrum through the local basilar membrane response. If the nerve fiber is tuned to frequency ω, then the sound power on that nerve in response to input spectrum $R_x(\psi)$ is approximately given by

$$P(\omega) \approx \frac{1}{\pi}\int_0^\pi R_x(\psi)|\Gamma_\omega(\psi)|^2 d\psi \tag{12.183}$$

In Eq. 12.183, $\Gamma_\omega(\psi)$ is the frequency response of the auditory filter centered at center frequency ω. Recall from Chap. 7 that the frequency response of the auditory filter is well represented by the response of a second-order gamma-tone filter,

$$|\Gamma_\omega(\psi)|^2 = \left|\left(\frac{\alpha(\omega)}{(j(\psi-\omega)-\alpha(\omega))}\right)^2 + \left(\frac{\alpha(\omega)}{(j(\psi+\omega)-\alpha(\omega))}\right)^2\right|^2 \tag{12.184}$$

$$\gamma[n] \propto ne^{-\alpha n}\cos(\omega n) \tag{12.185}$$

In Eq. 12.184, the parameter $\alpha(\omega)$ is the 6dB bandwidth of the auditory filter centered at frequency ω, expressed in units of radians/sample. Recall from Chap. 7 that the

6dB bandwidth of a gammatone filter is proportional to its equivalent rectangular bandwidth:

$$\alpha(\omega) = \frac{8}{\pi}\mathrm{ERB}(\omega) \tag{12.186}$$

The equivalent rectangular bandwidth, expressed in radians/sample, was shown by Moore and Glasburg (JASA, 1983) to be

$$\mathrm{ERB}(\omega) \approx 10^{-6}F_s\left(\omega + \frac{2\pi 312}{F_s}\right)\left(\omega + \frac{2\pi 14700}{F_s}\right) \tag{12.187}$$

Quantization error will be inaudible if the difference between the quantized signal and the original is less than about 1dB, i.e.

$$10\log_{10}\left(\frac{\int_0^\pi (R_x(\psi)+R_e(\psi))|\Gamma_\omega(\psi)|^2 d\psi}{\int_0^\pi R_x(\psi)|\Gamma_\omega(\psi)|^2 d\psi}\right) < 1\mathrm{dB} \tag{12.188}$$

Eq. 12.188 is equivalent to

$$\int_0^\pi R_e(\psi)|\Gamma_\omega(\psi)|^2 d\psi < \frac{1}{4}\int_0^\pi R_e(\psi)|\Gamma_\omega(\psi)|^2 d\psi \tag{12.189}$$

Suppose that $e[n]$ is caused by quantization noise with a relatively large number of bits per sample. In that case, $R_e(\omega)$ is relatively flat within a sub-band, so we can use this approximation:

$$\int_0^\pi R_e(\psi)|\Gamma_\omega(\psi)|^2 d\psi \approx R_e(\omega)\int_0^\pi |\Gamma_\omega(\psi)|^2 d\psi = R_e(\omega)\mathrm{ERB}(\omega) \tag{12.190}$$

because the ERB is defined to be the integral of the squared magnitude filter response. Substituting Eq. 12.190 into Eq. 12.188 yields the result that noise is inaudible whenever $R_e(\omega) < M_x(\omega)$, where

$$M_x(\omega) = \frac{1}{4\mathrm{ERB}(\omega)}\int_0^\pi R_x(\psi)|\Gamma_\omega(\psi)|^2 d\psi \tag{12.191}$$

Schroeder proposed a series of approximations that lead to a computationally efficient method for computing the masking spectrum. First, when performing computations in the positive-frequency half of the power spectrum, ignore the negative-frequency half of the auditory filter:

$$|\Gamma_\omega(\psi)|^2 \approx \left|\frac{\alpha(\omega)}{(j(\psi-\omega)-\alpha(\omega))}\right|^4, \quad \psi > 0 \tag{12.192}$$

Second, Schroeder proposed performed computing the masking spectrum in Bark units. Bark-frequency is a nonlinear function, $\beta(\omega)$, defined so that one ERB corresponds to one Bark, i.e.

$$\frac{d\omega}{d\beta} = \mathrm{ERB}(\omega) \tag{12.193}$$

Ignoring the quadratic term in Eq. 12.187, and integrating Eq. 12.193, yields:

$$\beta(\omega) \approx 11 \log \left(1 + \frac{F_s \omega}{640\pi}\right) \tag{12.194}$$

$$\omega(\beta) \approx \frac{640\pi}{F_s} \left(e^{\beta/11} - 1\right) \tag{12.195}$$

The power spectrum of the masking signal, $R_x(\omega)$, can be written in Bark. Because $R_x(\omega)$ is a density rather than a mass function, the mapping from $R_x(\omega)$ to $R_x(\beta)$ needs to take into consideration the nonlinearity, as shown here:

$$R_x(\beta) = \frac{d\omega}{d\beta} R_x(\omega) = \mathrm{ERB}(\omega) R_x(\omega) \tag{12.196}$$

If Eq. 12.196 is applied to the positive-frequency auditory filters (Eq. 12.192), the result turns out to be remarkably simple (by design—after all, this is the reason that Schroeder chose "Bark frequency" as the units in which to work):

$$\Gamma_\beta(\phi) = \Gamma(\phi - \beta) \tag{12.197}$$

where the "smoothing function" $\Gamma(\phi)$ is given by

$$|\Gamma(\phi)|^2 \approx \left| \frac{\frac{8}{\pi}}{(j\phi - \frac{8}{\pi})} \right|^4 \tag{12.198}$$

By plugging Eq. 12.197 into Eq. 12.191, we find that the masking spectrum is computed as

$$M_x(\beta) = \frac{1}{4\mathrm{ERB}(\omega)} \int_0^{\beta(\pi)} R_x(\phi) |\Gamma(\beta - \phi)|^2 d\phi \tag{12.199}$$

Given $M_x(\beta)$, one can then compute the linear-frequency masking spectrum as

$$M_x(\omega) = \frac{d\beta}{d\omega} M_x(\beta(\omega)) = \frac{M_x(\beta(\omega))}{\mathrm{ERB}(\omega)} \tag{12.200}$$

Eq. 12.199 is a convolution in frequency. Eq. 12.199 can lead to two different types of really dramatic computational savings. First, if we assume that the low-frequency components of the signal don't matter very much, then it is possible to write that

$$M_x(\beta) \approx \frac{1}{2\pi} \int_{-\beta(\pi)}^{\beta(\pi)} R_x(\phi) |\Gamma(\beta - \phi)|^2 d\phi = \mathcal{F}\left\{r_x(b) r_\gamma(b)\right\} \tag{12.201}$$

where $r_x(b)$ and $r_\gamma(b)$ are the inverse Fourier transforms of $R_x(\beta)$ and $|\Gamma(\beta)|^2$, respectively. Eq. 12.201 can be computed by taking the inverse FFT of $R_x(\beta)$, windowing it, and then taking the forward FFT.

Unfortunately, Eq. 12.201 is usually not as good an approximation as Eq. 12.199, because Eq. 12.201 smooths the low-frequency components of the power spectrum ($R_x(\beta)$ at $\beta < 1$) in a way that is perceptually unrealistic, and often unacceptable. Schroeder proposed, instead, that audio coders should implement Eq. 12.199 directly, but that the masking smoother doesn't need to be quite as complicated as that given in Eq. 12.198. Instead, he proposed using the triangular smoother:

$$|\Gamma(\beta)|^2 = \max(0, 1 - |\beta|) \tag{12.202}$$

Example 12.4.3 Masking Spectrum of a Sinusoid

$$R_x(\omega) = \frac{\pi}{2}\delta(\omega - 0.64\pi) \tag{12.203}$$

Suppose that the sampling frequency is $F_s = 8000$Hz. Then a tone at $\omega = 0.64\pi$ is a tone at 2560Hz, which corresponds to $11\log(9) = 24.2$ Bark.

$$R_x(\beta) = \frac{\pi}{2}\text{ERB}(0.64\pi)\delta(\beta - 11\log(9)) \tag{12.204}$$

Where the ERB at 2560Hz comes out to be 265Hz, or $0.066pi$.

Using Schroeder's convolution-in-frequency method, the masking spectrum is computed as

$$\begin{aligned} M_x(\beta) &= \frac{1}{4\text{ERB}(0.64\pi)}\int_0^{\beta(\pi)} R_x(\psi)|\Gamma(\beta-\psi)|^2 d\psi && (12.205)\\ &= \frac{\pi}{8}\left|\Gamma(\beta - 11\log(9))\right|^2 && (12.206)\\ &= \max\left(0, \frac{\pi}{8}\left(1 - |\beta - 11\log(9)|\right)\right) && (12.207)\end{aligned}$$

where the last line uses Schroeder's triangular spread-of-masking function.

$M_x(\omega)$ can be computed by inverse-frequency-warping $M_x(\beta)$. If $\Gamma(\beta)$ is a triangle centered at $\beta(0.11\pi)$, then $\Gamma(\omega)$ will be a pseudo-triangle centered at $\omega = 0.64\pi$:

$$\begin{aligned} M_x(\omega) &= \max\left(0, \frac{\pi}{8\text{ERB}(\omega)}\left(1 - |\beta(\omega) - 11\log(9)|\right)\right) && (12.208)\\ &= \max\left(0, \frac{1}{\alpha}\left(1 - |\beta(\omega) - 11\log(9)|\right)\right) && (12.209)\\ &\approx \max\left(0, \frac{1}{\alpha}\left(1 - \frac{|\omega - 0.64\pi|}{0.066\pi}\right)\right) && (12.210)\end{aligned}$$

In the solution above, the first approximation makes the relatively benign assumption that $\text{ERB}(\omega) \approx \text{ERB}(0.11\pi)$ for all ω within one ERB of 0.11π. The second approximation assumes that the pseudo-triangle is approximately a triangle in ω, with a base width equal to twice the ERB. The second approximation is obviously a little more drastic than the first, but it is not unreasonable, considering that the triangular spread-of-masking function is itself only an approximate representation of the true gammatone filter shape.

The solution demonstrates that a pure tone at a relatively high frequency (2560Hz) masks lower-amplitude noise within one ERB on either side of the tone, i.e., within about ±10% of the tone's center frequency. Outside of this one-ERB range, a pure tone masks very little of the quantization noise.

12.4.7 Sinusoidal Transform Coding

Audio synthesis in a sinusoidal transform coder consists of three steps. First, L or more continuous frequency "tracks" are constructed by stringing together harmonics from consecutive frames. Second, amplitudes and phases of the peaks in each track are smoothly interpolated, in order to avoid abrupt discontinuities at phrase boundaries. Finally, the smoothly interpolated amplitudes and phases are inserted into a Fourier series equation in order to synthesize the output audio waveform.

STFT of Periodic Signals

Suppose $x[n]$ is periodic with fundamental frequency $\omega_0 = 2\pi/\tau_0$, i.e.

$$x[n] = \sum_{l=1}^{L} \gamma_l e^{jl\omega_0 n} \tag{12.211}$$

Then

$$\begin{aligned} X_f(\omega) &= \sum_{n=-\infty}^{\infty} x[n+fS]w[n]e^{-j\omega n} \\ &= \sum_{n=-\infty}^{\infty} w[n]e^{-j\omega n} \sum_{l=1}^{L} \gamma_l e^{jl\omega_0(n+fS)} \\ &= \sum_{l=1}^{L} \gamma_l e^{jl\omega_0 fS} \sum_{n=-\infty}^{\infty} w[n]e^{-jn(\omega - l\omega_0)} \\ &= \sum_{l=1}^{L} \gamma_{lf} W(\omega - l\omega_0) \end{aligned}$$

where the frame-synchronous Fourier series coefficient is defined as

$$\gamma_{lf} = \gamma_l e^{jl\omega_0 fS} \tag{12.212}$$

"Pitch-synchronous analysis" is analysis using a window length N that is an integer multiple of the pitch period τ_0. Pitch-synchronous analysis makes it possible to design much simpler, low-noise analysis algorithms. The problem with pitch-synchronous analysis is that it is only possible if the pitch period τ_0 is known in advance. Many music analysis algorithms first estimate τ_0 using an arbitrary N, then go back and perform the STFT a second time using $N = 2\tau_0$.

The rectangular window spectrum $W_R(\omega)$ has nulls at $\omega = 2\pi k/N$ for any nonzero integer k. The Hanning and Hamming windows have nulls at $\omega = 2\pi k/N$ for $k \geq 2$. The triangular window $W_T(\omega)$ has nulls at $\omega = 2\pi k/N$ for even values of k.

Suppose that N is an even multiple of τ_0, meaning that N/τ_0 is an even integer. Consider the N-point DFT of the window:

$$W\left(\frac{2\pi k}{N}\right) = \begin{cases} W(0) & k = 0 \\ 0 & k = \text{multiple of } N/\tau_0 \\ \text{possibly nonzero} & \text{otherwise} \end{cases} \tag{12.213}$$

For example, in the typical "pitch-synchronous" case that $N/\tau_0 = 2$, the DFT of the window is equal to zero for all frequency samples such that k is even. Remember that the $k = 0$ term is just the DC value of the window:

$$W(0) = \sum_n w[n] \tag{12.214}$$

Putting this information into the formula for the STFT, we get

$$X_f\left(\frac{2\pi k}{N}\right) = \sum_{l=1}^{L} \gamma_{lf} W\left(\frac{2\pi k}{N} - \frac{2\pi l}{\tau_0}\right) = \sum_{l=1}^{L} \gamma_{lf} W\left(\frac{2\pi}{N}(k - lN/\tau_0)\right) \tag{12.215}$$

At the frequencies such that $k = lN/\tau_0$ (ω is an integer multiple of ω_0), the STFT simplifies to

$$X_f\left(\frac{2\pi k}{N}\right) = \gamma_{lf} W(0) \quad \text{if } k = lN/\tau_0 \tag{12.216}$$

Equation 12.216 suggests an algorithm for audio synthesis. First, estimate the pitch period τ_0. Second, compute the STFT using $N/\tau_0 = 2$. Third, compute coefficients of the Fourier series using the formula

$$\gamma_{lf} = \frac{X_f(2\pi l/\tau_0)}{W(0)} \tag{12.217}$$

Finally, resynthesize the time-domain waveform using the Fourier series formula:

$$x[n] = \sum_{l=1}^{L} \gamma_{lf} e^{jl\omega_0 n}, \quad fS \leq n \leq (f+1)S - 1 \tag{12.218}$$

The Quatieri and McAulay method implements the analysis algorithm in equation 12.217 once per frame, with the following simplification: Quatieri and McAulay normalize the window in advance, so that $W(0) = 1$.

Analysis

The analysis part of the sinusoidal transform coder computes three sets of parameters in every frame: a set of sine wave frequencies ω_{lf}, the corresponding sine wave amplitudes A_{lf}, and the corresponding phases θ_{lf} (l is the harmonic number, f is the frame number). The frame rate is usually about 30-100 frames/second. The number of peaks per frame ranges between about $L = 10$ (for low-bandwidth, low-quality coding) up to as much as $L = N/2$ (in which case the number of parameters transmitted is equal to the dimension of the STFT; this case is used for high-quality time-scale modification and other problems that do not require bandwidth compression).

$$\omega_{lf} = \arg\max X_f(\omega), \quad l = 1, ..., L, \quad 0 \le \omega \le \pi \tag{12.219}$$

$$\gamma_{lf} = \frac{X_f(\omega_{lf})}{W(0)} \tag{12.220}$$

$$A_{lf} = |\gamma_{lf}| \tag{12.221}$$

$$\theta_{lf} = \angle\gamma_{lf} \tag{12.222}$$

Notice that there is no need to compute or transmit the amplitudes and phases of peaks in the negative-frequency part of the spectrum; by conjugate symmetry,

$$|X_f(-\omega_{lf})| = A_{lf} W(0) \tag{12.223}$$

$$\angle X_f(-\omega_{lf}) = -\theta_{lf} \tag{12.224}$$

Track Synthesis

Audio synthesis in a sinusoidal transform coder consists of three steps. First, L or more continuous frequency "tracks" are constructed by stringing together harmonics from consecutive frames. Second, amplitudes and phases of the peaks in each track are smoothly interpolated, in order to avoid abrupt discontinuities at phrase boundaries. Finally, the smoothly interpolated amplitudes and phases are inserted into a Fourier series equation in order to synthesize the output audio waveform.

The process of tracking assigns each harmonic to a "track number" $i(l, f)$ dependent on both the peak number l and the frame number f. Tracks are "grown" from left to right, by connecting peaks in consecutive frames if their peak frequencies are not too far apart. The tracking algorithm begins with $i(l, 1) = 1$, meaning that in the first frame, track number equals peak number. The total number of tracks is initialized to $I = L$.

After initializing the tracks, the following steps are performed for every frame $f \ge 1$. First, compute the forward match $l(k)$ and the backward match $k(l)$:

$$l(k) = \arg\min |\omega_{l(k)f} - \omega_{k,f-1}|, \quad 1 \le k \le L \tag{12.225}$$

$$k(l) = \arg\min |\omega_{lf} - \omega_{k(l),f-1}|, \quad 1 \le l \le L \tag{12.226}$$

If any pair is joined by both forward-match and backward-match ($k(l(k)) = k$), and if the peak frequencies are not too far apart ($|\omega_{lf} - \omega_{k,f-1}| \leq \Delta$ for some threshold Δ), then they should be joined together into the same track:

$$i(l(k), f) = i(k, f-1) \tag{12.227}$$

If the forward-match and backward-match disagree, but the peaks are not too far apart ($|\omega_{lf} - \omega_{k,f-1}| \leq \Delta$), then try the second-best forward match. If that fails, try the second-best backward match.

If any peak $\omega_{k,f-1}$ can not be coupled with any forward match such that $|\omega_{lf} - \omega_{k,f-1}| \leq \Delta$, then the track $i(k, f-1)$ "dies out" and is never re-used.

If any peak $\omega_{l,f}$ can not be coupled with any backward match such that $|\omega_{lf} - \omega_{k,f-1}| \leq \Delta$, then the peak is assigned to a new track "born" in frame number f. Mechanically, "birth" of a new track is represented by the steps

$$I = I + 1, \quad i(l, f) = I \tag{12.228}$$

Once all peaks have been assigned to tracks, the inverse map must be constructed. In every frame, construct the inverse lookup table

$$l(i, f) = \begin{cases} l & \text{if } \exists\, l : i(l, f) = i \\ 0 & \text{otherwise} \end{cases} \tag{12.229}$$

Interpolation

Once the peaks have been assigned to tracks, peak amplitudes and phases are smoothly interpolated between frame boundaries. In this section, the peak number l is assumed to always mean "the peak corresponding to track number i." For example, ω_{lf} is a shorthand for $\omega_{l(i,f),f}$, while $\omega_{l,f+1}$ is a shorthand for $\omega_{l(i,f+1),f+1}$.

Amplitude of track i in frame f is linearly interpolated between the peaks $l(i, f)$ and $l(i, f+1)$:

$$A_i[n] = \left(\frac{n - fS}{S}\right) A_{lf} + \left(\frac{(f+1)S - n}{S}\right) A_{l,f+1}, \quad fS \leq n \leq (f+1)S - 1 \tag{12.230}$$

Tracks that "die" in frame f are linearly interpolated down to zero amplitude. Tracks that are "born" in frame $f+1$ are linearly interpolated up from zero amplitude.

Frequency is the derivative of phase: θ_{lf} is the phase in radians of peak number l at time $n = fS$, and ω_{lf} is the phase derivative of the same peak measured in radians per sample. Frequency and phase can be interpolated together using a smooth cubic function

$$\theta_i[n] = \alpha_{if} + \beta_{if}(n - fS) + \gamma_{if}(n - fS)^2 + \delta_{if}(n - fS)^3, \quad fS \leq n \leq (f+1)S - 1 \tag{12.231}$$

where the constants α_{if}, β_{if}, γ_{if}, and δ_{if} are chosen so that both phase and frequency match the boundary conditions at both ends of the frame:

$$\begin{aligned}
\theta_i[fS] = &\quad \theta_{lf} &= \alpha_{if} \qquad (12.232)\\
\dot{\theta}_i[fS] = &\quad \omega_{lf} &= \beta_{if} \qquad (12.233)\\
\theta_i[(f+1)S] = &\quad \theta_{l,f+1} + 2\pi M &= \alpha_{if} + \beta_{if}S + \gamma_{if}S^2 + \delta_{if}S^3 \qquad (12.234)\\
\dot{\theta}_i[(f+1)S] = &\quad \omega_{l,f+1} &= \beta_{if} + 2\gamma_{if}S + 3\delta_{if}S^2 \qquad (12.235)
\end{aligned}$$

Equations 12.232 and 12.235 would completely specify the four parameters α_{if}, β_{if}, γ_{if}, and δ_{if}, except that the phase at the end of the frame is ambiguous: it is possible to add any integer multiple of 2π to the phase without changing $e^{j\theta_l}$. The phase ambiguity is resolved by choosing the parameter M so that the ending phase, $\theta_{l,f+1} + 2\pi M$, is as close as possible to the starting phase θ_{lf} plus the average phase derivative:

$$M = \arg\min \left| \theta_{lf} + \frac{(\omega_{l(i,f),f} + \omega_{l(i,f),f})}{2} S - (\theta_{l,f+1} + 2\pi M) \right| \qquad (12.236)$$

If a frequency track is born or dies during frame f, the frequency of the track stays the same from beginning to end of frame f: consequently, the phase changes linearly all through the frame.

Synthesis

Once peaks have been arranged into tracks, and the phase and amplitude of each track are known for every frame, then synthesis uses the time-varying Fourier series formula:

$$\hat{x}[n] = \sum_{i=1}^{I} A_i[n] \cos(\theta_i[n]), \quad fS \le n \le (f+1)S - 1 \qquad (12.237)$$

where $\hat{x}[n]$ is the synthesized approximation to $x[n]$, S is the frame spacing, and i is the "track" number.

Time-Scale and Pitch Modification

When a sampled waveform is played back at B times its original sampling frequency, the waveform changes in two ways. First, the waveform speeds up by a factor of B. Second, the pitch of the waveform also increases by a factor of B. The pitch change associated with speedup or slowdown of a waveform is sometimes called "wow," or more colloquially, "the chipmunk effect."

Using sinusoidal transform coding, it is possible to change the time scale of a waveform without changing its pitch. The algorithm has three steps:

1. Perform analysis and peak tracking as described above.
2. Change the frame skip parameter from S samples to S/B samples.

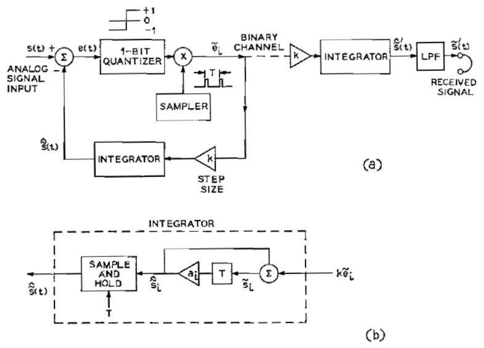


Figure 12.25: Delta modulator with single integration

3. Perform amplitude interpolation, phase interpolation, and Fourier series synthesis exactly as described above, but using S/B instead of S as the frame skip parameter.

Likewise, it is possible to use sinusoidal transform coding to perform pitch modification. The simplest pitch modification algorithm is as follows:

1. Modify the time-scale of the waveform by a factor of $1/B$.

2. Play back the waveform at a sampling rate B times its original sampling rate.

12.5 Predictive Quantization

12.5.1 Delta Modulation

Considerable interest attaches to realizing the advantages of digital transmission in economical ways. Multi-bit quantizers, such as used in PCM, are relatively expensive. In telephone communication they normally are not dedicated to individual customers, but typically are shared in time-division multiplex. This requires individual analog transmission to a central point where the digitizing occurs.

In many instances it is desirable to digitize the signal immediately at the source (for example, in some rural telephone systems). Inexpensive digital encoders which can be dedicated to individual customers are therefore required. Delta modulation is one solution.

Delta modulation (DM) may be considered perhaps the simplest form of DPCM. Quantization of the error signal is to one-bit only (i.e., a simple comparator), and a single or double integrator is typically used as the predictor network, as shown in Fig. 12.25a. The transmitted binary samples, e_i, are either +1 or -1 and represent the sign of the error, $e(t)$. The integrator can be implemented many ways, including a simple analog storage capacitor. A digital implementation, using the terminology employed in the earlier discussion of predictive quantizing, is shown in Fig. 12.25b.

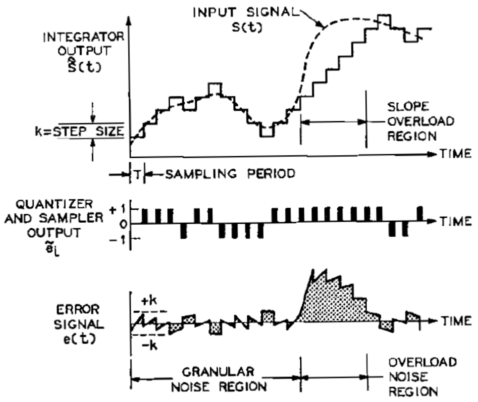


Figure 12.26: Waveforms for a delta modulator with single integration

The box T is a one-sample delay and $a_l = 1$ for an ideal integrator. A sample-and-hold converts the discrete samples to a continuous function.

The local estimate provided by the integrator, $\hat{\hat{s}}(t)$, is the staircase function shown in Fig. 12.26. The step size of the staircase function is determined by the amplifier constant, k. The step-size is typically chosen small compared to the input signal magnitude. Two types of distortion can occur in the estimate -granular distortion and slope overload. The former is determined by the step size of the quantization (that is, by the amplifier k). The latter is caused by the inability of the encoder to follow the signal when its slope magnitude exceeds the ratio of step size to sampling period,

$$|\dot{s}| > k/T. \tag{12.238}$$

These two types of distortion are indicated in Fig. 12.26.

Granular distortion can be made small by using a small step size. Slope overload can be reduced by using a large step size or by running the sampler (clock) faster. The latter of course increases the transmitted bit rate. In typical designs, for a prescribed bit rate, the step size is selected to effect a compromise "mix" between quantizing distortion and slope overload. Perceptually, more overload noise power is tolerable than granular noise power (JAYANT and ROSENBERG). During granular distortion the samples of the error signal tend to be uncorrelated and the error signal power spectrum tends to be uniform.

For high-quality speech transmission, say with signal-to-noise ratio of the order of 40 dB, the resulting bit rate for simple DM is relatively high, typically greater than 200 Kbps. Tolerable channel error rates are typically 10^{-4}. The signal-to-quantizing noise present in the received signal is strongly dependent upon the final low-pass filter. If simple low-pass filters are used for desampling, the transmission bit rate must be pushed into the Mbps range to achieve high quality. Such high bit rates cannot be supported in many transmission facilities. Consequently, there is strong interest in techniques for reducing the bit rate of DM while at the same time retaining most of

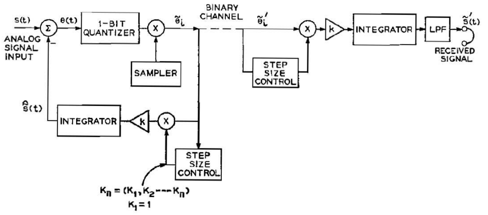


Figure 12.27: Adaptive delta modulator with single integration

its advantages in circuit simplicity. Adaptive delta modulation (ADM) is one such solution.

In ADM the quantizer step size is varied according to a prescribed logic. The logic is chosen to minimize quantizing and slope distortion when the sampler is run at a relatively slow rate[6]. The additional control is typically effected by a step size multiplier incorporated in the feedback loop, as shown in Fig. 12.27. As in simple DM, the feedback network may be a single or double integration. The step control logic may be discrete or continuous (Jayant [1970], Greefkes [1957], Greefkes and de Jager [1968], de Jager [1952], Abate [1967]), and it may act with a short time constant (i.e., sample-by-sample) or with a time constant of syllabic duration (Greefkes [1957], Tomozawa and Kaneko [1968]). Normally the step size is controlled by information contained in the transmitted bit stream, but it may be controlled by some feature of the input signal; for example, the slope magnitude averaged over several msec (de Jager [1952]). In this case, the control feature must be transmitted explicitly along with the binary error signal.

The receiver duplicates the feedback (predictor) branch of the transmitter, including an identical step size control element. In the absence of errors in the transmission channel, the receiver duplicates the transmitter's estimate of the input signal. Desampling by a low-pass filter to the original signal bandwidth completes the detection.

The manner in which discrete adaptation is implemented is illustrated in Fig. 12.28. As long as the slope of the input signal is small enough that the signal can be followed with the minimum step size, k, the multiplier is set to $K_n = K_1 = 1$. When the input signal slope becomes too great, the step size multiplier is increased to permit more accurate following and to minimize slope overload. In the logic illustrated, an increase in step size is made whenever three successive samples of $\tilde{e}_i$ have the same polarity. At the point of greatest input signal slope, a step multiplication by K_3 is attained. Further increases can be accomplished in successive samples, if needed, until a maximum multiplication of K_N is achieved. Any situation where the current channel bit and the past two bits are not the same results in a reduction in step size. Reductions can likewise be accomplished successively until the minimum value $K_n = K_1 = 1$ is again attained.

---

[6] Adaptation is normally not applied to the feedback network, but this is an attractive possibility for further improvement in the encoding.

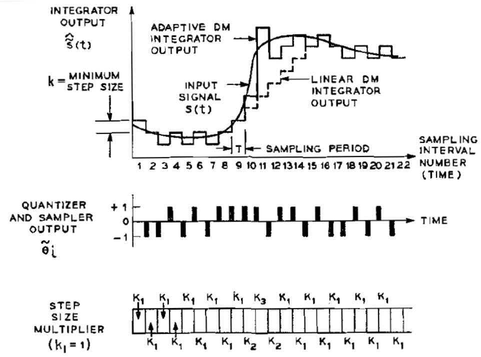


Figure 12.28: Waveform for an adaptive delta modulator with discrete control of the step size

Exponential adaptation logics have been found valuable for speech encoding (Jayant [1970]). In this case, the multiplier is typically $K_n = P^{n-l}$, $n = 1, \ldots N$. A typical value of P is in the order of 1.5 to 2.0. As few as eight ($N = 8$) discrete multiplier values are found adequate in some applications of exponential ADM.

Because of the ability to "shift gears" automatically, ADM can be designed to yield signal quality comparable to 7-bit log PCM at bit rates commensurate with PCM; typically, 56 Kbps for a 4 KC signal band. At lower bit rates, ADM can surpass PCM in signal-to-noise (S/N) performance. This relation results because S/N for ADM varies roughly as the cube of the sampling rate. For PCM the growth in S/N is 6 dB/bit of quantizing. At low bit rates, ADM wins out. However, the range of normally useful quality is restricted to rates greater than 20 Kbps. A S/N comparison is shown for ADM and PCM in Fig. 12.29.

Because delta modulators can be implemented very economically in digital circuitry, they constitute an attractive means for initial analog-todigital conversion of signals. However, other formats of digital encoding are frequently used in digital communication systems. Techniques for direct digital conversion from one signal format to another, with no intervening analog detection, are therefore of great interest. Present work in digital communication includes direct digital transformation between simple DM, ADM, linear PCM, log PCM, and DPCM (Shipley [1971], Goodman [1969]). These and related studies aim to establish coding relations which make the transmission system and switching network "transparent" to the signal, regardless of its digital form.

12.5.2 Differential PCM (DPCM)

Successive speech samples are highly correlated. The long-term average spectrum of voiced speech is reasonably well approximated by the function $S(f) = 1/f$ above about 500 Hz; the first-order inter-sample correlation coefficient is approximately 0.9. In differential PCM, each sample $s(n)$ is compared to a prediction $s_p(n)$, and

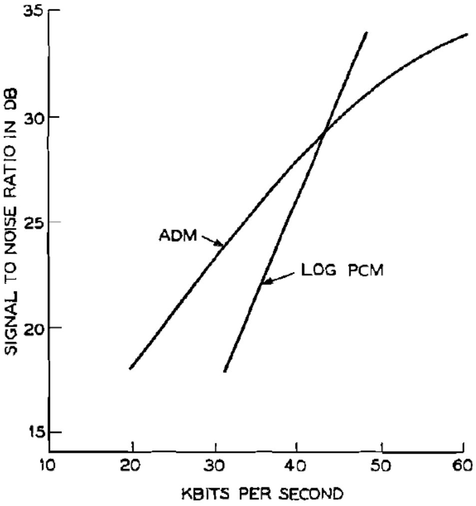


Figure 12.29: Signal-to-noise ratios as a function of bit rate. Performance is shown for exponentially adaptive delta modulation (ADM) and logarithmic PCM. (After (Jayant [1970]))

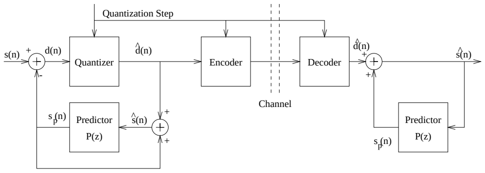


Figure 12.30: Schematic of a DPCM coder

the difference is called the prediction residual $d(n)$ (Figure 12.30). $d(n)$ has a smaller dynamic range than $s(n)$, so for a given error power, fewer bits are required to quantize $d(n)$.

Accurate quantization of $d(n)$ is useless unless it leads to accurate quantization of $s(n)$. In order to avoid amplifying the error, DPCM coders use a technique copied by many later speech coders: the encoder includes an embedded decoder, so that the reconstructed signal $\hat{s}(n)$ is known at the encoder. By using $\hat{s}(n)$ to create $s_p(n)$, DPCM coders avoid amplifying the quantization error:

$$\begin{aligned} d(n) &= s(n) - s_p(n) && (12.239)\\ \hat{s}(n) &= \hat{d}(n) + s_p(n) && (12.240)\\ e(n) = s(n) - \hat{s}(n) &= d(n) - \hat{d}(n) && (12.241) \end{aligned}$$

Two existing standards are based on DPCM. In the first type of coder, continuously varying slope delta modulation (CVSD), the input speech signal is upsampled to either 16kHz or 32kHz. Values of the upsampled signal are predicted using a one-tap predictor, and the difference signal is quantized at one bit per sample, with an adaptively varying Δ. CVSD performs badly in quiet environments, but in extremely noisy environments (e.g. helicopter cockpit), CVSD performs better than any LPC-based algorithm, and for this reason it remains the US Department of Defense recommendation for extremely noisy environments (Kohler [1997], Tardelli and Kreamer [1996]).

DPCM systems with adaptive prediction and quantization are referred to as adaptive differential PCM systems (ADPCM). A commonly used ADPCM standard is G.726 which can operate at 16, 24, 32, or 40 kbps (2-5 bits/sample) (ITU-T [1990a]). G.726 ADPCM is frequently used at 32 kbps in land-line telephony. The predictor in G.726 consists of an adaptive second-order IIR predictor in series with an adaptive sixth-order FIR predictor. Filter coefficients are adapted using a computationally simplified gradient descent algorithm. The prediction residual is quantized using a semi-logarithmic companded PCM quantizer at a rate of 2-5 bits per sample. The quantization step size adapts to the amplitude of previous samples of the quantized prediction error signal; the speed of adaptation is controlled by an estimate of the type of signal, with adaptation to speech signals being faster than adaptation to signaling tones.

12.5.3 Differential Pulse Code Modulation

Predictive quantizing, or feedback around the quantizer, is a method used in a wide class of digital encoders for reducing the redundancy of a signal. The idea is to form an estimate of the sample of the input signal, and quantize the difference between the signal and its estimate. For accurate estimates, the variance of the difference, or error signal, is less than that of the input and fewer bits are required to transmit

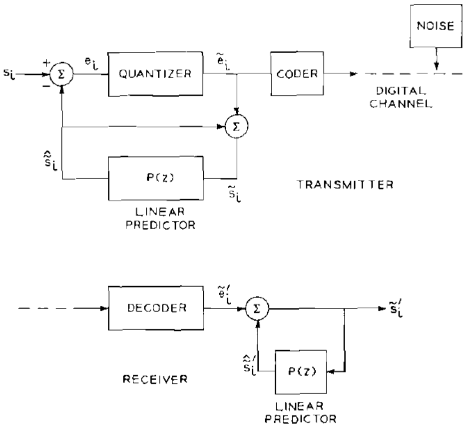


Figure 12.31: Predictive quantizing system. (After (McDonald [1966]))

the error. Estimators typically include linear prediction networks (both adaptive and nonadaptive) and single or multiple integrators. Differential pulse code modulation (DPCM) and delta modulation (DM) are special cases of predictive quantizing, the latter using merely a 1-bit quantizer for the error signal.

Estimation or prediction of the signal requires knowledge of input signal statistics. In a nonadaptive predictor these data are built into a fixed feedback network. In adaptive prediction, the network is changed as the input signal changes its characteristics. Digital transmission can be made relatively free of channel errors in well-designed systems. The controlling impairment is consequently noise introduced by the quantization process.

Fig. 12.31 shows a predictive quantizing system (McDonald [1966]). Input signal samples are s_i; the local (transmitter) estimate of the signal is $\hat{\tilde{s}}_i$; the error signal is e_i, which when quantized is $\tilde{e}_i$. The locally reconstructed signal is $\tilde{s}_i = (e_j + \hat{\tilde{s}}_j)$. For transmission, $\tilde{e}_i$ coded into a prescribed digital format and transmitted. Any digital errors in transmission cause a corrupted version of the error signal, $\tilde{e}'_i$, to be received. Detection produces the reconstructed signal $\tilde{s}'_i$.

This type of differential quantizing has the important feature that the quantization noise in the reconstructed signal is the same as that in the error signal–that is, quantization noise does not accumulate in the reconstructed signal. Quantization noise samples are

$$q_i = (e_i - \tilde{e}_i) \tag{12.242}$$

$$\begin{aligned} &= (s_i - \hat{\tilde{s}}_i - \tilde{e}_i) \\ &= (s_i - \tilde{s}_i) \end{aligned}$$

The quantization noise in the transmitted error signal is therefore identical to the quantization noise in the reconstructed signal.

A logical measure of the effectiveness of the predictor in reducing signal redundancy is the amount by which the power of the error signal is reduced below that of the input signal. This ratio is

$$\xi^2 \equiv \frac{E[s_i^2]}{E[e_i^2]} \tag{12.243}$$

where $E[x]$ denotes the expected value of x. To assess this figure one needs to know explicitly the predictor characteristics. Linear prediction represents a well known class of feedback networks. For linear prediction the signal estimate is formed from a linear combination of past values of the reconstructed input signal. That is,

$$\hat{\hat{s}}_i = \sum_{j=1}^{N} a_j \tilde{s}_{i-j}{}^{7} \tag{12.244}$$

$$\begin{aligned} &= \sum_{j=1}^{N} a_j \left[s_{i-j} - (e_{i-j} - \tilde{e}_{i-j})\right] \\ &= \sum_{j=1}^{N} a_j s_{i-j} - \sum_{j=1}^{N} a_j q_{i-j} \end{aligned}$$

for an N-th order predictor. The variance of the error signal is

$$E\left[e_i^2\right] = E\left[\left(s_i - \hat{\hat{s}}_i\right)^2\right]. \tag{12.245}$$

If the correlation between error samples is vanishingly small (i.e., if the power spectrum of the error is uniform) and if the correlation between input and error samples is negligible, then

$$E\left[e_i^2\right] \approx E\left[\left(s_i - \sum_{j=1}^{N} a_j s_{i-j}\right)^2\right] + e[q_i^2] \sum_{j=1}^{N} a_j^2. \tag{12.246}$$

For a given signal, therefore, maximizing ξ^2 is equivalent to minimizing $E[e_i^2]$. Differentiation of $E[e_i^2]$ with respect to a_j and setting the resulting equations to zero gives

$$\begin{aligned} &\rho_1 = (1 + 1/R)a_1 + a_2\rho_1 + a_3\rho_2 + \cdots + a_N\rho_{N-1} \\ &\rho_2 = a_1\rho_1 + (1 + 1/R)a_2 + a_3\rho_1 + \cdots + a_N\rho_{N-2} \\ &\vdots \\ &\rho_N = a_1\rho_{N-1} + a_2\rho_{N-2} + a_3\rho_{N-3} + \cdots + (1 + 1/R)a_N \end{aligned} \tag{12.247}$$

---

[7] The absence of an a_0 term implies delay around the loop.

where $R = E[s_i^2]/E[q_i^2]$ is the signal-to-quantizing noise ratio, and $\rho_j = E[s_i s_{i-j}]/E[s_i^2]$ is the signal autocovariance. The minimum of $E[e_i^2]$ can be written (McDonald [1966]).

$$E[e_i^2]_{min} = E[s_i^2]\left[1 - \sum_{j=1}^{N} a_j\left(\frac{\rho_j}{(1+1/R)}\right)\right],$$

so that

$$\xi^2\Big|_{max} = \left[1 - \sum_{j=1}^{N} a_j\rho_j/(1+1/R)\right]^{-1}. \tag{12.248}$$

The quantization noise power $E[q_i^2]$ depends upon properties of the quantizer. For example, for a linear quantizer of L steps, of step size Δ_l and step probability P_l. the quantizing noise power can be shown to be (Carlson [1968])

$$E[q_i^2] = \sum_{l=1}^{L} P_l\frac{\Delta_l^2}{12}. \tag{12.249}$$

For relatively fine quantizing, the quantizer noise is negligible compared to other terms in $E[e_i^2]$. Historically, a commonly-used feedback network in DPCM systems is a simple integrator or accumulator. For this case $N = 1$ and

$$\begin{aligned} a_1 &= 1, \\ a_j &= 0, \quad j \neq 1 \end{aligned}$$

and

$$\hat{\hat{s}}_i = \sum_{j=1}^{\infty} \tilde{e}_{i-j}$$

$$e_i = s_i - \left(\hat{\hat{s}}_{i-1} + \tilde{e}_{i-1}\right). \tag{12.250}$$

The error power from (12.246) is

$$E[e_i^2] = E[s_i^2][2(l-\rho_1)] + E[q_{i-1}^2]. \tag{12.251}$$

Neglecting the quantizing noise,

$$\xi^2 \approx \frac{1}{2(1-\rho_1)} \tag{12.252}$$

The optimum $N = 1$ predictor (in the least error power sense) is however

$$a_1 = \frac{\rho_1}{(1+1/R)},$$

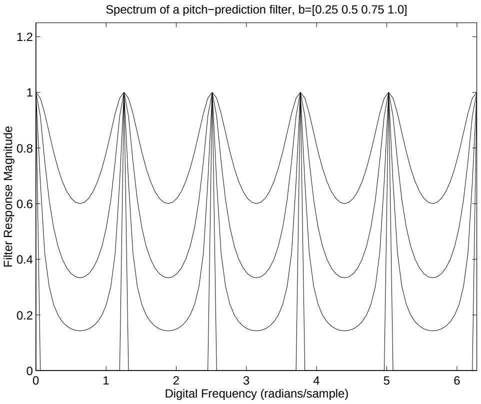


Figure 12.32: Normalized magnitude spectrum of the pitch prediction filter for several values of the prediction coefficient.

for which

$$\xi^2 \approx \frac{1}{(1-\rho_1^2)}. \tag{12.253}$$

The optimum predictor therefore shows a slight advantage (for the case $N = 1$) over the simple ideal integrator (McDonald [1966]).

Computer studies on speech show that DPCM with a fixed linear predictor network optimized according to the preceding discussion gives approximately $10\log_{10}\xi^2 =$ 10dB. Over 9 dB of this improvement is achieved by an $N = 2$ optimum predictor. Compared to a straight PCM encoding, this means that 1 to 2 bits per sample may be saved in the encoding.

Predictive coding and quantizing has been applied in several forms to the digital transmission of speech. Optimum nonadaptive linear predictors for speech have been studied to reduce the bit rate for transmission below that of conventional PCM (McDonald [1966], Haskew [1969], Fujisaki [1960]). Adaptive predictive coding has also been used in which the predictor is designed to represent the pitch of voiced sounds and the shape of the signal spectrum (Schroeder [1968], Kelly and Miller [1967]). Predictive quantizing can be implemented with adaptive quantization as well as with adaptive prediction.

12.5.4 Pitch Prediction Filtering

In an LPC-AS coder, the LPC excitation is allowed to vary smoothly between fully voiced conditions (as in a vowel) and fully unvoiced conditions (as in /s/). Intermediate levels of voicing are often useful to model partially voiced phonemes such as /z/.

The partially voiced excitation in an LPC-AS coder is constructed by passing an uncorrelated noise signal $c(n)$ through a pitch prediction filter (Atal [1982], Ra-

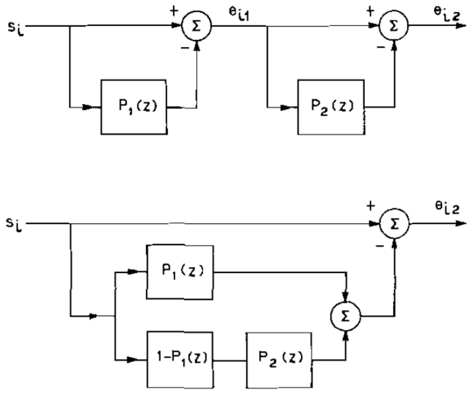


Figure 12.33: Two stage predictor for adaptive predictive coding. (After (Schroeder [1968]))

machandran and Kabal [1987]). A typical pitch prediction filter is

$$u(n) = gc(n) + bu(n - T_0) \tag{12.254}$$

where T_0 is the pitch period. If $c(n)$ is unit-variance white noise, then according to Equation 12.254 the spectrum of $u(n)$ is

$$|U(e^{j\omega})|^2 = \frac{g^2}{1 + b^2 - 2b\cos\omega T_0} \tag{12.255}$$

Figure 12.32 shows the normalized magnitude spectrum $(1 - b)|U(e^{j\omega})|$ for several values of b between 0.25 and 1. As shown, the spectrum varies smoothly from a uniform spectrum, which is heard as unvoiced, to a harmonic spectrum which is heard as voiced, without the need for a binary voiced/unvoiced decision.

In LPC-AS coders, the noise signal $c(n)$ is chosen from a “stochastic codebook” of candidate noise signals. The stochastic codebook index, the pitch period, and the gains b and g are chosen in a closed-loop fashion in order to minimize a perceptually weighted error metric. The search for an optimum T_0 typically uses the same algorithm as the search for an optimum $c(n)$. For this reason, the list of excitation samples delayed by different candidate values of T_0 is typically called an “adaptive codebook” (Singhal and Atal [1984]).

12.5.5 Adaptive Predictive Coding

Adaptive predictive coding has been used to reduce signal redundancy in two stages: first by a predictor that removes the quasi-periodic nature of the signal, and second by a predictor that removes formant information from the spectral envelope ((Schroeder [1968])). The first predictor is simply a gain and delay adjustment, and the second

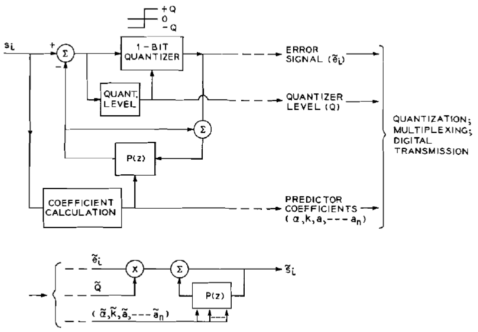


Figure 12.34: Adaptive predictive coding system. (After (Schroeder [1968]))

is a linear combination of past values of the first predictor output. The equivalent operations are shown in Fig. 12.33, where

$$\begin{aligned} P_1(z) &= \alpha z^{-k} \\ P_2(z) &= \sum_{j=1}^{N} a_j z^{-j} \end{aligned}$$

$$P(z) = \{P_1(z) + P_2(z)\,[1 - P_1(z)]\} \tag{12.256}$$

This predictor is used in the DPCM encoder form with a two-level (1 bit) quantizer for the error signal, as shown in Fig. 12.34. The quantizer level is variable and is adjusted for minimum quantization noise power in the error signal. The quantizer representation level Q is set to the average absolute value of the error samples being quantized, i.e.,

$$Q = \frac{1}{N} \sum_{j=1}^{N} |e_j|. \tag{12.257}$$

The coefficients for predictor $P_2(z)$ are calculated as described previously. Those for $P_1(z)$, i.e., α and k, are obtained by minimizing the error power from the first predictor

$$\epsilon_1^2 = \sum_{j=1}^{N} (s_j - \alpha s_{j_k})^2. \tag{12.258}$$

The minimum is given by

$$\alpha = \left. \frac{\sum_{j=1}^{N} s_j s_{j_k}}{\sum_{s_{j-k}^2}} \right|_{k=\text{optimum}}, \tag{12.259}$$

where the optimum k maximizes the normalized correlation

$$\rho = \frac{\sum_j s_j s_{j-k}}{\left\{ \sum_j s_j^2 \sum_j s_{j-k}^2 \right\}^{\frac{1}{2}}} \tag{12.260}$$

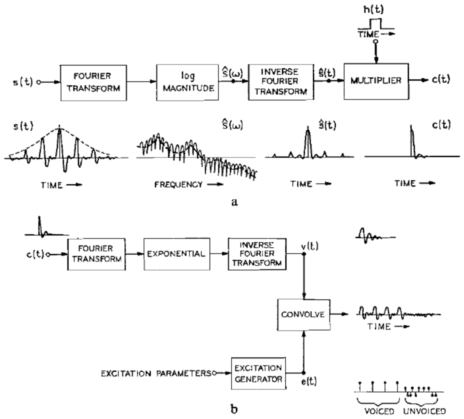


Figure 12.35: Analysis and synthesis operations for the homomorphic vocoder. (After (Oppenheim [1969]))

The optimum k is found by a search of computed and tabulated values of p.

One implementation of the predictive system has been made for digital transmission at 9600 bps and at 7200 bps (Kelly and Miller [1967]). The system was optimized in extensive computer-simulation studies. It used the following parameters and quantization to achieve digital transmission at 9600 bps: signal bandwidth = 2950 Hz; sampling rate = 6 kHz; prediction optimization interval = 10 msec ($N = 60$ samples); $P_1(z)$ predictor quantization: $\alpha = 3$ bits, $k = 7$ bits (determined by maximum delay of 20 msec, or 120 samples at 6 kHz, for the computation of p; quantizer level= 4 bits; four $P_2(z)$ coefficients at 5 bits each; error signal = 60 bits/frame (i.e. 60 samples at 6 kHz); parameter normalization = 2 bits (to normalize the $P_2(z)$ coefficients to a range of 1 for quantizing accuracy). The transmission coding therefore included a total of 96 bits/frame and a frame rate of 100sec^{-1}, for a total bit rate of 9600 bps. By sampling at a slower frame rate, and using fewer predictor coefficients [for $P_2(z)$] and fewer bits for the error signal, the total bit rate could be reduced to 7200 bps.

In subjective tests it was found that the 9600 bps predictive coding is equivalent in quality to 4.5 bit log PCM, corresponding to a signal-to-quantizing noise ratio of 16.9 dB. At 7200 bps, the predictive coder was found equivalent in quality to 4.1 bit log PCM, with a corresponding signal-to-quantizing ratio of 14.7 dB. Sensitivity to digital errors in the transmission channel was also studied. Resulting error rates and associated qualities were found to be: 10^{-3} and lower, satisfactory; 10^{-2}, marginal performance; 10^{-1}, unacceptable (Kelly and Miller [1967]).

12.6 Parametric Models of the Spectral Envelope

12.6.1 Homomorphic Vocoders

The analyzer and synthesizer operations for a complete homomorphic vocoder are shown in Fig. 12.35. Fig. 12.35a illustrates the analysis. At successive intervals (typically every 20 msec), the input speech signal is multiplied by a data window (a 40 msec Hamming window in this case) and the short-time Fourier transform is computed[8]. For each analysis interval the logarithm of the spectral magnitude is taken to produce the log-spectrum $\hat{S}(\omega)$. A further inverse Fourier transform produces the real, even time function $\hat{s}(t)$ which is defined as the cepstrum (see Sections 5.5.1 and 5.6). The low-time parts of $\hat{s}(t)$ characterize the slow fluctuations in $\hat{S}(\omega)$ due to the vocal-tract resonances, and the high-time parts of $\hat{s}(t)$ characterize the rapid fluctuations in $\hat{S}(\omega)$ due to vocal excitation properties. The high-time part of $\hat{s}(t)$ is used for voiced-unvoiced analysis and for pitch extraction, in accordance with the techniques described in Sections 5.5.1 and 5.6.

The final step in the analysis is to derive an equivalent minimum-phase description of the vocal-tract transmission by truncating and saving the positive low-time part of the cepstrum[9]. This is accomplished by multiplication with the time window $h(t)$. The result is $c(t)$ which together with the excitation information constitute the transmission parameters. The transform of $c(t)$ has a spectral magnitude illustrated by the dashed curve in $\hat{S}(\omega)$.

Synthesis is accomplished from $c(t)$ and the excitation information as shown in Fig. 12.35b. Periodic pulses, generated at the analyzed pitch, are used for synthesis of voiced sounds, and uniformly spaced pulses of random polarity are used for unvoiced sounds. The transmitted $c(t)$ is Fourier transformed, exponentiated (to undo the log-taking of the analysis), and an inverse transform yields a minimum-phase approximation to the vocal-tract impulse response. This impulse response is convolved with the excitation pulses to produce the output signal.

The system of Fig. 12.35 was implemented digitally on a general-purpose computer. Fast Fourier transform techniques and fast convolution techniques were used for the computations. The spectral analyses consisted of 512-point discrete Fourier transforms corresponding to a spectral resolution of approximately 20 Hz. Cepstrum computations also consisted of 5l2-point inverse transforms. Spectra and cepstra were computed at 20-msec intervals along the input speech waveform and $c(t)$ was described by the first 32 points of the cepstrum. Linear interpolation over the 20 msec intervals was used for the excitation and impulse response data. Listening tests performed on the system in a back-to-back mode yielded judgments of good quality and natural sound. In a separate experiment the e(t) data were reduced to 26 in

---

[8] See Section 5.1.1 for properties of the short-time Fourier transform.

[9] The minimum-phase properties of this function are not obvious. A proof can be found in (Oppenheim *et al.* [1968])

number and quantized to six bits each for a transmission rate of 7800 bits/sec. At this bit rate no noticeable degradation was reported (Oppenheim [1969]).

A further study of the homomorphic vocoder utilized a time-varying data window for analysis and a digital implementation for transmission at 3700 bits/sec (J. C. Hammett [1971]). At this bit rate, a signal of good quality was reported, with some reduction in naturalness.

Another study has applied predictive coding (see Section 12.5) to the transmission of the homomorphic vocoder signals. Using this technique, transmission of spectral information was digitally implemented for a data rate of 4000 bits/sec with modest impairment in quality. Listening tests concluded that spectral information digitized to around as 5000 bits/sec permits a quality indistinguishable from the unquantized system (Weinstein [1966]).

12.6.2 Maximum Likelihood Vocoders

All vocoder devices attempt to represent the short-time spectrum of speech as efficiently as possible. Exact reproduction of the waveform is not necessary. Some devices, such as channel vocoders, depend upon a frequency-domain transformation of the speech information, while others, such as correlation vocoders (Section 12.4) and orthogonal function vocoders (Section8.6), use strictly a time-domain representation of the signal.

In all vocoder devices, the greatest step toward band conservation derives from observing the source-system distinctions in the production of speech signals (see Fig. 12.1). Vocal excitation information and system function data are treated separately, and decisions about voiced-unvoiced excitation and pitch-period measurement are typically made. Devices which do not make the source-system distinction and which do not perform pitch extraction–such as the voice-excited vocoder and some transmission methods described in later sections of this chapter–derive their bandsaving solely from the ear's acceptance of a signal having a short-time spectrum similar to that of the original speech. Their representation of the signal is commensurately less efficient.

Differences among vocoder devices lie in how they attempt to represent the perceptually-important information in the short-time speech spectrum. The channel vocoder merely samples the spectrum at prescribed frequency intervals and transmits these values. An orthonormal expansion of the amplitude spectrum aims to give adeq uate definition of the spectrum through a few coefficients of a prescribed set of basis functions. The formant vocoder assumes a pole-zero model for the vocal transmission and aims to locate the first few formant frequencies to effect an efficient description of the whole spectrum. The time-domain approach of the correlation vocoder transmits samples of the correlation function and synthesizes a wave composed of the even, truncated correlation funciton. The Laguerre vocoder, another time-domain method, uses an orthonormal expansion of the short-time correlation function and attempts

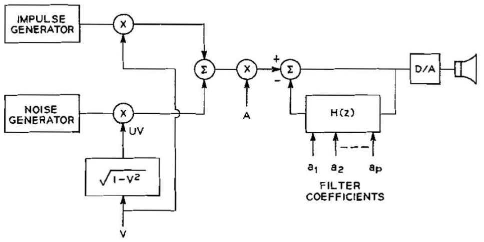


Figure 12.36: Synthesis method for the maximum likelihood vocoder. Samples of voiced and voiceless excitation are supplied to a recursive digital filter of p-th order. Digital-to-Analog (D/A) conversion produces the analog output. (After (Itakura and Saito [1968]))

to represent it by a few coefficients.

Another technique, called the Maximum Likelihood Method (Itakura and Saito [1968]), attempts to combine the advantages of time-domain processing and formant representation of the spectrum. The method is also amenable to digital implementation.

An all-pole model of the power spectrum of the speech signal is assumed. Zeros are omitted because of their lesser importance to perception and because their effect can be represented to any accuracy by a suitable number of poles. The synthesizer includes a recursive digital filter, shown in Fig. 12.36, whose transmission function in z-transform notation is

$$\begin{aligned} T(z) &= \frac{1}{1+H(z)} \\ &= \left[\frac{1}{1+a_1z^{-1}+a_2z^{-2}+\cdots+a_pz^{-p}}\right] \end{aligned} \qquad (12.261)$$

where $z^{-l} = e^{-sD}$ is the delay operator, D is the sampling interval and s is the complex frequency (see, for example, Section 11.5).

The complex roots of the denominator polynomial are the complex formants (bandwidths and frequencies) used to approximate the speech signal. The coefficients, a_i, of the denominator polynomial are obtained from time-domain calculations on samples of a short segment of the speech waveform; namely, $s_l, s_2 \ldots s_N$, where $N \gg p$. Under the assumption that the waveform samples s_i are samples of a random gaussian process, a maximum likelihood estimate is obtained for the a_i's. This estimate corresponds to minimization of a function of the logarithmic difference between the power spectrum of the filter $|T(z)|^2$ and the short-time power spectrum of the signal samples

$$S(\omega) = \frac{1}{2\pi N}\left|\sum_{n=1}^{N} s_n e^{jn\omega D}\right|^2 . \qquad (12.262)$$

The minimization results in a fit which is more sensitive at the spectral peaks

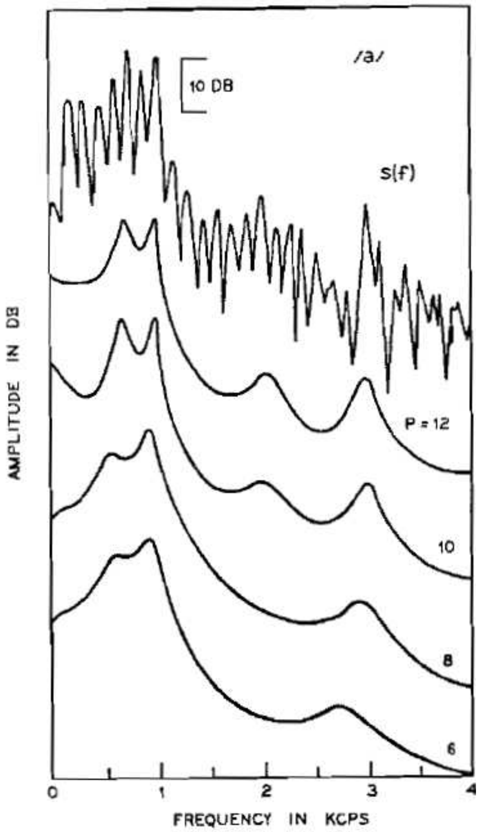


Figure 12.37: Approximations to the speech spectrum envelope as a function of the number of poles of the recursive digital filter. The top curve, $S(f)$, is the measured short-time spectral density for the vowel /ɑ/ produced by a man at a fundamental frequency 140 Hz. The lower curves show the approximations to the spectral envelope for $p = 6, 8, 10$ and 12. (After (Itakura and Saito [1970]))

than in the valleys between formants. Perceptually this is an important feature of the method. The fit of the all-pole model to the envelope of the speech spectrum is illustrated in Fig. 12.37.

The maximum likelihood estimate of the filter coefficients is obtained from the short-time correlation function

$$\Phi_i = \frac{1}{N} \sum_{j=1}^{N-i} s_j s_{j+i}, \quad (i = 0, 1, \ldots, N-1) \tag{12.263}$$

by solving the set of equations

$$\sum_{i=1}^{p} \Phi_{|i-j|} a_i = -\Phi_j, \quad (j = 1, 2, \ldots, p) \tag{12.264}$$

The maximum likelihood estimate also produces the amplitude scale factor for matching the speech signal power spectrum, namely

$$A^2 = \sum_{i=-p}^{p} A_i \Phi_i,$$

where

$$A_i = \sum_{j=0}^{p} a_j a_{j+|i|}; \quad a_0 = 1, \quad a_k = 0 (k > p). \tag{12.265}$$

As shown in Fig. 12.36, excitation of the synthesizer follows vocoder convention and uses a pulse generator and a noise generator of the same average power. Extraction of pitch period T is accomplished by a modified correlation method which has advantages similar to the cepstrum method, but relies strictly upon time domain techniques and does not require transformation to the frequency domain. A voicing amplitude signal, V, is also derived by the pitch extractor. The voiced and voiceless excitations are mixed according to the amplitude of the voicing signal, V. The unvoiced (noise) excitation level is given by $UV = \sqrt{l - V^2}$. The mixing ratio therefore maintains constant average excitation power. Overall control of the mixed excitation, by amplitude signal A, completes the synthesis[10].

Typical parameters for the analysis and synthesis are: sampling rate of input speech, $1/D = 8$ kHz; number of poles, $p = 10$; and number of analyzed samples, $N = 240$ (i.e., 30 msec duration). For transmission purposes, the control parameters are quantized to: 9 bits for each of the 10 a_i's, and 6 bits for each of the three excitation signals. Sampling these quantized parameters at 50 sec^{-1} yields a 5400 bit/sec encoding of the signal for digital transmission. The technique is demonstrated to be substantially better than digitized channel vocoders (Itakura and Saito [1968]).

Furthermore, the maximum likelihood method has been shown to be valuable for automatic extraction of formant frequencies and formant bandwidths. The complex roots z_i of $[1 + H(z)]$ in (12.261) give the real and imaginary parts of the formant frequencies, i. e., their bandwidths and center frequencies. Given $H(z)$ as defined by the coefficients a_i, a root-finding algorithm is applied to determine the z_i. Formant tracking tests on real speech show that the method with $p = 10$ produces accurate estimates of formant bandwidths and frequencies. An example of automatic formant tracking for a five-vowel sequence is shown in Fig. 12.38 (Itakura and Saito [1970]).

12.6.3 Linear Prediction Vocoders

Another time-domain vocoder method for speech analysis and synthesis employs the properties of linear prediction (Atal and Hanauer [1971a,b]). This method also utilizes

---

[10] Note that while the coefficients a_i are derived from the short-time correlation function Φ_i, the synthesis method utilizes a recursive filter and avoids the "truncation" distortion present in the open-loop synthesis of the correlation vocoder (see Section 12.4).

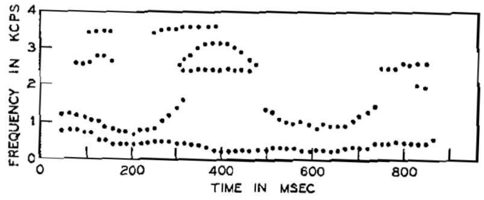


Figure 12.38: Automatic tracking of formant frequencies determined from the polynomial roots for $p = 10$. The utterance is the five-vowel sequence /a,o,i,u,e/. (After (Itakura and Saito [1970]))

an all-pole recursive digital filter excited either by a pitch-modulated pulse generator or a noise generator to synthesize the signal. The filter coefficients in this case represent an optimum linear prediction of the signal. The coefficients are determined by minimizing the mean square error between samples of the input signal and signal values estimated from a weighted linear sum of past values of the signal[11]. That is, for every sample of the input signal, s_n, an estimate $\hat{s}_n$ is formed such that

$$\hat{s}_n = \sum_{k=1}^{p} a_k s_{n-k}.$$

The filter coefficients, a_k, are determined by minimizing $\overline{(s_n - \hat{s}_n)^2}$ over an analysis interval that is typically a pitch period, but which may be as small as 3 msec for $p = 12$ and a sampling rate of 10kHz. The a_k's are given as a solution of the matrix equation

$$\Phi a = \psi \tag{12.266}$$

where a is a p-dimensional vector whose k-th component is a_k, Φ is a $(p \times p)$ covariance matrix with term ϕ_{ij} given by

$$\phi_{ij} = \sum_{n} s_{n-i} s_{n-j}, \quad i = 1, \ldots, p; j = 1, \ldots, p \tag{12.267}$$

and ψ is a p-dimensional vector with the j-th component $\psi_j = \phi_{j0}$ and the sum extends over all speech samples N in a given analysis interval. Since the matrix Φ is symmetric and positive definite, Eq. (12.266) can be solved without matrix inversion. These relations are similar to those obtained from the Maximum Likelihood method [See Eq. (12.264)] except for the difference in the matrix Φ. The two solutions approach each other for the condition $N \gg p$.

Synthesis is accomplished as shown in Fig. 12.39. Excitation either by pitch-modulated pulses or by random noise is supplied to a recursive filter formed from the linear predictor. The amplitude level, A, of the excitation is derived from the rms

---

[11] A general discussion of the theory of optimum linear prediction of signals is given in Section 12.5.

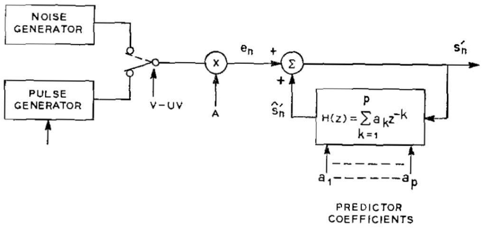


Figure 12.39: Synthesis from a recursive digital filter employing optimum linear prediction. (After citeAtal71b)

value of the input speech wave. The filter transmission function, is

$$T(z) = \frac{1}{1 - H(z)}, \tag{12.268}$$

where

$$H(z) = \sum_{k=1}^{p} a_k z^{-k}$$

which, except for the sign convention, is the same as the Maximum Likelihood method (Section 12.6.2). The filter coefficients a_k account both for the filtering of the vocal tract and the spectral properties of the excitation source. If e_n is the nth sample of the excitation, then the corresponding output sample of the synthesizer is

$$s'_n = e_n + \sum_{k=1}^{p} a_k s'_{n-k}$$

where the primes distinguish the synthesized samples from the original speech samples. The complex roots of $[1 - H(z)]$ in (12.268) therefore include the bandwidths and frequencies of the speech formants. The filter coefficient data can be transmitted directly as the values of the a_k, or in terms of the roots of $[l - H(z)]$. The latter requires a root-finding calculation. Alternatively, the coefficient data can be transmitted in terms of the correlation functions ϕ_{ij}. Further, it can be shown that the recursive filter function describes an equivalent hard-walled pipe composed of right-circular sections in cascade. Its area is expected to be similar to that of the real vocal tract. The area data therefore provide an equivalent form for the coefficient information. Because they can insure a stable filter function, the area and correlation functions are attractive for transmission and interpolation of the control data.

In one implementation, extraction of the pitch period, T, is accomplished by calculating the short-time autocorrelation function of the input speech signal after it has been raised to the third power. This exponentiation emphasizes the pitch periods of voiced passages. The voiced-unvoiced decision, V-UV, is based on the peak amplitude of the correlation function and on the density of zero crossings in the

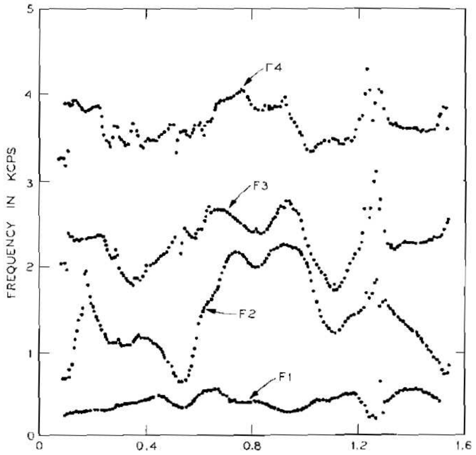


Figure 12.40: Formant frequencies determined from the recursive filter coefficients. The utterance is the voiced sentence “We were away a year ago” produced by a man at an average fundamental frequency of 120 Hz. (After (Atal and Hanauer [1971b]))

speech wave. Another implementation uses the error $(s_n - \hat{s}_n)$ and a peak-picking algorithm to determine the pitch period. Good-quality synthesis at a digital bit rate as low as 3600 bps has been reported for $p = 12$ (Atal and Hanauer [1971b]).

Because the roots of polynomial $[1 - H(z)]$ describe the complex formant frequencies, the linear prediction method is also effective for extracting formant bandwidths and center frequencies. For $p = 12$ the accuracy in obtaining formant frequencies is considered to be within perceptual tolerances. An example of formant extraction from a voiced sentence is shown in Fig. 12.40 (Atal and Hanauer [1971b]).

The Linear Prediction Vocoder and the Maximum Likelihood Vocoder implement their analysis-synthesis procedures in much the same way. Markel has pointed out that they are fundamentally similar, and that both utilize an analysis technique devised earlier by Prony (Markel [1971]). Further, an inverse digital filter, designed along the principles of Eq. (12.261) and (12.268) and Fig. 12.36 and 12.39, has also been found useful for automatic formant extraction (Markel [1972]).

12.6.4 Articulatory Vocoders

An attractive approach to the general vocoder problem is to code speech in terms of articulatory parameters. Such a description has the advantage of duplicating, on a direct basis, the physiological constraints that exist in the human vocal tract. Continuous signals that describe the vocal transmission would then produce all sounds, consonants and vowels.

The idea is to transmit a set of data which describes the tract configuration and its excitation as functions of time. The nature of the analysis–although neither its completeness nor sufficiency–is exemplified by the articulatory-domain spectral

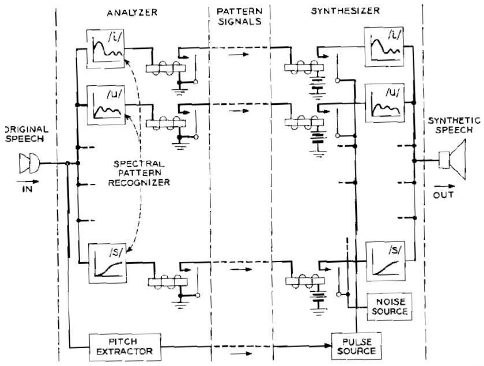


Figure 12.41: Phonetic pattern-matching vocoder. (After (Dudley [1958]))

matching techniques described in Section 5.7, Chapter 5. The synthesizer could be a controllable vocal tract model, such as described in Section 11.5, Chapter 11, or some equivalent device. At the present time no complete vocoder system based upon these principles has been demonstrated. However, the approach appears to be promising and to have much to recommend it. Its success will depend largely upon the precision with which articulatory data can be obtained automatically from the acoustic signal. As the discussion of Chapter 5 has indicated, computer techniques may provide the requisite sophistication for the analysis.

12.6.5 Pattern-Matching Vocoders

Another variation of the vocoder involves classification of the frequency*vs* amplitude spectral information of the channel signals into a limited number of discrete patterns citeSmith57. In one such study (Dudley [1958]), spectral pattern are associated with phonetic units of speech. The sound analysis is carried out according to the pattern recognition scheme described in Section 10.1, Chapter V. At any instant, the best match between the short-time speech spectrum and a set of stored spectral patterns is determined. A code representing the matching pattern is signaled to a vocoder synthesizer, along with conventional pitch and voiced-unvoiced data. New information is signalled only when the phonetic pattern changes. At the receiver, a set of spectral amplitude signals, approximating the signalled pattern, are applied to the modulators of the synthesizer. The pitch signal supplies the appropriate excitation. Filter circuits are included to provide smooth transitions from one sound pattern to the next.

An early version of the device used a ten-channel vocoder and only ten stored patterns. It is illustrated in Fig. 12.41. The stored patterns corresponded to the steady-state spectra of four consonant continuants and six vowels (/s,f,r,n/, and /i,ɪ,ɛ,ɑ,o,u/, respectively). For one speaker (from whose speech the spectral patterns were derived), digits uttered in isolation were recognized by two listeners with scores of 97 and 99 per cent correct, respectively. On common monosyllables, however, the intelligibility fell to around 50%. The addition of six more patterns increased the score by a small amount. The bandwidth required for transmission was only on the order of 50Hz, or around 60 times less than that for a conventional voice channel! While the intelligibility and quality of the speech processed by the device are clearly inadequate for most applications, the implementation does indicate the possibilities of narrow-band transmission for restricted message ensembles and limited speaker populations.

The obvious question suggested by the rather surprising performance with only ten stored patterns is how many stored spectral patterns would be needed to approach the performance of the conventional vocoder? At least one investigation has aimed to examine the question (Smith [1957, 1963]). The outputs of the analyzer of a channel vocoder are sampled at 50Hz, normalized in amplitude, and quantized. The digital description of the short-time spectrum is then compared to a large library of digital patterns stored in a rapid-access memory. No requirement is imposed that these spectral patterns correspond to specific phonetic units of speech. Using digital processing techniques, the best fitting pattern is selected and its code transmitted. The objective is to determine the smallest population of patterns necessary to meet given performance criteria. The processing cannot, of course, result in better speech quality than provided by the conventional vocoder. It may, however, afford a useful bandsaving beyond that the of channel vocoder. Digital data rates for the transmission of the spectral patterns and excitation are estimated to be on the order of 400 to 800 bits/sec (Smith [1957, 1963]).

12.6.6 Formant Vocoders

The results of the acoustic analyses in Chapter 4 suggest that one efficient way to code speech is in terms of the vocal mode pattern. The results show, for example, that adjacent values of the short-time amplitude spectrum are not independent, but are closely correlated. In fact, specification of the complex poles and zeros is equivalent to specifying the spectrum at all frequencies. The formant vocoder aims to exploit this fact and to code the speech signal in terms of the mode pattern of the vocal tract. Because it does not use multiple control signals to describe strongly correlated points in the speech spectrum, the formant-vocoder hopes to achieve a band-saving in excess of that accomplished by the channel vocoder. The practicability of formant vocoders depends upon how well formant-mode data, or the equivalent, can be automatically derived. In addition, excitation information must be provided as in the channel vocoder.

A number of formant-vocoder systems have been designed and instrumented. Although it is not possible to treat each in detail, this section proposes to indicate typical circuit realizations and the results obtained from them.

Formant-vocoders generally divide into two groups–essentially defined by the synthesis philosophies set forth in Chapter 11. That is, the classification relates to the cascade and parallel connections of the synthesis circuits. The cascade approach

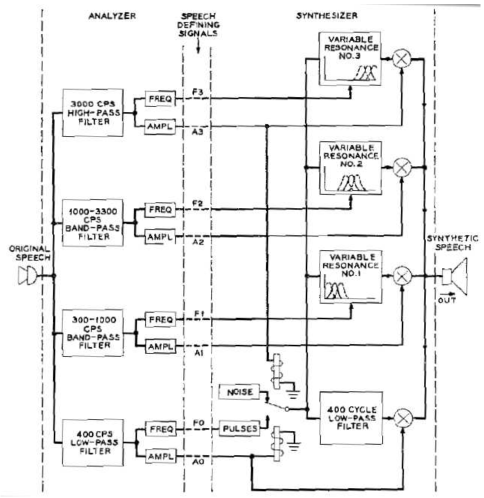


Figure 12.42: Parallel-connected formant vocoder. (After (Munson and Montgomery [1950]))

strives to reconstruct the signal by simulating, usually termwise, the perceptually significant pole and zero factors of the vocal transmission. The complex frequencies of the poles and zeros, and the excitation data (pitch and voiced-unvoiced) are the coding parameters.

The parallel connection attempts to reconstruct the same signal in a different, but equivalent, way–namely, from information on the frequencies of the formants (poles) and their spectral amplitudes (residues). Ideally, the mode frequencies and their residues are specified in complex form. The complex residues are equivalent to specification of the spectral zeros. The discussion of Section 11.4, Chapter 11, has set down in some detail the relations between the cascade and parallel representations of the speech signal. If the requisite data for either synthesis arrangement can be obtained automatically and with sufficient accuracy, the formant vocoder has the potential for producing intelligible speech of perhaps better quality than that of the channel vocoder. Because it attempts to duplicate the vocal mode structure, it innately has the potential for a better and more natural description of the speech spectrum.

One of the earliest, complete formant-vocoder systems was a parallel arrangement (Munson and Montgomery [1950]). It is illustrated in Fig. 12.42. At the analyzer, the input speech band is split into four subbands. In each band, the average frequency of axis-crossings, F, and the average rectified-smoothed amplitude, A, are measured[12]. Signal voltages proportional to these quantities are developed. These eight parameters, which approximate the amplitudes and frequencies of the formants and of voicing, are transmitted to the synthesizer.

---

[12] Note in this design the highest two bands normally contain more than a single formant. Their amplitude and frequency measures primarily reflect the most prominent formants in these ranges.

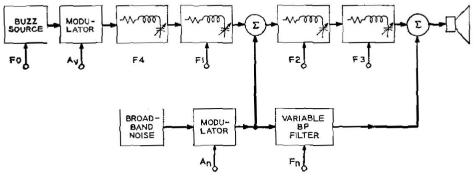


Figure 12.43: Cascade-connected formant vocoder. (After (House [1956]))

The synthesizer contains excitation circuitry, three variable resonators connected in parallel, and a fourth parallel branch with a fixed low-pass filter. Voiced (pulse) excitation of the parallel branches is signalled by the voicing amplitude, $A0$. The $A0$ control also determines the amplitude of the signal passing the fixed low-pass branch of the circuit. As in the channel vocoder, the frequency of the pulse source is prescribed by $F0$. Unvoiced (noise) excitation of the parallel branches is determined by amplitude $A3$. The amplitudes and frequencies of the three formant branches are continuously controlled and their outputs combined.

Intelligibility scores reported for the system were approximately 100% for vowel articulation and about 70% for consonant articulation. The total bandwidth occupancy of the eight control signals was about 300 Hz, or about the same as for the channel vocoder. A number of different versions of parallel-connected formant vocoders have subsequently been constructed (for example, (Chang [1956], Campanella *et al.* [1962], Ayers [1959], Stead and Jones [1961], Howard [1956])). Two of these will receive further comment in the following section on digitalizing and multiplexing.

An early effort at realizing a cascade system also investigated the effects of severe band-limitation of the control signals (Flanagan [1971]). One synthesizer configuration considered in the study is shown in Fig. 12.43. The control data employed were pitch F0; amplitude of voicing AV; three formant frequencies F1, F2, F3 (covering the range approximately 100 to 3000 Hz); a single, relatively-broad, fricative noise resonance FN (the major resonance in the range 3000 to 7000 Hz); and the amplitude of noise excitation AN.

The formant frequency data were obtained from a peak-picking analyzer as described in Section 5.5, Chapter 5. The amplitude of voicing was determined from the rectified-smoothed signal in a lowpass band of the original speech, and the amplitude of noise excitation was determined from the rectified-smoothed signal in the 3000 to 7000 Hz band. Pitch was measured with a fundamental-extracting circuit, as in the channel vocoder. Each of the seven control signals was band-limited to slightly less than 10 Hz by a low-pass filter, so that the total bandwidth occupancy was on the order of 60 Hz.

All voiced sounds were produced by the upper resonator string of the circuit, following strictly the cascade approach. The unvoiced sounds were produced by a cascade-parallel connection which introduced zeros, as well as poles, into the transmission. Data on frequencies of zeros, as such, were not transmitted.

Although the band saving was high, detailed articulation testing of the system showed its performance to be relatively poor. In nonsense monosyllables, the vowel articulation was on the order of 82%. For the consonants, the mean score was 27%. Confusion-matrix analysis of listener responses showed that voiced-unvoiced errors were few. Errors in discriminating voiced-stops and nasals, however, were relatively numerous, the synthesizer being congenitally incapable of simulating these sounds. In addition, errors in discriminating affricates and stops were due in large part to temporal imprecision resulting from the severe band-limitation of the control signals.

A more recent, digital computer simulation of an improved version of the synthesizer corrects some of the shortcomings (Coker [1965]). It provides for an additional pole-zero pair in the voiced branch and a controllable zero in the unvoiced branch (see Fig. 11.14 and Section 11.4, Chapter 11). When combined with a sophisticated digitally-simulated formant analyzer, the performance as a complete real-time formant vocoder is unusually good (Coker [1965]). Theformant analysis in the computer is accomplished by a detailed matching of the real speech spectrum by a pole-zero model spectrum, similar to the analysis-by-synthesis procedure. (See Section 5.5.1.) The digital processing provides much greater accuracy than can be obtained with analog equipment. The precision in the formant tracking, and the more detailed accounting for system and excitation characteristics by means of the additional pole-zero pair, contribute significantly to the quality of the synthetic speech.

A further word may be appropriate concerning the relative merits of parallel versus cascade connections, and about the approach which may result in the most efficient and practical set of parameters. The vocal transmission for vowel sounds contains only poles. The residues in these poles are therefore functions only of the pole frequencies. Given the formant frequencies, any formant amplitude specification is redundant because the amplitudes are implied by the frequencies. In this case, the cascade synthesizer provides correct formant amplitudes automatically from formant frequency data alone. For nonvowel sounds the vocal transmission can have zeros, one or two of which may prove to be perceptually significant. To simulate these factors, the cascade synthesizer requires controllable antiresonances. Again, given the proper pole and zero frequencies, spectral amplitudes are automatically accounted for.

The parallel synthesizer, on the other hand, requires the significant pole frequencies and, ideally, the complex residues in these poles. The residues, in effect, specify the spectral zeros. The contribution to perception of the residue phases is modest but not negligible (Flanagan [1965]). (See Section 11.4.7.) A relevant question about formant synthesis is then "Which is easier to analyze automatically, the frequencies of spectral zeros or the amplitudes and phases of spectral maxima?" The question is complicated by one other matter–the excitation source. What are its perceptually important characteristics? Are they easier to include in one model than in the other? At the present stage of study, the ultimate practical choice is not clear.

12.7 Quantized Linear Prediction Coefficients

Recall that the linear prediction synthesis filter is given by:

$$H(z) = \frac{1}{1 - \sum_{i=1}^{p} \alpha_i z^{-i}} = \prod_{i=1}^{p} \frac{1}{1 - r_i z^{-1}} \tag{12.269}$$

Discussion in Sec. 5.2.4 demonstrated that $H(z)$ is stable if and only if $|r_i| < 1$, that is, if and only if the reflection coefficients are $|k_i| < 1$. Filter coefficients generated using the Levinson-Durbin recursion 5.2 are guaranteed to be stable. In order to avoid instability in the speech decoder, however, it is necessary to quantize the LPCs using a method that guarantees stability, in other words, using a method that guarantees $|k_i| < 1$.

A small quantization error in one of the direct-form coefficients $\hat{a}_i$ can easily make $\hat{H}(z)$ unstable. For example, this filter is stable:

$$H(z) = \frac{1}{1 - 0.4z^{-1} + 0.1z^{-2} + 0.28z^{-3} + 0.49z^{-4}} \tag{12.270}$$

If a_4 is changed from 0.49 to 0.52, the filter is unstable:

$$\hat{H}(z) = \frac{1}{1 - 0.4z^{-1} + 0.1z^{-2} + 0.28z^{-3} + 0.52z^{-4}} \tag{12.271}$$

If a_1 is different, however, the same change in a_4 leaves a stable filter:

$$\hat{H}(z) = \frac{1}{1 + 0.4z^{-1} + 0.1z^{-2} + 0.28z^{-3} + 0.52z^{-4}} \tag{12.272}$$

As demonstrated in Eqs. 12.270 through 12.272, there is no simple test that can be applied to the direct-form LPCs in order to determine whether or not the LPC synthesis filter is stable. In order to guarantee stability of the synthesis filter, therefore, it is prudent to quantize either k_i or r_i, and design the quantizer levels so that $|\hat{k}_i|$ or $|\hat{r}_i|$ is always less than 1.0.

12.7.1 Log Area Ratios

Stability of the synthesis filter is not the only consideration. Changes in k_i have a much larger effect on the synthesized speech spectrum if $|k_i| \approx 1$ than if $|k_i| << 1$, as shown in figure 12.7.1.

The sensitivity problem shown in Fig. 12.7.1 can be solved using companded quantization. If a reflection coefficient is near unit magnitude (and therefore quantization errors have large effect), it should be quantized in a way that guarantees small quantization errors. If a reflection coefficient is near zero magnitude (and therefore

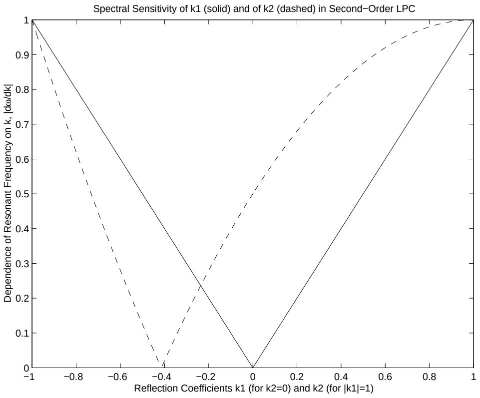


Figure 12.44: Spectral sensitivity to changes in the reflection coefficients.

quantization errors have little effect), it may be quantized coarsely. Sec. 12.2.3 demonstrated that level-dependent quantization errors can be achieved by companding the coefficient prior to quantization:

$$k_i \rightarrow \boxed{\text{Expand}} \rightarrow g_i \rightarrow \boxed{\text{Linear PCM}} \rightarrow \hat{g}_i \rightarrow \boxed{\text{Compress}} \rightarrow \hat{k}_i \tag{12.273}$$

For example, the log-area ratio transform in Eq. 12.274 stretches the k_i axis near values of ±1:

$$g_i = \log\left(\frac{1-k_i}{1+k_i}\right) \tag{12.274}$$

Remember that the reflection coefficients k_i can be used to define a stylized vocal tract model, with reflection coefficients of $r_L = 1$ at the lips, $r_g = 0$ at the glottis, and $r_i = -k_{p-i}$ $(i = 0, \ldots, p-1)$, as shown in figure 12.7.1.

If the cross-sectional areas are A_i, A_{i+1}, then

$$r_i = -k_{p-i} = \frac{A_{i+1} - A_i}{A_{i+1} + A_i}, \tag{12.275}$$

and

$$\frac{A_{i+1}}{A_i} = \frac{1-k_{p-i}}{1+k_{p-i}}. \tag{12.276}$$

So the "expanded" reflection coefficient is really a "log area ratio:"

$$g_i = \log\left(\frac{1-k_i}{1+k_i}\right) = \log\left(\frac{A_{p-i+1}}{A_{p-i}}\right) \tag{12.277}$$

12.7.2 Line Spectral Frequencies

Log-area-ratio quantization has the advantage of simplicity, but a slightly more complicated method of LPC quantization has been adopted into most recent low-bit-rate

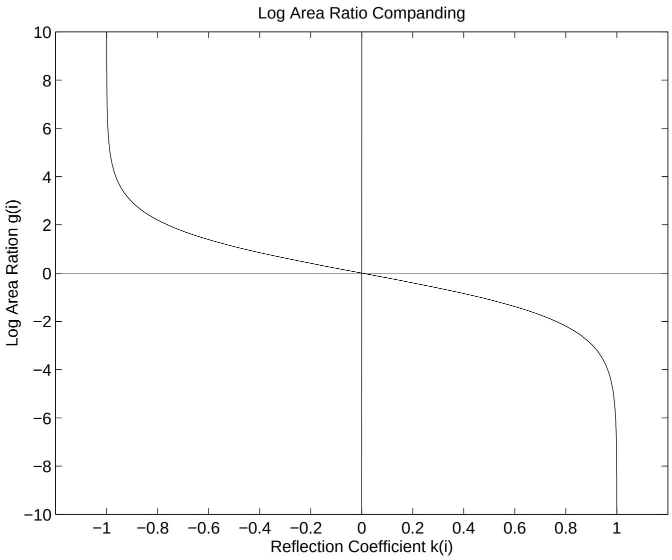


Figure 12.45: Log area ratio companding

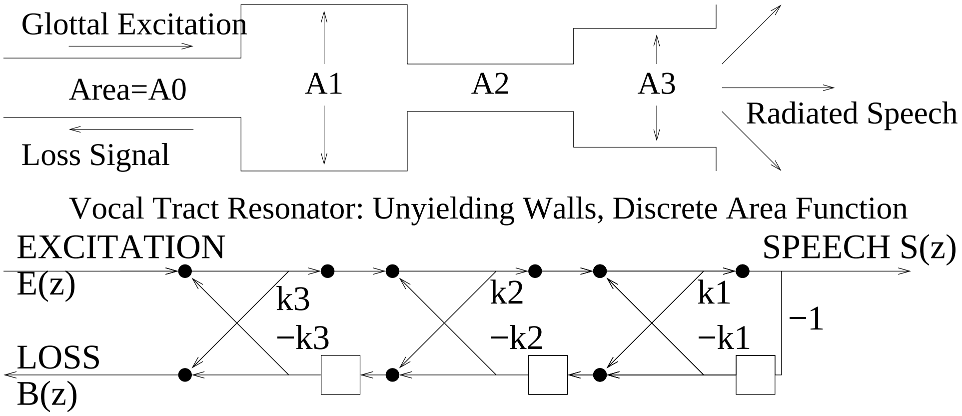


Figure 12.46: Acoustic resonator and lattice model with a matched impedance termination at the glottis.

coding standards. Linear Prediction can be viewed as an inverse filtering procedure in which the speech signal is passed through an all-zero filter $A(z)$. The filter coefficients of $A(z)$ are chosen such that the energy in the output, i.e. the residual or error signal, is minimized. Alternatively, the inverse filter $A(z)$ can be transformed into two other filters $P(z)$ and $Q(z)$. These new filters turn out to have some interesting properties, and the representation based on them, called the *line-spectrum pairs* (Sugamura and Itakura [1981], Soong and Juang [1984]), has been used in speech coding and synthesis applications.

Let $A(z)$ be the frequency response of an LPC inverse filter of order p.

$$A(z) = -\sum_{i=0}^{p} a_i z^{-i} \tag{12.278}$$

with $a_0 = -1$.

The a_i's are real and all the zeros of $A(z)$ are inside the unit circle.

If we use the lattice formulation of LPC we arrive at a recursive relation between the mth stage ($A_m(z)$) and the one before it ($A_{m-1}(z)$). For the p-th order inverse filter, we have:

$$A_p(z) = A_{p-1}(z) - k_p z^{-p} A_{p-1}(z^{-1})$$

By allowing the recursion to go one more iteration, we obtain:

$$A_{p+1}(z) = A_p(z) - k_{p+1} z^{-(p+1)} A_p(z^{-1}) \tag{12.279}$$

Suppose that we choose the new reflection coefficient to be a perfect, lossless reflection at the lips, i.e., $k_{p+1} = \pm 1$; then we can define two new polynomials as follows:

$$P(z) = A(z) - z^{-(p+1)} A(z^{-1}) \tag{12.280}$$

$$Q(z) = A(z) + z^{-(p+1)} A(z^{-1}) \tag{12.281}$$

Since k_{p+1} is a lossless termination, and every tube section is lossless, the impedance $Z_T(z)$ must also be lossless. This means that the poles and zeros e^{jp_n} and e^{jq_n} are on the unit circle.

Physically, $P(z)$ and $Q(z)$ can be interpreted as the inverse transfer function of the vocal tract for the *open glottis* and *closed glottis* boundary conditions, respectively (Furui [1989]), and $P(z)/Q(z)$ is the driving-point impedance of the vocal tract as seen from the glottis (Hasegawa-Johnson [2000]).

If p is odd, the formulae for p_n and q_n are as follows:

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) = \prod_{n=1}^{(p+1)/2} (1 - e^{jp_n} z^{-1})(1 - e^{-jp_n} z^{-1}) \tag{12.282}$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) = (1 - z^{-2}) \prod_{n=1}^{(p-1)/2} (1 - e^{jq_n} z^{-1})(1 - e^{-jq_n} z^{-1}) \tag{12.283}$$

The LSFs have some interesting characteristics: the frequencies $\{p_n\}$ and $\{q_n\}$ are related to the formant frequencies; the dynamic range of $\{p_n\}$ and $\{q_n\}$ is limited and the two alternate around the unit circle ($0 \leq p_1 \leq q_1 \leq p_2 ...$); $\{p_n\}$ and $\{q_n\}$ are correlated so that intra-frame prediction is possible; and they change slowly from one frame to another, hence, inter-frame prediction is also possible. The interleaving nature of the $\{p_n\}$ and $\{q_n\}$ allow for efficient iterative solutions (Kabal and Ramachandran [1986]).

Almost all LPC-based coders today use the LSFs to represent the LP parameters. Considerable recent research has been devoted to methods for efficiently quantizing the LSFs, especially using Vector Quantization (VQ) techniques. Typical algorithms include predictive VQ, split VQ (Paliwal and Atal [1993]), and multi-stage VQ (Paksoy *et al.* [1992], LeBlanc *et al.* [1993]). All of these methods are used in the ITU standard ACELP coder G.729: the moving-average vector prediction residual is quantized using a 7-bit first-stage codebook, followed by second-stage quantization of two subvectors using independent 5-bit codebooks, for a total of 17 bits per frame (ITU-T [1996a], Salami *et al.* [1998]).

The Augmented-Tube Interpretation of Line Spectral Frequencies

Setting $A_{p+1} = 0$ in the concatenated tube model yields:

$$\frac{S(z)}{U(z)} = \frac{S(z)}{E^{(p)}(z) + z^{-(p+1)}B^{(p)}(z)} = \frac{1}{P(z)}, \qquad P(z) \equiv A(z) + z^{-(p+1)}A(z^{-1}) \quad (12.284)$$

Setting $A_{p+1} = \infty$ yields:

$$\frac{S(z)}{U(z)} = \frac{S(z)}{E^{(p)}(z) - z^{-(p+1)}B^{(p)}(z)} = \frac{1}{Q(z)}, \qquad Q(z) \equiv A(z) - z^{-(p+1)}A(z^{-1}) \quad (12.285)$$

Because of symmetry, the roots of both $P(z)$ and $Q(z)$ are on the unit circle. If p is even:

$$P(z) = (1 + z^{-1}) \prod_{n=1}^{p/2} (1 - e^{jp_n} z^{-1})(1 - e^{-jp_n} z^{-1}), \qquad p_n \text{ real} \quad (12.286)$$

$$Q(z) = (1 - z^{-1}) \prod_{n=1}^{p/2} (1 - e^{jq_n} z^{-1})(1 - e^{-jq_n} z^{-1}), \qquad q_n \text{ real} \quad (12.287)$$

The frequencies p_n and q_n, for $1 \leq n \leq p/2$, are called the line spectral frequencies (LSFs). The LSFs have the following useful characteristics:

- LSFs are real, so they are easier to quantize than the LPC roots r_i, which are complex.

- If and only if $1/A(z)$ is stable, the LSFs satisfy: $0 < p_1 < q_1 < p_2 < q_2 < \ldots < \pi$
- The LSFs tend to track the LPC root frequencies $\arg(r_i)$, but...
- The LSFs vary more slowly and smoothly than the LPC roots r_i.
- Efficient algorithms exist for calculating the LSFs.

12.8 Parametric Models of the Spectral Fine Structure

The characteristics of the vocoder excitation signal $u(n)$ change quite rapidly. The energy of the signal may change from zero to nearly full amplitude within one millisecond at the release of a plosive sound, and a mistake of more than about 5ms in the placement of such a sound is clearly audible. The LPC coefficients, on the other hand, change relatively slowly. In order to take advantage of the slow rate of change of LPC coefficients without sacrificing the quality of the coded residual, most LPC-AS coders encode speech using a frame-subframe structure, as depicted in Figure 12.47. A frame of speech is approximately 20ms in length, and is composed of typically 3-4 subframes. The LPC excitation is transmitted once per subframe, while the LPC coefficients are only transmitted once per frame. The LPC coefficients are computed by analyzing a window of speech which is usually longer than the speech frame (typically 30-60ms). In order to minimize the number of future samples required to compute LPC coefficients, many recent LPC-AS coders use an asymmetric window which may include several hundred milliseconds of past context, but which emphasizes the samples of the current frame (Florencio [1993], Salami *et al.* [1998]).

The perceptually weighted original signal $s_w(n)$ and weighted reconstructed signal $\hat{s}_w(n)$ in a given subframe are often written as L-dimensional row vectors S and $\hat{S}$, where the dimension L is the length of a sub-frame:

$$S_w = [s_w(0), \ldots, s_w(L-1)], \qquad \hat{S}_w = [\hat{s}_w(0), \ldots, \hat{s}_w(L-1)] \tag{12.288}$$

The core of an LPC-AS coder is the closed-loop search for an optimum coded excitation vector U, where U is typically composed of an "adaptive codebook" component representing the periodicity, and a "stochastic codebook" component representing the noise-like part of the excitation. In general, U may be represented as the weighted sum of several "shape vectors" X_m, $m = 1, \ldots, M$, which may be drawn from several codebooks, including possibly multiple adaptive codebooks and multiple stochastic codebooks.

$$U = GX, \qquad G = [g_1, g_2, \ldots], \qquad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \end{bmatrix} \tag{12.289}$$

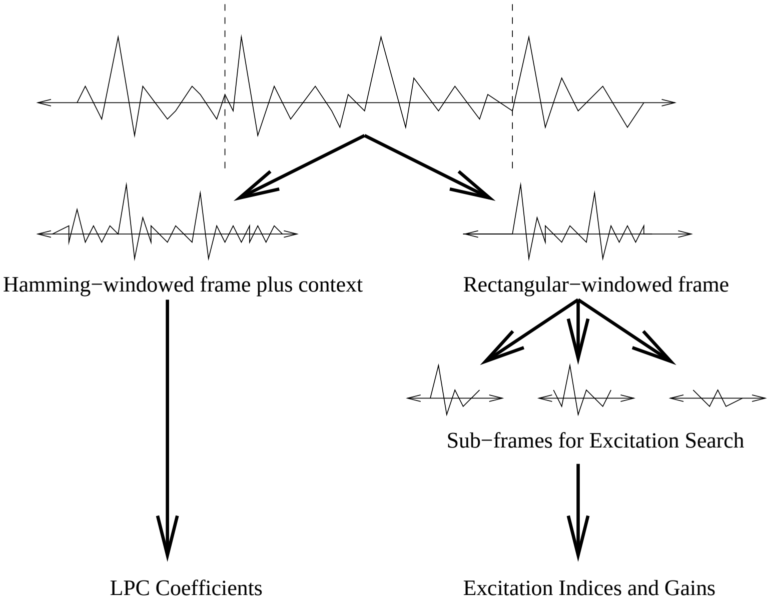


Figure 12.47: The frame/sub-frame structure of most LPC analysis by synthesis coders

The choice of shape vectors and the values of the gains g_m are jointly optimized in a closed-loop search, in order to minimize the perceptually weighted error metric $|S_w - \hat{S}_w|^2$.

The value of S_w may be computed prior to any codebook search by perceptually weighting the input speech vector. The value of $\hat{S}_w$ must be computed separately for each candidate excitation, by synthesizing the speech signal $\hat{s}(n)$, and then perceptually weighting to obtain $\hat{s}_w(n)$. These operations may be efficiently computed, as described below.

Zero State Response and Zero Input Response

Let the filter $H(z)$ be defined as the composition of the LPC synthesis filter and the perceptual weighting filter, thus $H(z) = W(z)/A(z)$. The computational complexity of the excitation parameter search may be greatly simplified if $\hat{S}_w$ is decomposed into the zero-input response (ZIR) and zero-state-response (ZSR) of $H(z)$ (Trancoso and Atal [1986]). Note that the weighted reconstructed speech signal is

$$\hat{S}_w = [\hat{s}_w(0), \ldots, \hat{s}_w(L-1)], \qquad \hat{s}_w(n) = \sum_{i=0}^{\infty} h(i)u(n-i) \tag{12.290}$$

where $h(n)$ is the infinite-length impulse response of $H(z)$. Suppose that $\hat{s}_w(n)$ has already been computed for $n < 0$, and the coder is now in the process of choosing the optimal $u(n)$ for the subframe $0 \le n \le L-1$. The sum above can be divided into two parts: a part which depends on the current subframe input, and a part which does not:

$$\hat{S}_w = \hat{S}_{ZIR} + UH \tag{12.291}$$

where $\hat{S}_{ZIR}$ contains samples of the zero input response of $H(z)$, and the vector UH contains the zero state response. The zero input response is usually computed by implementing the recursive filter $H(z) = W(z)/A(z)$ as the sequence of two IIR filters, and allowing the two filters to run for L samples with zero input. The zero state response is usually computed as the matrix product UH, where

$$H = \begin{bmatrix} h(0) & h(1) & \dots & h(L-1) \\ 0 & h(0) & \dots & h(L-2) \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & h(0) \end{bmatrix}, \qquad U = [u(0), \dots, u(L-1)] \tag{12.292}$$

Given a candidate excitation vector U, the perceptually weighted error vector E may be defined as

$$E_w = S_w - \hat{S}_w = \tilde{S} - UH \tag{12.293}$$

where the target vector $\tilde{S}$ is

$$\tilde{S} = S_w - \hat{S}_{ZIR} \tag{12.294}$$

The target vector only needs to be computed once per subframe, prior to the codebook search. The objective of the codebook search, therefore, is to find an excitation vector U which minimizes $|\tilde{S} - UH|^2$.

Optimum Gain and Optimum Excitation

Recall that the excitation vector U is modeled as the weighted sum of a number of codevectors X_m, $m = 1, \dots, M$. The perceptually weighted error is therefore:

$$|E|^2 = |\tilde{S} - GXH|^2 = \tilde{S}\tilde{S}' - 2GXH\tilde{S}' + GXH(GXH)' \tag{12.295}$$

where prime denotes transpose. Minimizing $|E|^2$ requires optimum choice of the shape vectors X and of the gains G. It turns out that the optimum gain for each excitation vector can be computed in closed form. Since the optimum gain can be computed in closed form, it need not be computed during the closed loop search: instead, one can simply assume that each candidate excitation, if selected, would be scaled by its optimum gain. Assuming an optimum gain results in an extremely efficient criterion for choosing the optimum excitation vector (Atal [1986]).

Suppose we define the following additional bits of notation:

$$R_X = XH\tilde{S}', \qquad \Sigma = XH(XH)' \tag{12.296}$$

Then the mean squared error is

$$|E|^2 = \tilde{S}\tilde{S}' - 2GR_X + G\Sigma G' \tag{12.297}$$

For any given set of shape vectors X, G is chosen so that $|E|^2$ is minimized, which yields

$$G = R_X' \Sigma^{-1} \tag{12.298}$$

If we substitute the minimum-MSE value of G into Equation 12.297, we get

$$|E|^2 = \tilde{S}\tilde{S}' - R_X' \Sigma^{-1} R_X \tag{12.299}$$

Hence, in order to minimize the perceptually weighted MSE, we choose the shape vectors X in order to maximize the covariance-weighted sum of correlations,

$$X_{opt} = \arg\max \left(R_X' \Sigma^{-1} R_X \right) \tag{12.300}$$

When the shape matrix X contains more than one row, the matrix inversion in Equation 12.300 is often computed using approximate algorithms (Atal and Remde [1982]). In the VSELP coder (Gerson and Jasiuk [1991]), X is transformed using a modified Gram-Schmidt orthogonalization so that Σ has a diagonal structure, thus simplifying the computation of Equation 12.300.

12.8.1 Voice-Excited Vocoders

Despite their high potential for transmitting intelligible speech with bandwidth savings on the order of ten-to-one, or more, vocoders have been applied only in special communication situations. Little or no commercial use has been made, largely because speech quality and naturalness suffer in the processing[13]. The resulting synthetic speech tends to have a "machine accent," and its naturalness is less than that of a conventional voice circuit.

The seat of the difficulty is largely the extraction of excitation information–that is, the pitch measurement and the voiced-unvoiced discrimination. The difficult problem of automatic pitch extraction is well known. The device must faithfully indicate the fundamental of the voice over a frequency range of almost a decade (if male and female voices are to be handled) and over a large range of signal intensity. Practically, the pitch extractor must cope with unfavorable conditions where the speech signal may be produced in noisy and reverberant environments. In addition, the signal may suffer band limitation that eliminates the first several lowest harmonics, requiring that the fundamental frequency be generated from some nonlinear operation. These difficulties are compounded by the human ear's ability to detect small imprecisions in pitch data. (See Section 9.2.4, Chapter 9.)

---

[13] Other considerations include the cost of terminal equipment compared to the cost of bandwidth.

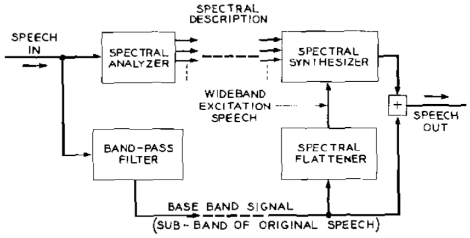


Figure 12.48: Block diagram of voice-excited vocoder. (After (E. E. David [1956], Schroeder *et al.* [1962]))

Some of the many approaches that have been made to the pitch extraction problem have been briefly outlined in Section 5.6, Chapter 5. It suffices here to say that solutions are yet to be implemented to bring the quality of the spectrum channel vocoder up to the quality of conventionally-coded voice circuits. The same general remark applies to the voiced-unvoiced discrimination which is also signalled in the pitch channel.

One method for avoiding the difficulties inherent in automatic analysis of excitation data is the voice-excited vocoder (E. E. David [1956], Schroeder *et al.* [1962]). In this device excitation information is transmitted in an unprocessed, subband of the original speech. At the receiving end, this baseband is put through a nonlinear distortion process to spectrally flatten and broaden it. It is then used as the source of excitation for regular vocoder channels covering the frequency range above the baseband. A block diagram of the arrangement is shown in Fig. 12.48.

The flattened excitation band reflects the spectral line structure of the quasi-periodic voiced sounds and the continuous spectral character of the unvoiced sounds. Because it is derived from a real speech band, it inherently preserves the voiced-unvoiced and pitch information. At some sacrifice in bandwidth, the overall quality of the processed signal can be made comparable to conventional voice circuits. A higher quality signal is therefore realized together with a part of the bandsaving advantage of the channel vocoder.

In one implementation of the device the baseband is taken as 250 to 940 Hz. The frequency range 940 to 3650 Hz, above the baseband, is covered by 17 vocoder channels. The first 14 of these channels have analyzing bandwidths of 150 Hz, and the upper three are slightly wider. The total transmission band occupancy is 1000 to 1200 Hz, yielding a bandwidth compression of about three-to-one. The method of spectral rlattening is shown in Fig. 12.49. The transmitted baseband is rectified and applied to the bandpass filters of the vocoder synthesizer. The filter outputs are peak-clipped to remove amplitude fluctuations. They are then applied as inputs to amplitude modulators which are controlled by the vocoder channel signals.

Intelligibility and speech quality tests, using speech from a carbon button microphone, were carried out to compare the voice-excited vocoder to telephone handset

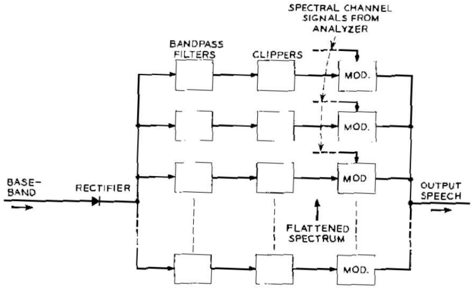


Figure 12.49: Block diagram of the spectral flattener. (After (E. E. David [1956], Schroeder *et al.* [1962]))

speech band limited to the same frequency range citeDavid56, Schroeder62c. To provide a more sensitive test, and to keep intelligibility substantially below 1001%, masking noise was added to provide an 18 db speech-to-noise ratio. Phonetically balanced (PB) words were used in articulation tests (see Section 9.6, Chapter 9). For male speakers, the intelligibility of the voice-excited vocoder was found to be 6.1% less than the carbon-microphone speech of the same bandwidth. For female speakers the intelligibility was 10.1% less than the carbon-microphone speech.

Over-all speech quality of the voice-excited vocoder was assessed, along with that for three other transmission methods, by presenting listeners with sentences in isolation. The subjects were asked to rate each sentence "as good as normal telephone" or "worse than normal telephone." In 72% of the cases, the voice-excited vocoder was rated as good as normal telephone. In the same test, for comparison, a long distance carrier telephone circuit rated 82%, an 1800 Hz lowpass circuit rated 36%, and a regular 18-channel vocoder rated 17%. The results show the voiced-excited system to be better than the spectrum channel vocoder and to approach the quality of conventional voice circuits. Its application, as with similar methods, depends upon desired trade-offs between cost of terminal equipment, amount of bandsaving and signal quality.

Multi-Pulse LPC (MPLPC)

In the multipulse LPC algorithm (Atal and Remde [1982], ITU-T [1996b]), the shape vectors are impulses. U is typically formed as the weighted sum of 4-8 impulses per subframe.

The number of possible combinations of impulses grows exponentially in the number of impulses, so joint optimization of the positions of all impulses is usually impossible. Instead, most MPLPC coders optimize the pulse positions one at a time, using something like the following strategy. First, the weighted zero state response of $H(z)$ corresponding to each impulse location is computed. If C_k is an impulse located at

$n = k$, the corresponding weighted zero state response is

$$C_k H = [0, \ldots, 0, h(0), h(1), \ldots, h(L-k-1)] \tag{12.301}$$

The location of the first impulse is chosen in order to optimally approximate the target vector $\tilde{S}_1 = \tilde{S}$, using the methods described in the previous Section. After selecting the first impulse location k_1, the target vector is updated according to

$$\tilde{S}_m = \tilde{S}_{m-1} - C_{k_{m-1}} H \tag{12.302}$$

Additional impulses are chosen until the desired number of impulses is reached. The gains of all pulses may be re-optimized after the selection of each new pulse (Singhal and Atal [1984]).

Variations are possible. The multipulse coder described in ITU standard G.723.1 transmits a single gain for all of the impulses, plus sign bits for each individual impulse. The G.723.1 coder restricts all impulse locations to be either odd or even; the choice of odd or even locations is coded using one bit per subframe (ITU-T [1996b]). The regular pulse excited LPC algorithm, which was the first GSM full-rate speech coder, synthesized speech using a train of impulses spaced one per 4 samples, all scaled by a single gain term (Kroon *et al.* [1986]). The alignment of the pulse train was restricted to one of four possible locations, chosen in a closed-loop fashion together with a gain, an adaptive codebook delay, and an adaptive codebook gain.

Singhal and Atal demonstrated that the quality of MPLPC may be improved at low bit rates by modeling the periodic component of an LPC excitation vector using a pitch prediction filter (Singhal and Atal [1984]). Using a pitch prediction filter, the LPC excitation signal becomes

$$u(n) = bu(n-D) + \sum_{m=1}^{M} c_{k_m}(n) \tag{12.303}$$

where the signal $c_k(n)$ is an impulse located at $n = k$, and b is the pitch prediction filter gain. Singhal and Atal proposed choosing D before the locations of any impulses are known, by minimizing the following perceptually weighted error:

$$|E_D|^2 = |\tilde{S} - bX_D H|^2, \quad X_D = [u(-D), \ldots, u((L-1)-D)] \tag{12.304}$$

The G.723.1 multi-pulse LPC coder and the GSM full-rate RPE-LTP coder both use a closed-loop pitch predictor, as do all standardized variations of the CELP coder (see below). Typically, the pitch delay and gain are optimized first, and then the gains of any additional excitation vectors (e.g. impulses in an MPLPC algorithm) are selected to minimize the remaining error.

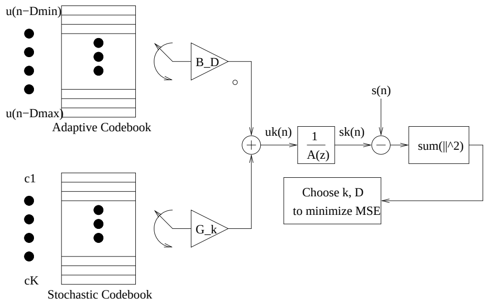


Figure 12.50: The code-excited LPC algorithm (CELP) constructs an LPC excitation signal by optimally choosing input vectors from two codebooks: an "adaptive" codebook, which represents the pitch periodicity, and a "stochastic" codebook, which represents the unpredictable innovations in each speech frame.

Code-excited LPC (CELP)

LPC analysis finds a filter $1/A(z)$ whose excitation is uncorrelated for correlation distances smaller than the order of the filter. Pitch prediction, especially closed-loop pitch prediction, removes much of the remaining inter-sample correlation. The spectrum of the pitch prediction residual looks like the spectrum of uncorrelated Gaussian noise, but replacing the residual with real noise (noise which is independent of the original signal) yields poor speech quality. Apparently, some of the temporal details of the pitch prediction residual are perceptually important. Schroeder and Atal proposed modeling the pitch prediction residual using a stochastic excitation vector $c_k(n)$ chosen from a list of stochastic excitation vectors, $k = 1, \ldots, K$, known to both the transmitter and receiver (Schroeder and Atal [1985]):

$$u(n) = bu(n - D) + gc_k(n) \tag{12.305}$$

The list of stochastic excitation vectors is called a stochastic codebook, and the index of the stochastic codevector is chosen in order to minimize the perceptually weighted error metric $|E_k|^2$. Rose and Barnwell discussed the similarity between the search for an optimum stochastic codevector index k and the search for an optimum predictor delay D (Rose and Barnwell III [1986]), and Kleijn et al. coined the term "adaptive codebook" to refer to the list of delayed excitation signals $u(n - D)$ which the coder considers during closed-loop pitch delay optimization (Figure 12.50).

The CELP algorithm was originally not considered efficient enough to be used in real-time speech coding, but a number of computational simplifications were proposed which resulted in real-time CELP-like algorithms. Trancoso and Atal proposed efficient search methods based on the truncated impulse response of the filter $W(z)/A(z)$ (Trancoso and Atal [1986], Atal [1986]). Davidson and Lin separately proposed center-clipping the stochastic codevectors, so that most of the samples in each codevector are zero (Davidson and Gersho [1986], Lin [1986]). Lin also proposed structuring the stochastic codebook so that each codevector is a slightly-shifted version of the previous codevector; such a codebook is called an overlapped codebook (Lin [1986]). Overlapped stochastic codebooks are rarely used in practice today, but overlapped-codebook search methods are often used to reduce the computational complexity of an adaptive codebook search. In the search of an overlapped codebook, the correlation R_X and autocorrelation Σ may be recursively computed, thus greatly reducing the complexity of the codebook search (Kleijn *et al.* [1988]).

Most CELP coders optimize the adaptive codebook index and gain first, and then choose a stochastic codevector and gain in order to minimize the remaining perceptually weighted error. If all of the possible pitch periods are longer than one sub-frame, then the entire content of the adaptive codebook is known before the beginning of the codebook search, and the efficient overlapped codebook search methods proposed by Lin may be applied (Lin [1986]). In practice, the pitch period of a female speaker is often shorter than one sub-frame. In order to guarantee that the entire adaptive codebook is known before beginning a codebook search, two methods are commonly used. First, the adaptive codebook search may simply be constrained to only consider pitch periods longer than L samples. In this case, the adaptive codebook will lock on to values of D which are an integer multiple of the actual pitch period; if the same integer multiple is not chosen for each subframe, the reconstructed speech quality is usually good. Second, adaptive codevectors with delays of $D < L$ may be constructed by simply repeating the most recent D samples as necessary to fill the subframe.

SELP, VSELP, ACELP, and LD-CELP

Rose and Barnwell demonstrated that reasonable speech quality is achieved if the LPC excitation vector is computed completely recursively, using two closed-loop pitch predictors in series, with no additional information (Rose and Barnwell III [1986]). In their "self-excited LPC" algorithm (SELP), the LPC excitation is initialized during the first sub-frame using a vector of samples known at both the transmitter and receiver. For all frames after the first, the excitation is the sum of an arbitrary number of adaptive codevectors:

$$u(n) = \sum_{m=1}^{M} b_m u(n - D_m) \tag{12.306}$$

Kleijn et al. developed efficient recursive algorithms for searching the adaptive codebook in SELP coder and other LPC-AS coders (Kleijn *et al.* [1988]).

Just as there may be more than one adaptive codebook, it is also possible to use more than one stochastic codebook. The vector-sum excited LPC algorithm (VSELP) models the LPC excitation vector as the sum of one adaptive and two stochastic codevectors (Gerson and Jasiuk [1991]).

$$u(n) = bu(n - D) + \sum_{m=1}^{2} g_m c_{k_m}(n) \tag{12.307}$$

The two stochastic codebooks are each relatively small (typically 32 vectors), so that each of the codebooks may be searched efficiently. The adaptive codevector and the two stochastic codevectors are chosen sequentially. After selection of the adaptive codevector, the stochastic codebooks are transformed using a modified Gram-Schmidt orthogonalization, so that the perceptually weighted speech vectors generated during the first stochastic codebook search are all orthogonal to the perceptually weighted adaptive codevector. Because of this orthogonalization, the stochastic codebook search results in the choice of a stochastic codevector which is jointly optimal with the adaptive codevector, rather than merely sequentially optimal. VSELP is the basis of the Telecommunications Industry Associations digital cellular standard IS-54.

The algebraic CELP (ACELP) algorithm creates an LPC excitation by choosing just one vector from an adaptive codebook and one vector from a fixed codebook. In the ACELP algorithm, however, the fixed codebook is composed of binary-valued or trinary-valued algebraic codes, rather than the usual samples of a Gaussian noise process (Adoul *et al.* [1987]). Because of the simplicity of the codevectors, it is possible to search a very large fixed codebook very quickly using methods which are a hybrid of standard CELP and MPLPC search algorithms. ACELP is the basis of the ITU standard G.729 coder at 8 kbps. ACELP codebooks may be somewhat larger than the codebooks in a standard CELP coder; the codebook in G.729, for example, contains 8096 codevectors per subframe.

Most LPC-AS coders operate at very low bit rates, but require relatively large buffering delays. The low delay CELP coder (LD-CELP) operates at 16 kbps (ITU-T [1992], Chen *et al.* [1992]) and is designed to obtain the best possible speech quality, with the constraint that the total algorithmic delay of a tandem coder and decoder must be no more than two milliseconds. LPC analysis and codevector search are computed once per two milliseconds (16 samples). Transmission of LPC coefficients once per two milliseconds would require too many bits, so LPC coefficients are computed in a recursive backward-adaptive fashion. Before coding or decoding each frame, samples of $\hat{s}(n)$ from the previous frame are windowed, and used to update a recursive estimate of the autocorrelation function. The resulting autocorrelation coefficients are similar to those that would be obtained using a relatively long asymmetric analysis window. LPC coefficients are then computed from the autocorrelation function using the Levinson-Durbin algorithm.

12.8.2 The LPC-10e Vocoder

The 2.4 kbps LPC-10e vocoder (Figure 12.51) is one of the earliest and one of the longest-lasting standards for low-bit-rate digital speech coding (DDVPC [1984], Campbell and Tremain [1986]). This standard was originally proposed in the 1970s, and was not officially replaced until the selection of the MELP 2.4 kbps coding standard in 1996 (Kohler [1997]). Speech coded using LPC-10e sounds metallic and

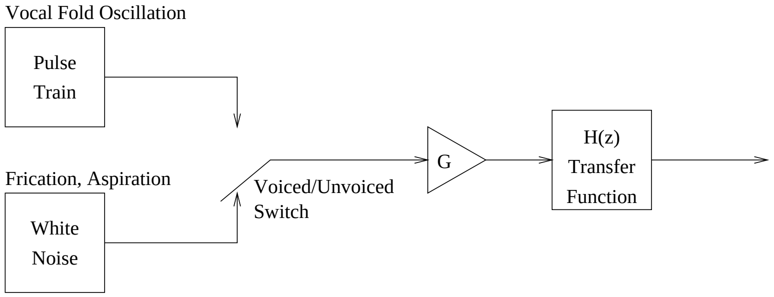


Figure 12.51: A simplified model of speech production, whose parameters can be transmitted efficiently across a digital channel.

synthetic, but it is intelligible.

In the LPC-10e algorithm, speech is first windowed using a Hamming window of length 22.5ms. The gain (G) and coefficients (a_i) of a linear prediction filter are calculated for the entire frame using the Levinson-Durbin recursion. Once G and a_i have been computed, the LPC residual signal $d(n)$ is computed:

$$d(n) = \frac{1}{G}(s(n) - \sum_{i=1}^{p} a_i s(n-i)) \tag{12.308}$$

The residual signal $d(n)$ is modeled using either a periodic train of impulses (if the speech frame is voiced) or an uncorrelated Gaussian random noise signal (if the frame is unvoiced). The voiced/unvoiced decision is based on the average magnitude difference function (AMDF),

$$\Phi_d(m) = \frac{1}{N-|m|} \sum_{n=|m|}^{N-1} |d(n) - d(n-|m|)| \tag{12.309}$$

The frame is labeled as voiced if there is a trough in $\Phi_d(m)$ which is large enough to be caused by voiced excitation. Only values of m between 20 and 160 are examined, corresponding to pitch frequencies between 50Hz and 400Hz. If the minimum value of $\Phi_d(m)$ in this range is less than a threshold, the frame is declared voiced, and otherwise it is declared unvoiced (Campbell and Tremain [1986]).

If the frame is voiced, then the LPC residual is represented using an impulse train of period T_0, where

$$T_0 = \arg \min_{m=20}^{160} \Phi_d(m) \tag{12.310}$$

If the frame is unvoiced, a pitch period of $T_0 = 0$ is transmitted, indicating that an uncorrelated Gaussian random noise signal should be used as the excitation of the LPC synthesis filter.

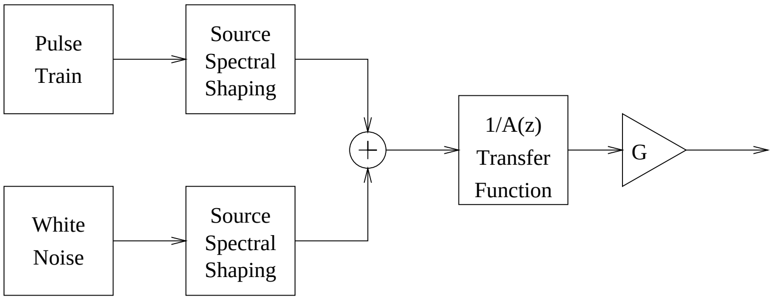


Figure 12.52: The MELP speech synthesis model

12.8.3 Mixed Excitation Linear Prediction (MELP)

The Mixed Excitation Linear Prediction (MELP) coder (McCree *et al.* [1996]) was selected in 1996 by the United States Department of Defense Voice Processing Consortium (DDVPC) to be the U.S. Federal Standard at 2.4 kbps replacing LPC-10e. The MELP coder is based on the LPC model with additional features that include mixed excitation, aperiodic pulses, adaptive spectral enhancement, pulse dispersion filtering, and Fourier magnitude modeling (McCree and Barnwell III [1995]). The synthesis model for the MELP coder is illustrated in Figure 12.52. LP coefficients are converted to LSFs and a Multi-Stage Vector Quantizer (MSVQ) is used to quantize the LSF vectors. For voiced segments a total of 54 bits that represent: LSF parameters (25), Fourier magnitudes of the prediction residual signal (8), gain (8), pitch (7), bandpass voicing (4), aperiodic flag (1), and a sync bit are sent. The Fourier magnitudes are coded with an 8-bit VQ and the associated codebook is searched with a perceptually-weighted Euclidean distance. For unvoiced segments, the Fourier magnitudes, bandpass voicing, and the aperiodic flag bit are not sent. Instead, 13 bits that implement Forward Error Correction (FEC) are sent. The performance of MELP at 2.4 kbps is similar to or better than that of the Federal Standard at 4.8 kbps (FS 1016) (Supplee *et al.* [1997]). Versions of MELP coders operating at 1.7 kbps (McCree and Martin [1998]) and 4.0 kbps (Stachurski *et al.* [1999]) have recently been reported.

12.8.4 Multi-Band Excitation (MBE)

In Multi-Band Excitation (MBE) coding the voiced/unvoiced decision is not a binary one; instead, a series of voicing decisions are made for independent harmonic intervals (Griffin and Lim [1988]). Since voicing decisions can be made in different frequency bands individually, synthesized speech may be partially voiced and partially unvoiced. An improved version of the MBE was introduced in (Hardwick and Lim [1988], Brandstein *et al.* [1990]) and referred to as the IMBE coder. The IMBE at 2.4 kbps produces better sound quality than the LPC-10e. The IMBE was adopted as the Inmarsat-M coding standard for satellite voice communication at a total rate of 6.4 kbps, including 4.15 kbps of source coding and 2.25 kbps of channel coding (Wong [1991]). The Advanced MBE (AMBE) coder was adopted as the Inmarsat Mini-M standard at a 4.8 kbps total data rate, including 3.6 kbps of speech and 1.2 kbps of channel coding (Dimolitsas [1995], Goldberg and Riek [2000]). In (Das and Gersho [1999]) an enhanced multiband excitation (EMBE) coder was presented. The distinguishing features of the EMBE coder include signal-adaptive multimode spectral modeling and parameter quantization, a two-band signal-adaptive frequency-domain voicing decision, a novel VQ scheme for the efficient encoding of the variable-dimension spectral magnitude vectors at low-rates, and multi-class selective protection of spectral parameters from channel errors. The 4 kbps EMBE coder accounts for both source (2.9 kbps) and channel (1.1 kbps) coding and was designed for satellite-based communication systems.

12.8.5 Prototype Waveform Interpolative (PWI) Coding

A different kind of coding technique which has properties of both waveform and LPC-based coders was proposed in (Kleijn [1991], Kleijn and Granzow [1991]) and is called Prototype Waveform Interpolation (PWI). PWI uses both interpolation in the frequency domain and forward-backward prediction in the time domain. The technique is based on the assumption that, for voiced speech, a perceptually accurate speech signal can be reconstructed from a description of the waveform of a single, representative pitch cycle per interval of 20-30 ms. The assumption exploits the fact that voiced speech can be interpreted as a concentration of slowly evolving pitch-cycle waveforms. The prototype waveform is described by a set of linear-prediction (LP) filter coefficients describing the formant structure and a prototype excitation waveform, quantized with analysis-by-synthesis procedures. The speech signal is reconstructed by filtering an excitation signal consisting of the concatenation of (infinitesimal) sections of the instantaneous excitation waveforms. By coding the voiced and unvoiced components separately, a 2.4 kb/s version of the coder performed similarly to the 4.8 kb/s FS1016 standard (Kleijn and Haagen [1995]).

Recent work has aimed at reducing the computational complexity of the coder for rates between 1.2 and 2.4 kbps by including a time-varying waveform sampling rate and a cubic B-spline waveform representation (Kleijn *et al.* [1996], Shoham [1997]).

12.8.6 Voice-Excited Formant Vocoders

The voice-excitation technique described in Section 12.8.1 has also been applied to a parallel-connected formant vocoder (Flanagan [1960b]). The circuit arrangement is shown in Fig. 12.53. In this implementation, a baseband of about 400 Hz (300 to 700 Hz) is transmitted in unprocessed form. Three formant vocoder channels cover the frequency range 800 to 3200, and the amplitude and frequency of three spectral maxima in this range are transmitted. Formant extraction is accomplished according

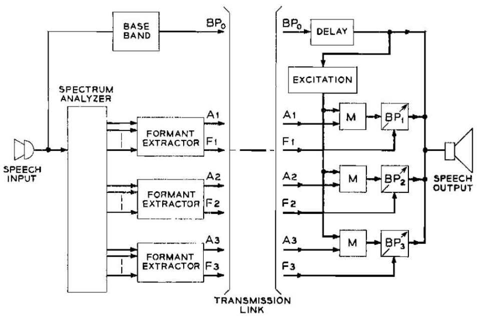


Figure 12.53: Voice-excited formant vocoder. (After (Flanagan [1960b]))

to the maximum-picking technique described in Section 5.5, Chapter 5. All control signals are low-passed to 17 Hz. The total bandwidth occupancy is therefore slightly more than 500 Hz.

At the synthesizer the baseband is spectrally broadened. It is peakclipped, differentiated, half-wave rectified and used to trigger a one-shot multivibrator. The pulse output of the multivibrator provides the excitation source for the formant channels. Unvoiced sounds create shot noise from the multivibrator. Voiced sounds produce periodic pulse trains which sometimes may have more than one pulse per fundamental period. The technique generally provides an improvement in the quality and naturalness of the formant vocoder transmission. However, because the baseband is such a large percentage of the total bandwidth, it is almost as economical to use conventional vocoder channels above the baseband.

A related voice-excited technique uses the spectral shape of the first formant region to shape the second and third formants (Greefkes and de Jager [1968]). A baseband about 300 to 800 Hz is separated and transmitted in unprocessed form. In two other (formant) bands, 800 to 2000 Hz and 2000 to 3200 Hz, zero-crossing counters and rectifier-integrator circuits determine signals representing the amplitudes and frequencies of the formants. These four signals are lowpassed to 40 Hz each, and are sent with the baseband to the receiver.

The synthesizer reconstructs a spectrum in which the baseband (essentially the first formant) is produced in its original position. A second formant is synthesized in a separate parallel branch by heterodyning the baseband to the measured second formant frequency position. A third is generated in a similar fashion. The output speech is obtained by adding the three parallel branches in accordance with the measured amplitudes. The spectral components of the heterodyned bands generally become inharmonic, and the pitch frequency is preserved in them only to the extent of line spacing. Perceptually, the degradation of pitch information is less than might be expected, since the baseband is retained in original form with its correct line structure, and it is an effective masker.

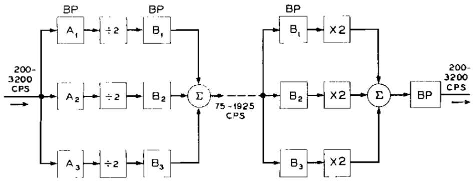


Figure 12.54: Block diagram of the Vobanc frequency division-multiplication system. (After (Bogert [1956]))

12.8.7 Frequency-Dividing Vocoders

A class of vocoder devices aims to avoid the difficult problems of pitch tracking and voiced-unvoiced switching that conventional vocoders use. The intent is to settle for modest savings in bandwidth in order to realize a simpler implementation and a synthetic signal of higher quality. The voice-excited vocoder described in Section 12.8.1 represents one such effort. The band saving which accrues is due primarily to the ear's criteria for representing the short-time spectrum of the signal.

Frequency division is a well-known process for reducing the bandwidth of signals whose spectral widths are determined primarily by large-index frequency modulation. While speech is not such a signal, sub-bands of it (for example, formant bands or individual voice harmonics) have similarities to large-index frequency modulation. Frequency division by factors of two or three are possible before intelligibility deteriorates substantially.

Frequency division generally implies possibilities for frequency multiplication. Similarly, spectral division-multiplication processes suggest possibilities for compression and expansion of the signal's time scale. Reproduction of a divided signal at a rate proportionately faster restores the frequency components to their original values, and compresses the time scale by a factor equal to the frequency divisor.

Vobanc

Various methods–including electrical, mechanical, optical and digital–have been used to accomplish division and multiplication. All cannot be described in detail. Several, however, serve to illustrate the variety of designs and applications.

One frequency-division method for bandwidth conservation is the Vobanc (Bogert [1956]). Although constructed practically using heterodyne techniques, the principle involved is shown in Fig. 12.54. The speech band 200 to 3200 Hz is separated into three contiguous band-pass channels, A_l, A_2, A_3. Each channel is about 1000 Hz wide and normally covers the range of a speech formant. Using a regenerative modulator, the signal in each band is divided by two and limited to one-half the original frequency range by BP filters B_1, B_2, B_3. The added outputs of the filters yield a transmission signal which is confined to about one-half the original bandwidth.

At the receiver, the signal is again filtered into the three bands, B_1, B_2, and

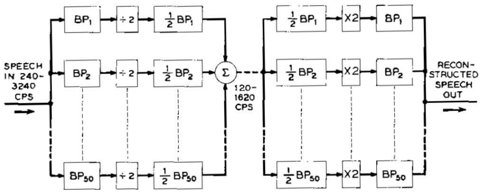


Figure 12.55: Block diagram of "harmonic compressor." (After (Schroeder *et al.* [1962]))

B_3. The bands are restored by frequency doubling and are combined to provide the output signal. In consonant articulation tests with 48 listeners and 10 talkers, the Vobanc consonant articulation was approximately 80 per cent. In the same test, an otherwise unprocessed channel, band-limited to 200 to 1700 Hz, scored a consonant intelligibility of about 66 per cent.

Other systems similar in band-division to Vobanc have been investigated (Seki [1958], Marcou and Daguet [1956a,b]). One proposal, called Codimex, considers potential division by factors as high as eight (Daguet [1962]), although practical division appears limited to factors of two or three.

Harmonic Compressor

Another complete division-multiplication transmission system, designed with a sufficient number of filters to operate on individual voice harmonics, has been investigated by digital simulation (Schroeder *et al.* [1962]). This method, called the "harmonic compressor," uses 50 contiguous bandpass filters, each 60 Hz wide, covering the range 240 to 3240 Hz. The circuit is shown in Fig. 12.55. It is designed to achieve a bandwidth reduction of two-to-one. On the transmitter side, the signals from the bandpass filters are divided by two and combined for transmission over one-half the original bandwidth. At the receiver the components are again separated by filtering and restored by multiplication by two. All filters and operations arc simulated in a large digital computer. From informal listening tests, the quality and intelligibility of the transmitted speech are judged to fall between that of a voice-excited vocoder with a 700 bps baseband and an unprocessed signal of the same bandwidth. A time speed up by a factor of two can also be applied to the transmitted signal to restore it to the original frequency range.

A related investigation in which attention is focused upon the individual harmonic components of the signal has considered optical methods, mechanical string-filter methods, and ultrasonic storage devices for frequency division-multiplication (Vilbig [1950, 1952], Vilbig and Haase [1956a,b]). A part of this same effort produced an electrical "speech stretcher" (Gould [1951]). The idea is to expand the time scale of speech by the arrangement shown in Fig. 12.56. The speech signal is filtered by 32 contiguous BP-filters covering the range 75 to about 7000 Hz. The filter band-

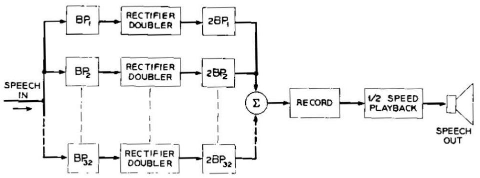


Figure 12.56: A "speech stretcher" using frequency multiplication to permit expansion of the time scale. (After (Gould [1951]))

widths are approximately 100 Hz wide up to 1000 Hz, and increase logarithmically to 7000 Hz. Full-wave rectification doubles the frequency components of each band. Band-pass filtering at twice the original bandwidth eliminates much of the harmonic distortion. Recording the combined signal and playing back at one-half speed restores the components to their original frequency positions. The time scale of the signal, however, is expanded by two.

12.9 Rate-Distortion Tradeoffs for Speech Coding

12.9.1 Multiplexing and Digitalization

The problems in multiplexing the voice-excited vocoder are essentially similar to those discussed in Section 12.4.1 for the channel vocoder. The main difference is the unprocessed baseband. For economical transmission in a frequency multiplex system, it should be left unaltered or produced as a single sideband modulation. Transmission of the spectrum-defining channel signals can be the same in both cases.

One design of a voice-excited vocoder uses 500 Hz of unprocessed baseband and 13 spectrum channels above the baseband (Howell *et al.* [1961]). The baseband is transmitted by single sideband modulation, and the channel signals are transmitted by vestigial sideband. Another analog implementation uses an unprocessed baseband of 250 to 925 Hz and 10 vocoder channels covering the range to approximately 3000 Hz (Golden [1963]). The channel signals are double-sideband amplitude modulated onto 10 carriers spaced by 60 Hz in the range 925 to 1630 Hz. A bandwidth compression of approximately two-to-one is thereby realized.

Digital simulation and current computer techniques have also been used to design and study a complete voice-excited vocoder (Golden [1963]). To realize the digital simulation, the sampled-data equivalents of all filters and all circuits of an analog 10-channel voice-excited vocoder were derived (see, for example, Section 11.5, Chapter 11). Transformation of the continuous system into the sampled-function domain permits its simulation in terms of discrete operations which can be programmed in a digital computer. In the present instance, the entire vocoder was represented inside the computer, and sampled-quantized input speech signals were processed by the program.

The immense advantage that this technique offers for research and design of signal-processing systems cannot be overemphasized. The entire transmission system can be simulated and evaluated before COnstructing a single piece of hardware. The usual price paid is non-real time operation of the system. The time factor for the present simulation was 172 to 1, or 172 sec of computation to process one second of speech. However, as digital techniques develop and as computers become even faster, this time factor will shrink proportionately.

Another vocoder development has resulted in a time-multiplexed, fully digitalized voice-excited vocoder (L. A. Yaggi [1962]). The device is designed to operate at a data rate of 9600 bits/sec and to use PCM encoding. The system operates with a baseband whose upper cutoff is, optionally, either 800 Hz or 950 Hz. For the former, 12 vocoder channels cover the range to 4000 Hz; for the latter, 11 channels are used. The baseband signal is sampled at the Nyquist rate and quantized to 5 bits. The spectrum channels are sampled at 50 sec-1 (64 sec-1 for the 950 Hz baseband); the lower three are quantized to 3 bits, and the higher ones to 2 bits. Amplitude normalization of the spectrum signals is also used. Comparable choices have been made in alternative digital implementations (Tierney [1965]).

Other coding techniques which, like the voice-excited vocoder, avoid the pitch tracking problem include the phase vocoder, the vobanc and the analytic rooter. These methods are discussed in later sections.

12.9.2 Multiplexing of Formant Vocoders

One real-time formant vocoder that has been given extensive tests is the parallel configuration shown in Fig. 12.57 (Stead and Jones [1961]). Besides being tested in the unmultiplexed "back-to-back" connection, this system has also been examined in a fully digitalized version using time-division PCM techniques (Weston [1962]). The components of the system have several similarities with devices discussed previously. In one version, the synthesizer is based upon an earlier development (Lawrence [1953]). The formant-frequency extractor is based upon the peak-picking technique described in Section 5.5, Chapter 5. The overall implementation and circuit design are unusually refined, and considerable effort is made to insure adequate dynamic range for the extraction of frequency and amplitude data. In the analog form, low-pass filters confine the eight control parameters to approximately 20 Hz each, resulting in a total bandwidth occupancy of about 160 Hz. Typical intelligibility scores for phonetically-balanced words and for relatively naive listeners are reported to average approximately 70%.

As mentioned earlier, the advantages of digital transmission are several. Not least is the ability to regenerate the signal repeatedly–essentially free of accumulating distortion. Problems in switching, time sharing and security are also amenable to straightforward solutions with the signal in digital form. One difficulty, however, is that transmission of the digital signal requires more bandwidth than does the analog

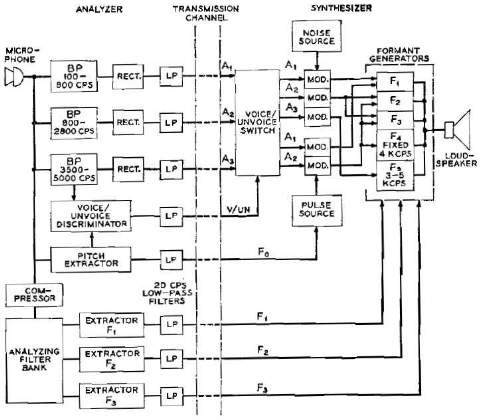


Figure 12.57: A complete formant-vocoder system utilizing analog and digital transmission techniques. (After (Stead and Jones [1961], Weston [1962]))

Table 12.5: Quantization of formant-vocoder signals. (After STEAD and WESTON)

Parameter	Number of levels	Bits
Fl:	16	4
F2:	16	4
F3:	8	3
AI:	8	3
A2:	8	3
A3:	8	3
V/UN:	2	1
F0:	64[14]	6
TOTAL		27

form. For example, a 3000 Hz speech band sampled at the Nyquist rate (6000 sec^{-1}) and quantized to 6 or 7 bits may require–without further coding–a bandwidth on the order of 50000 Hz. If, through appropriate coding, the data rate could be brought down to the order of 1000 bits/sec, the digital signal could easily be transmitted over ordinary 3000 Hz voice channels. The formant vocoder holds promise for providing such a coding.

In the formant vocoder of Fig. 12.57, the control parameters were band-limited to 20 Hz. For digitalizing the control signals, however, a sampling rate of 32 sec^{-1} was found to be a safe working minimum. This rate suggests that the control parameters have little significant energy above about 16 Hz. The amplitude quantization found acceptable for digitalizing the control data of this system is shown in Table 12.5.

In evaluating the digital transmission, 16 levels were thought too generous for the first formant frequency, but 8 levels were too coarse. For the three amplitude

Table 12.6: Estimated precision necessary in quantizing formant-vocoder parameters. The estimates are based upon just-discriminable changes in the parameters of synthetic vowels; amplitude parameters are considered to be logarithmic measures. (After (Flanagan [1957c]))

Parameter	Number of levels	Bits
F1:	14	3.8
F2:	14	3.8
F3:	9	3.2
A1:	3	1.6
A2:	3	1.6
A3:	2	1.0
F0:	40	5.3
TOTAL:		20.3

parameters, the 8 levels each were also thought too generous and that additional saving could be effected by coding the functions on a log-amplitude scale[15]

It is relevant to compare the practical figures of Table 12.5 with earlier estimates of the precision necessary in quantizing similar parameters (Flanagan [1957c]). The earlier estimates were based upon the just-perceptible changes which listeners could detect in the formant parameters of synthetic vowels (see Section 9.2, Chapter 9). The quantizing accuracy estimated to be necessary is given in Table 12.6.

In view of the limitations of the perceptual data on which the estimates are based, the correspondence with the numbers in Table 12.5 is surprisingly close. It suggests that psychoacoustic measures of the type discussed in Chapter 9 might be judiciously applied with some confidence to estimate system performance.

After sampling and quantizing, the data of Fig. 12.57 are PCM encoded for transmission. At a sampling rate of 32 sec^{-1}, the control data, exclusive of pitch, are specified with 672 bits/sec. A 6-bit pitch parameter produces a total rate of 864 bits/sec, a rate which could be transmitted by conventional 3000 Hz channels. Although detailed testing has not been carried out, the digitally transmitted signal is reported to differ only slightly in intelligibility and quality from the analog connection. One interesting observation about the system is that the spectrum of the quantizing noise associated with digitalizing the control signals does not lie in the audio band. Rather, the noise corresponds to a sort of quasi-random uncertainty in the synthesis process. Its subjective effects have not been fully explored.

A preliminary study has considered the effects of digital errors in the PCM-encoded parameters of a formant vocoder (Campanella *et al.* [1962]). The system

---

[15] These observations have been confirmed, extended and quantified in greater depth by computer-perceptual experiments on the bandlimitation and quantization of formant data in synthesis (Rosenberg [1971a]).

on which the tests were performed is similar to that shown in Fig. 12.57, except the voiced-unvoiced decision is transmitted by the pitch signal. The total bandwidth occupancy of the control signals is 140 Hz. The formant parameters are quantized to 3 bits. Pitch is quantized to 5 bits. A 43.5 $\sec^{-1}$ scan rate in the time multiplexing produces a data rate of 1000 bits/sec. Under error free conditions, articulation scores of about 80% on PB words are claimed for this bit rate. A digital error rate of 3% degrades the articulation score by 15%. This impairment is found to be equivalent to reducing the signal-to-noise ratio of the analog parameters to 9.5 db.

12.9.3 Time-Assignment Transmission of Speech

In two-way conversation, one party is normally silent and listening on the average of one-half the time. In addition, natural speech has many pauses and silent intervals. A given talker, therefore, transmits a signal only on the order of 35 to 40 per cent of the total time. In longdistance communication, where amplification of the signal is necessary, the two-way communication channels are normally four-wire circuits-or two unilateral transmission paths. Each party has a transmit circuit and a receive circuit. Because of the relative inactivity of each talker, a single one-way channel is not used on the order of 60 to 65 per cent of the time. When a large group of such connections are accessible from single transmit and receive locations, the statistical properties of the conversation ensemble make a significant amount of time and bandwidth available for signal transmission. A method for practicably utilizing this capacity is called Time Assignment Speech Interpolation, or “TASI” (O’Neil [1959], Bullington and Fraser [1959]).

The TASI system has available a group of unilateral transmit and receive circuits–typically the line-pairs in an undersea cable. The system is to serve a greater number of talkers than the number of unilateral circuits. The incoming transmit circuit of each talker is equipped with a fast-acting speech detector, or voice switch. When the detector indicates the presence of speech on its line, an automatic electronic switch connects the line to an available transmit path of the TASI group. Incoming signals for transmission are assigned transmit circuits until all have been filled. When the number of signals to be transmitted exceeds the number of transmit paths, the TASI switch searches the connections to find one that has fallen silent, disconnects it, and assigns that path to a channel which has a signal to transmit.

During pauses and silent intervals, a given talker loses his priority on the transmit link. He is reassigned a channel–often a different one–when he again becomes active. The TASI switch must consequently keep track of who is talking to whom, and it must identify the recipient of each signal presented for transmission. This message “addressing” information can be transmitted in the form of a very short identification signal, either before each talk spurt or over an auxiliary channel that serves the entire system.

A limit obviously exists to the number of incoming signals that can be transmitted

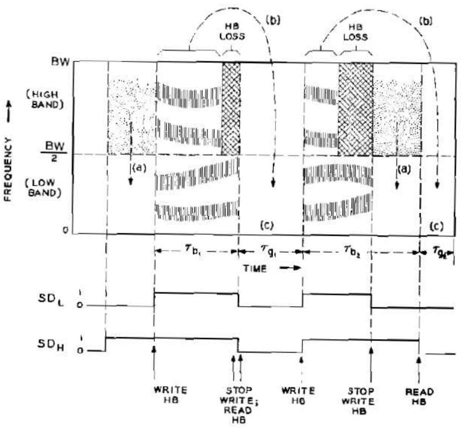


Figure 12.58: Schematic sound spectrogram illustrating the principle of the "one-man TASI." (After (Schroeder and Bird [1962]))

by a given group of transmit paths before some "freezeout" or loss of speech signals occurs. Among other things, this limit is a function of the size of the cable group, the circuit signal-to-noise ratio, and the sensitivity of the speech detectors. Several TASI systems have been put into practical operation on undersea cables. On a 36-channel cable, for example, the effective transmission bandwidth is on the order of two to three times that of the physical circuit.

As mentioned at the beginning of the section, natural pauses of phonemic, syllabic or longer durations occur in a single "one-way" speech signal. These pauses or gaps suggest that the TASI principle might be applied to a single speech channel to realize a band-saving. An experimental circuit, called a "one-man TASI," has considered this point (Schroeder and Bird [1962]). The system has been tested by simulation in a digital computer. Its principle of operation is illustrated by the schematic sound spectrogram in Fig. 12.58.

As shown in Fig. 12.58, suppose that a speech band of BW Hz is to be transmitted, but that a channel width of only BW/2 is available. The natural pauses and gaps in one BW/2 of the signal might be used to transmit information about the other BW/2 band of the signal. If the BW/2 bands are called high band (HB) and low band (LB), four signal possibilities exist. The processing strategies employed in the four situations are illustrated by corresponding letters on Fig. 12.58, and are:

a) When only HB signal (and no LB signal) is present, the HB is detected, heterodyned down to the LB range, and transmitted immediately over the BW/2 channel.

b) When HB and LB are detected simulataneously, the LB is transmitted immediately, while the HB is heterodyned down and read into a storage for transmission later. (See τ_b intervals in Fig. 12.58).

c) When neither HB nor LB signal is detected, a gap exists. (See τ_g intervals in Fig. 12.58.) During this interval, as much of the previouslystored HB is transmitted as there is time for. Generally some trailing edge of the HB will be lost. One set of speech-burst statistics gives average burst durations of about 130 msec followed by average silent intervals of 100 msec (BOLT and MACDONALD). On the basis of these statistics, about 3/13 of the HB signal would be expected to be lost. None of the LB signal is lost.

d) When LB only is present, it is transmitted immediately in the conventional manner.

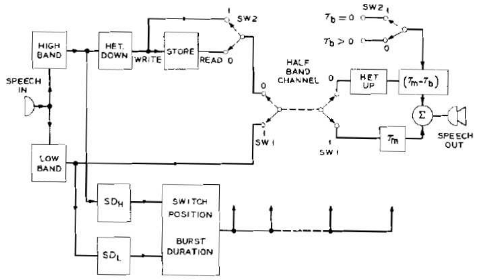
Two speech detectors, one for each band, are required. In the present study, they were full-wave rectifiers with 15-msec smoothing time constants. Their outputs operated threshold devices with prescribed hysteresis characteristics. The binary output signals from the detectors, shown as SDL and SDH in Fig. 12.58, must also be transmitted over a narrow-band channel so that the speech may be properly reassembled at the receiver. Because of the storage on the transmitter side, a fixed transmission delay is incurred before the reassembled signal is available.

The reassembly operations are evident in the block diagram of the complete system in Fig. 12.59. Two delay elements are used at the receiver. One is a fixed, maximum transmission delay τ_m in the LB channel. Its value is equal to or greater than the duration of the longest speech burst to be stored. The other is a variable delay whose value is the difference between τ_m and the last speech-burst duration τ_b. The various switch conditions–corresponding to the SDL and SDH signal outputs–are shown in the table.

In testing the system by simulation in a digital computer, the effective size of the HB store was taken as 500 msec. In the unlikely instance of a speech-burst duration longer than 500 msec, the high-band information was discarded, rather than reassembled in the wrong place. Typical operation of the system, as simulated in the computer, is shown by the spectrograms of Fig. 12.60. The utterance is “High altitude jets whiz past screaming.” In comparing what the complete system provides over and above a single BW/2 channel, one sees that a substantial amount of the high band is transmitted. All high frequencies from unvoiced bursts are present, and a large percentage of the voiced HB is preserved.

The price of the improvement is the complexity of the storage and switching and the 500-msec transmission delay.

Alternatively, the silent gaps in the speech signal may be used to interleave another signal, such as digital data read on demand from a buffer store. In one computer simulation of this technique (Hanauer and Schroeder [1966]), the speech envelope was used as a control to switch between speech and data. It was found possible to make available as much as 55% of the speech-signal time for interleaving the alternate information.



SD_L	SD_H	SW1	SW2	STORE
1	0	1	0	---
0	1	0	1	---
1	1	1	0	WRITE
0	0	0	0	READ

Figure 12.59: Block diagram of ”one-man TASI” system for 2:1 band-width reduction. (After (Schroeder and Bird [1962]))

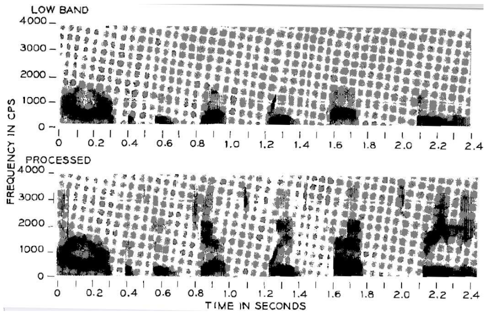


Figure 12.60: Sound spectrograms illustrating operation of the single channel speech interpolator

12.9.4 Multiplexing Channel Vocoders

Frequency-Space Multiplexing

The customary techniques for transmitting a multiplicity of simultaneous signals are frequency-space multiplexing and time-division multiplexing. In the former, the requisite amount of spectrum bandwidth is allocated to each signal. The individual signals are modulated onto separate carriers, which are transmitted simultaneously within the allocated channels and are demodulated at the receiver. In the latter, the several signals time-share a single transmission path of appropriate bandwidth.

Frequency multiplexing of vocoder signals is attractive from the standpoint of circuit simplicity and existing analog communication links. Certain relations can be observed to conserve spectrum space and, at the same time, provide accurate transmission. Since the vocoder signals normally contain a dc component, the modulation method must be chosen to preserve this response. Conventional double-sideband (I)S13) amplitude modulation would satisfy the response requirement, hut would not be economical of bandwidth. Conventional single-sidehand (SSB) modulation with suppressed carrier, although taking only half the bandwidth, would not reliably preserve the low-frequency components of the modulation. Vestigial sideband transmission might suffice. However, a two-phase (or quadrature) modulation method has been advanced as the best solution (Halsey and Swaffield [1948]).

A pair of channel signals DSB modulate separate carriers of the same frequency but differing in phase by $\pi/2$ radians. The two double-sideband signals then occupy the same frequency band. Provided the transmission path has attenuation and phase characteristics symmetrical about the carrier frequency, either signal-complex can be rejected at the receiver by demodulating (multiplying and integrating) with a synchronous quadrature carrier. Frequency and phase synchrony of the carriers at the transmitter and receiver are of course critical.

The quadrature method is generally not satisfactory for transmission of conventional voice signals. Practical stabilities are such that the crosstalk between circuits cannot be kept low enough. For vocoder signals, however, a crosstalk attenuation between spectrum channels of about 25 db seems adequate[16]. This figure is within the practical limits of the quadrature method. The signal-to-crosstalk ratio is the cotangent of the phase error between the modulating and demodulating carriers. Therefore, a crosstalk attenuation of 25 db, or more, requires a phase error of about 3.3 degrees, or less.

Time-Division Multiplexing

Time-division multiplexing involves the transmission of sample values of the channel signals taken in time sequence. According to the sampling theorem, the rate of

---

[16] The pitch channel is more sensitive to crosstalk. For it, an attenuation on the order of 40 db is desirable.

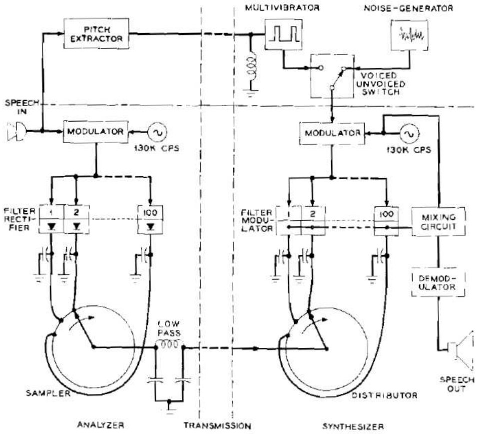


Figure 12.61: Channel vocoder utilizing time-multiplex transmission. (After (Vilbig and Haase [1956a]))

sampling must be at least twice the highest frequency contained in the channel signals. The vocoder signals are typically bandlimited to about 20 Hz, hence sampling rates on the order of 40Hz, or higher, are indicated. Practically, to provide adequate channel separation in the desampling (distributing) operation, a total transmission bandwidth about twice the sum of the input signals, that is, the same as for DSB frequency-multiplex, is required(Bennett [1941]). Even then, the crosstalk between channels may be only marginally acceptable. For example, in a 12channel system the signal-to-crosstalk ratio is only on the order of 20 db. Without further coding, therefore, this multiplexing method appears somewhat less attractive from the fidelity standpoint than the quadrature frequency-space multiplex. On the other hand, its simplicity, and the possibility for analog smoothing of the spectral shape, make it of interest.

One vocoder developed on the time-multiplex principle is called the Scan Vocoder (Vilbig and Haase [1956a,b]). It is illustrated in Fig. 12.61. One hundred spectrum channels, using high frequency (130 kc) magnetostriction rods as the filters, produce a short-time spectrum. The filter outputs are scanned at 30Hz and the time-multiplexed spectral envelope is smoothed by a 200Hz low-pass filter. The envelope signal is demultiplexed by a synchronously scanning distributor at the receiver. The pitch information is transmitted in a separate channel.

Digital Transmission of Vocoder Signals

Transmission of signals in the form of binary pulses has a number of advantages. One is the possibility for repeated, exact regeneration of the signal. Noise and distortion do not accumulate as they do in analog amplification. Quality of the digital signal can, within limits, be made independent of transmission distance. Another advantage is the freedom to "scramble" the message in complex ways for secure or private communication. The price paid for these important advantages is additional transmission bandwidth. Time-divison multiplexing, coupled with pulse code modulation (PCM) of the channel signals, is consequently an attractive means for vocoder transmission. The signal value in each sampled channel is represented by a sequence of binary pulses. The ordered and "framed" pulses are transmitted over a relatively broadband channel, synchronously distributed at the receiver, and reconverted from digital to analog form.

Although the digital signal requires comparatively greater bandwidth, the vocoded speech signal makes feasible full digital transmission over about the same bandwidth as normally used for nondigital conventional telephony. An important question is how many binary pulses are sufficient to represent each sample of the channel signals. The answer of course depends upon the quality of received signal that is acceptable. Current technology has used pulse rates from 1200 to 4800 bits/sec in particular applications (L. A. Yaggi and Mason [1963]). A typical design, for example, uses 18 spectrum channels which are sampled at 40Hz and which are normalized in amplitude. The number of binary digits used to specify the sampled values of channels 1 through 14 is three bits; for channels 15 through 18, two bits; for the over-all amplitude level, three bits, and for pitch and voiced-unvoiced indication, seven bits. Therefore, 60 bits are included in one scan or "frame," and 2400 bits/sec is the transmitted data rate. Numerous variations in the design for digital transmission can be found.

12.10 Network Issues

12.10.1 Voice over IP

Speech coding for the Voice over Internet Protocol (VoIP) application is becoming important with the increasing dependency on the internet. The first VoIP standard was published in 1998 as recommendation H.323 (ITU-T [1998b]) by the International Telecommunications Union (ITU-T). It is a protocol for multimedia communications over Local Area Networks using packet-switching, and the voice-only subset of it provides a platform for IP-based telephony. At high bit rates, H.323 recommends the coders G.711 (3.4 kHz at 48, 56, and 64 kbps) and G.722 (wideband speech and music at 7 kHz operating at 48, 56, and 64 kbps) while at the lower bit rates G.728 (3.4 kHz at 16 kbps), G.723 (5.3 and 6.5 kbps), and G.729 (8 kbps) are recommended (ITU-T [1998b]).

In 1999, a competing and simpler protocol named the Session Initiation Protocol (SIP) was developed by the Internet Engineering Task Force (IETF) Multiparty Multimedia Session Control working group and published as RFC 2543 (et al. [1999]). SIP is a signaling protocol for Internet conferencing and telephony, is independent of the packet layer, and runs over UDP or TCP although it supports more protocols and handles the associations between Internet end systems. For now, both systems will coexist but it is predicted that the H.323 and SIP architectures will evolve such that two systems will become more similar.

Speech transmission over the internet relies on sending 'packets' of the speech signal. Due to network congestion, packet loss can occur, resulting in audible artifacts. High-quality VoIP, hence, would benefit from variable-rate source and channel coding, packet loss concealment, and jitter buffer/delay management. These are challenging issues and research efforts continue to generate high-quality speech for VoIP applications (Hersent *et al.* [2000]).

12.10.2 Error Protection Coding

12.10.3 The Rate-Distortion Curve

12.10.4 Embedded and Multi-Mode Coding

When channel quality varies, it is often desirable to adjust the bit rate of a speech coder in order to match the channel capacity. Varying bit rates are achieved in one of two ways. In multi-mode speech coding, the transmitter and receiver must agree on a bit rate prior to transmission of the coded bits. In embedded source coding, on the other hand, the bitstream of the coder operating at low bit rates is embedded in the bitstream of the coder operating at higher rates. Each increment in bit rate provides marginal improvement in speech quality. Lower bit rate coding is obtained by puncturing bits from the higher rate coder and typically exhibits graceful degradation in quality with decreasing bit rates.

ITU Standard G.727 describes an embedded ADPCM coder, which may be run at rates of 40, 32, 24, or 16 kbps (5, 4, 3, or 2 bits/sample) (ITU-T [1990b]). Embedded ADPCM algorithms are a family of variable bit rate coding algorithms operating on a sample per sample basis (as opposed to, for example, a subband coder that operates on a frame-by-frame basis) that allows for bit dropping after encoding. The decision levels of the lower rate quantizers are subsets of those of the quantizers at higher rates. This allows for bit reduction at any point in the network without the need of coordination between the transmitter and the receiver.

The prediction in the encoder is computed using a more coarse quantization of $\hat{d}(n)$ than the quantization actually transmitted. For example, 5 bits/sample may be transmitted, but as few as 2 bits may be used to reconstruct $\hat{d}(n)$ in the prediction loop. Any bits not used in the prediction loop are marked as "optional" by the signaling channel mode flag. If network congestion disrupts traffic at a router between sender and receiver, the router is allowed to drop optional bits from the coded speech packets.

Embedded ADPCM algorithms produce code words that contain enhancement and core bits. The feed-forward (FF) path of the codec utilizes both enhancement bits and core bits, while the feed-back (FB) path uses core bits only. With this structure, enhancement bits can be discarded or dropped during network congestion.

An important example of a multi-mode coder is QCELP, the speech coder standard that was adopted by the TIA North American digital cellular standard based on Code Division Multiple Access (CDMA) technology (CDMA [1992]). The coder selects one of four data rates every 20 ms depending on the speech activity; for example, background noise is coded at a lower rate than speech. The four rates are approximately 1 kbps (eighth rate), 2 kbps (quarter rate), 4 kbps (half rate), and 8 kbps (full rate). QCELP is based on the CELP structure but integrates implementation of the different rates thus reducing the average bit rate. For example, at the higher rates, the LSP parameters are more finely quantized and the pitch and codebook parameters are updated more frequently (Gardner *et al.* [1993]). The coder provides good quality speech at average rates of 4 kbps.

Another example of a multi-mode coder is ITU standard G.723.1, which is an LPC-AS coder that can operate at 2 rates: 5.3 or 6.3 kbps (ITU-T [1996b]). At 6.3 kbps, the coder is a Multi-pulse LPC (MPLPC) coder while the 5.3 kbps coder is an Algebraic CELP (ACELP) coder. The frame size is 30 msec with an additional look ahead of 7.5 ms, resulting in a total algorithmic delay of 67.5 ms. The ACELP and MPLPC coders share the same LPC analysis algorithm and frame/sub-frame structure, so that most of the program code is used by both coders. As mentioned earlier, in ACELP, an algebraic transformation of the transmitted index produces the excitation signal for the synthesizer. In MPLPC, on the other hand, minimizing the perceptual-error weighting is achieved by choosing the amplitude and position of a number of pulses in the excitation signal. Voice Activity Detection (VAD) is used to reduce the bit rate during silent periods, and switching from one bitrate to another is done on a frame-by-frame basis.

Multi-mode coders have been proposed over a wide variety of bandwidths. Taniguchi et al. proposed a multi-mode ADPCM coder at bit rates between 10 kbps and 35 kbps (Taniguchi [1988]). Johnson and Taniguchi proposed a multi-mode CELP algorithm at data rates of 4.0-5.3 kbps in which additional stochastic codevectors are added to the LPC excitation vector when channel conditions are sufficiently good to allow high-quality transmission (Johnson and Taniguchi [1991]). The European Telecommunications Standards Institute (ETSI) has recently proposed a standard for Adaptive Multi-Rate coding at rates between 4.75 and 12.2 kbps.

12.10.5 Joint Source-Channel Coding

In speech communication systems, a major challenge is to design a system that provides the best possible speech quality throughout a wide range of channel conditions. One solution consists of allowing the transceivers to monitor the state of the communication channel and to dynamically allocate the bitstream between source and channel coding accordingly. For low SNR channels, the source coder operates at low bit rates, thus allowing powerful forward error control. For high SNR channels, the source coder uses its highest rate resulting in high speech quality, but with little error control. An adaptive algorithm selects a source coder and channel coder based on estimates of channel quality in order to maintain a constant total data rate (Taniguchi *et al.* [1990]). This technique is called adaptive multi-rate (AMR) coding, and requires the simultaneous implementation of an AMR source coder (Gersho and Paksoy [1999]), an AMR channel coder (Goeckel [1999], Goldsmith and Chua [1997]), and a channel quality estimation algorithm capable of acquiring information about channel conditions with a relatively small tracking delay.

The notion of determining the relative importance of bits for further unequal error protection (UEP) was pioneered by Rydbeck and Sundberg (Rydbeck and Sundberg [1976]). Rate-compatible channel codes, such as Hagenauer's rate compatible punctured convolutional codes (RCPC) (Hagenauer [1988]), are a collection of codes providing a family of channel coding rates. By puncturing bits in the bitstream, the channel coding rate of RCPC codes can be varied instantaneously, providing UEP by imparting on different segments different degrees of protection. Cox et al. (Cox *et al.* [1991]) address the issue of channel coding and illustrate how RCPC codes can be used to build a speech transmission scheme for mobile radio channels. Their approach is based on a subband coder with dynamic bit allocation proportional to the average energy of the bands. RCPC codes are then used to provide UEP.

Relatively few AMR systems describing source and channel coding have been presented. The AMR systems (Vainio *et al.* [1998], Uvliden *et al.* [1998], Paksoy *et al.* [1999], Ito *et al.* [1998]) combine different types of variable rate CELP coders for source coding with RCPC and cyclic redundancy check (CRC) codes for channel coding and were presented as candidates for the European Telecommunications Standards Institute (ETSI) GSM AMR codec standard. In (Sinha and Sundberg [1999]), UEP is applied to perceptually based audio coders (PAC). The bitstream of the PAC is divided into two classes and punctured convolutional codes are used to provide different levels of protection, assuming a BPSK constellation.

In (Bernard *et al.* [1998, 1999]), a novel UEP channel encoding scheme is introduced by analyzing how symbol-wise puncturing of symbols in a trellis code and the rate-compatibility constraint (progressive puncturing pattern) can be used to derive rate-compatible punctured trellis codes (RCPT). While conceptually similar to RCPC codes, RCPT codes are specifically designed to operate efficiently on large constellations (for which Euclidean and Hamming distances are no longer equivalent) by maximizing the residual Euclidean distance after symbol puncturing. Large constellation sizes, in turn, lead to higher throughput and spectral efficiency on high SNR channels. An AMR system is then designed based on a perceptually-based embedded subband encoder. Since perceptually based dynamic bit allocations lead to a wide range of bit error sensitivities (the perceptually least important bits being almost insensitive to channel transmission errors), the channel protection requirements are determined accordingly. The AMR systems utilize the new rate-compatible channel coding technique (RCPT) for UEP and operate on an 8-PSK constellation. The AMR-UEP system is bandwidth efficient, operates over a wide range of channel conditions and degrades gracefully with decreasing channel quality.

Systems using AMR source and channel coding are likely to be integrated in future communication systems since they have the capability for providing graceful speech degradation over a wide range of channel conditions.

12.11 Standards

Standards for land-line public switched telephone service (PSTN) networks are established by the International Telecommunication Union (ITU) (http://www.itu.int). The ITU has promulgated a number of important speech and waveform coding standards at high bit rates and with very low delay, including G.711 (PCM), G.727 and G.726 (ADPCM), and G.728 (LD-CELP). The ITU is also involved in the development of internetworking standards, including the voice over IP standard H.323. The ITU has developed one widely used low-bit-rate coding standard (G.729), and a number of embedded and multi-mode speech coding standards operating at rates between 5.3 kbps (G.723.1) and 40 kbps (G.727). Standard G.729 is a speech coder operating at 8 kbps, based on algebraic code excited LPC (ACELP) (ITU-T [1996a], Salami *et al.* [1998]). G.723.1 is a multi-mode coder, capable of operating at either 5.3 or 6.3 kbps (ITU-T [1996b]). G.722 is a standard for wideband speech coding, and the ITU will announce an additional wideband standard within a few months. The ITU has also published standards for the objective estimation of perceptual speech quality (P.861 and P.862).

The ITU is a branch of the International Standards Organization (ISO) (http://www.iso.ch). In addition to ITU activities, the ISO develops standards for the Moving Picture Experts Group (MPEG). The MPEG-2 standard included digital audio coding at three levels of complexity, including the layer-three codec commonly known as MP3 (Noll [1997]). The MPEG-4 motion picture standard includes a structured audio standard (ISO/IEC [1998a]), in which speech and audio "objects" are encoded with header information specifying the coding algorithm. Low bit-rate speech coding is performed using Harmonic Vector Excited Coding (HVXC) (ISO/IEC [1998b]) or Code Excited LPC (CELP) (ISO/IEC [1998d]), and audio coding is performed using time/frequency coding (ISO/IEC [1998c]). The MPEG home page is at drogo.cselt.stet.it/mpeg.

Standards for cellular telephony in Europe are established by the European Telecommunications Standards Institute (ETSI) (http://www.etsi.org). ETSI speech coding standards are published by the Global System for Mobile Telecommunications (GSM) subcommittee. All speech coding standards for digital cellular telephone use are based on LPC-AS algorithms. The first GSM standard coder was based on a precursor of CELP called regular-pulse excitation with long-term prediction (RPE-LTP) (Hellwig *et al.* [1989], Kroon *et al.* [1986]). Current GSM standards include the enhanced full-rate codec GSM 06.60 ([GSM], Jarvinen *et al.* [1997]) and the adaptive multi-rate codec ([GSM]); both standards use algebraic code excited LPC (ACELP). In the next few months, both the ITU and ETSI will announce new standards for wideband

	Rate (kbps)	BW (kHz)	Standards Organization	Standard Number	Algorithm	Year
Land-line Telephone	64	3.4	ITU	G.711	μ-law or A-law PCM	1988
	32	3.4	ITU	G.726	ADPCM	1990
	16-40	3.4	ITU	G.727	ADPCM	1990
Tele-conferencing	48-64	7	ITU	G.722	Split-band-ADPCM	1988
	16	3.4	ITU	G.728	Low Delay CELP	1992
Digital Cellular	13	3.4	ETSI	Full-rate	RPE-LTP	1992
	12.2	3.4	ETSI	EFR	ACELP	1997
	7.9	3.4	TIA	IS-54	VSELP	1990
	6.5	3.4	ETSI	Half-rate	VSELP	1995
	8.0	3.4	ITU	G.729	ACELP	1996
	4.75-12.2	3.4	ETSI	AMR	ACELP	1998
	1-8	3.4	CDMA-TIA	IS-96	QCELP	1993
Multimedia	5.3-6.3	3.4	ITU	G.723.1	MPLPC, CELP	1996
	2.0-18.2	3.4-7.5	ISO	MPEG-4	HVXC, CELP	1998
Satellite Telephony	4.15	3.4	INMARSAT	M	IMBE	1991
	3.6	3.4	INMARSAT	Mini-M	AMBE	1995
Secure Communications	2.4	3.4	DDVPC	FS1015	LPC-10e	1984
	2.4	3.4	DDVPC	MELP	MELP	1996
	4.8	3.4	DDVPC	FS1016	CELP	1989
	16-32	3.4	DDVPC	CVSD	CVSD	

Table 12.7: A Representative Sample of Speech Coding Standards

speech coding.

The Telecommunications Industry Association (http://www.tiaonline.org) published some of the first North American digital cellular standards, including the Vector Sum Excited LPC (VSELP) standard IS-54 (Gerson and Jasiuk [1991]). In fact, both the initial North American and Japanese digital cellular standards were based on the VSELP algorithm. Recently, the TIA has been active in the development of standard TR-41 for voice over IP.

The United States Department of Defense Voice Processing Consortium (DDVPC) publishes speech coding standards for United States government applications. As mentioned earlier, the original FS-1015 LPC-10e standard at 2.4 kbps (Campbell and Tremain [1986], DDVPC [1984]), originally developed in the 1970s, was replaced in 1996 by the newer MELP standard at 2.4 kbps (Supplee *et al.* [1997]). Transmission at slightly higher bit rates uses the FS-1016 CELP (CELP) standard at 4.8 kbps (DDVPC [1989], Jr. *et al.* [1989, 1991]). Waveform applications use the continuously variable slope delta-modulator (CVSD) at 16 kbps. Descriptions of all DDVPC standards and code for most are available at http://www.plh.af.mil/ddvpc/index.html.

12.12 Homework

Problem 12.1

a. Write a program, **XHAT = linpcm(X, XMAX, B)**, which quantizes **X** using a **B**-bit linear PCM quantizer, with maximum output values of **+/- XMAX**.

b. Record your own voice. Quantize the recorded waveform using linear PCM with 3, 4, 5, 6, and 7 bit quantization. In each case, set **XMAX=max(abs(male_sent))**. Plot SNR in decibels as a function of the number of bits.

   Listen to the error signals $e(n)$ produced at each bit rate. At low bit rates, $e(n)$ may be so highly correlated with $x(n)$ that it is actually intelligible. Are any of your error signals intelligible? Which ones?

Problem 12.2

Write a function, **T = mulaw(X, MU, XMAX)**, which compresses **X** using μ-law compression (R&S 5.3.2). Write another function, **X = invmulaw(T, MU, XMAX)**, which expands **t** to get **x** again.

Check your work by finding
XHAT=invmulaw(mulaw(X, 255, XMAX), 255, XMAX)

where **X** is a recording of your voice. Confirm that **XHAT** is identical to **X**.[17]

Quantize **T** using the function **linpcm**, and expand the quantized version to obtain a μ-law quantized version of the input sentence. In this way, create μ-law quantized sentences using 3, 4, 5, 6, and 7 bit quantizers, with a value of $\mu = 255$. Plot SNR as a function of the number of bits, and compare to the values obtained with linear PCM.

Problem 12.3

Write a program that accepts two input waveforms (an original and a coded waveform), and computes the SNR of the coded waveform.

Write another program that computes the segmental SNR of the coded waveform. This program should divide the signal **X** and error **E=XHAT-X** into frames of **N** samples each, compute the SNR (in decibels) within each frame, and return the average segmental SNR.

---

[17] . . . with allowance for floating-point roundoff error.

Compute the SNR and SEGSNR of uniform PCM and companded PCM at bit rates of 3, 4, 5, and 6 bits/sample, and plot the results. How do SEGSNR and SNR differ?

Sort the quantized utterances (including both linear and companded PCM) on a scale from lowest to highest SNR. Now sort them from lowest to highest SEGSNR. Finally, sort them on a scale from "worst sounding" to "best sounding." Which of the two objective measures (SNR or SEGSNR) is a better representation of the subjective speech quality?

Problem 12.4

Consider a signal $x(n)$ with a unit-variance Gaussian distribution:

$$p_{x(n)}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{12.311}$$

a. Design a 1-bit-per-sample scalar quantizer for $x(n)$. Find the decision boundary and reconstruction levels which minimize the expected mean squared error.

b. Generate a matlab data vector $x(n)$ consisting of 1000 Gaussian random numbers. Cluster your data into two regions using either the K-means algorithm or the binary-split algorithm.

   What are the centroids of your two regions? What is the decision boundary?

c. Comment on any differences between your answers in parts (a) and (b), and explain them if you can. You may or may not find it useful to calculate the mean and the histogram of your data vector $x(n)$.

Problem 12.5

A particularly good model of the distribution of speech audio samples is the "Laplacian" distribution, given by

$$p_x(x_0) = \frac{1}{2} e^{-|x_0|} \tag{12.312}$$

Assume that you are to quantize a signal with the PDF given in equation 12.312. Design a uniform, zero-centered, mid-tread quantizer to satisfy the following criteria:

- $P_{clip} = e^{-10}$.
- When there is no clipping, $\sigma_e^2 \leq \frac{1}{48}$.
- The bit rate is as low as possible in order to satisfy the criteria given above.

Specify the clipping threshold (T_Q), the bit rate per sample (B), the number of reconstruction levels (Q), and the spacing of reconstruction levels (Δ).

Problem 12.6

Consider a PQMF filterbank, with filters given by

$$h_k[n] = 2h[n]\cos\left(\frac{\pi(k+1/2)n}{K} + \phi_k\right)$$

where

$$H(\omega) = \begin{cases} 1 - \frac{K|\omega|}{\pi} & |\omega| < \frac{\pi}{K} \\ 0 & \text{otherwise} \end{cases}$$

Suppose that $K = 2$, and that

$$\phi_k = \left(k + \frac{1}{2}\right)\frac{\pi}{2}$$

Sketch $H_{0R}(\omega)$, $H_{0I}(\omega)$, $H_{1R}(\omega)$, and $H_{1I}(\omega)$, the real and imaginary parts of $H_0(\omega)$ and $H_1(\omega)$.

Problem 12.7

Consider a set of pseudo quadrature mirror filters designed with the following properties. There are $K = 2$ bands. The prototype lowpass filter $h[n]$ should have a passband of $[-\frac{\pi}{4}, \frac{\pi}{4}]$, and a transition band of $[\frac{\pi}{4}, \frac{\pi}{2}]$, but in order to keep things from getting too absurdly complicated, please assume the following prototype lowpass filter:

$$H(\omega) = \begin{cases} 1 & -\frac{\pi}{2} \le \omega \le \frac{\pi}{2} \\ 0 & \text{otherwise} \end{cases} \tag{12.313}$$

The two bandpass filters are designed as follows:

$$h_k[n] = 2h[n]\cos\left(\frac{\pi(2k+1)}{2K}n + \phi_k\right), \quad k \in \{0, 1\} \tag{12.314}$$

where the phase offset is given by

$$\phi_k = \frac{\pi(2k+1)}{4} \tag{12.315}$$

Suppose that the input to the sub-band coder is an impulse,

$$x[n] = \delta[n] \tag{12.316}$$

The remainder of this problem will ask you to sketch various spectra. In each case, be sure to sketch the requested spectrum either over the range $[-\pi, \pi]$ or over the range $[0, 2\pi]$ (one range will be easier to use for some spectra, the other will be easier for other spectra). Two types of labels are required for every plot: (1) label frequencies of every DTFT discontinuity between 0 and π, and (2) label the height of the graph at the three frequencies $\omega = 0$, $\frac{\pi}{2}$, and π.

a. Consider the two sub-band filtered signals $x_k[n] = h_k[n] * x[n]$. Let $X_k(\omega) = X_{kR}(\omega) + jX_{kI}(\omega)$, i.e., the real part is $X_{kR}(\omega)$ and the imaginary part is $X_{kI}(\omega)$. Write equations expressing $X_{0R}(\omega)$, $X_{0I}(\omega)$, $X_{1R}(\omega)$, and $X_{1I}(\omega)$ as functions of the prototype filter spectrum $H(\omega)$. Sketch these four spectra as functions of frequency.

b. Suppose that the sub-band signals are downsampled by a factor of K to create signals $d_k[n]$. The spectrum of $d_k[n]$ includes an "unaliased" term (the $i = 0$ term in the following equation), and an "aliasing" term (the $i = 1$ term in the following equation):

$$D_k(\omega) = \frac{1}{K} \sum_{i=0}^{K-1} X_k\left(\frac{\omega - 2\pi i}{K}\right) \tag{12.317}$$

Call the real and imaginary parts of the aliased and unaliased components $D^i_{kR}(\omega)$ and $D^i_{kI}(\omega)$, respectively, i.e.

$$D^i_{kR}(\omega) = X_{kR}\left(\frac{\omega - 2\pi i}{K}\right) \tag{12.318}$$

and likewise for $D^i_{kI}(\omega)$.

Draw the following twelve spectra:

- Six real parts: unaliased components $D^0_{kR}(\omega)$, aliasing components $D^1_{kR}(\omega)$, and their sum $D_{kR}(\omega) = D^0_{kR}(\omega) + D^1_{kR}(\omega)$.
- Six imaginary parts: unaliased components $D^0_{kI}(\omega)$, aliasing components $D^1_{kI}(\omega)$, and their sum $D_{kI}(\omega) = D^0_{kI}(\omega) + D^1_{kI}(\omega)$.

c. The first operation implemented by the decoder is to upsample the received signals $d_k[n]$, as follows:

$$v_k[n] = \begin{cases} d_k[n/K] & n = \text{ integer multiple of } K \\ 0 & \text{otherwise} \end{cases} \tag{12.319}$$

Draw the following twelve spectra:

- Six real parts: unaliased components $V^0_{kR}(\omega)$, aliasing components $V^1_{kR}(\omega)$, and their sum $V_{kR}(\omega) = V^0_{kR}(\omega) + V^1_{kR}(\omega)$.

- Six imaginary parts: unaliased components $V_{kI}^0(\omega)$, aliasing components $V_{kI}^1(\omega)$, and their sum $V_{kI}(\omega) = V_{kI}^0(\omega) + V_{kI}^1(\omega)$.

d. The second operation implemented by the decoder is to filter the upsampled spectra as follows:

$$\hat{X}_k(\omega) = KH_k^*(\omega)V_k(\omega) \tag{12.320}$$

Draw the following twelve spectra:

- Six real parts: unaliased components $\hat{X}_{kR}^0(\omega)$, aliasing components $\hat{X}_{kR}^1(\omega)$, and their sum $\hat{X}_{kR}(\omega) = \hat{X}_{kR}^0(\omega) + \hat{X}_{kR}^1(\omega)$.
- Six imaginary parts: unaliased components $\hat{X}_{kI}^0(\omega)$, aliasing components $\hat{X}_{kI}^1(\omega)$, and their sum $\hat{X}_{kI}(\omega) = \hat{X}_{kI}^0(\omega) + \hat{X}_{kI}^1(\omega)$.

e. The final operation implemented by the decoder is to add together the sub-band signals,

$$\hat{x}[n] = \sum_{k=0}^{K-1} \hat{x}_k[n] \tag{12.321}$$

Draw the following six spectra:

- Three real parts: unaliased components $\hat{X}_R^0(\omega)$, aliasing components $\hat{X}_R^1(\omega)$, and their sum $\hat{X}_R(\omega) = \hat{X}_R^0(\omega) + \hat{X}_R^1(\omega)$.
- Three imaginary parts: unaliased components $\hat{X}_I^0(\omega)$, aliasing components $\hat{X}_I^1(\omega)$, and their sum $\hat{X}_I(\omega) = \hat{X}_I^0(\omega) + \hat{X}_I^1(\omega)$.

Problem 12.8

Write a program which reads in a frame of speech, calculates LPC coefficients, filters the input by $A(z)$ to find an LPC residual, and then filters the residual by $H(z) = 1/A(z)$ to synthesize a speech waveform. Verify that the output is identical to the non-overlapping part of the input.

a. It is important to carry the state of all digital filters forward from frame to frame. What happens if the state of the LPC synthesis filter is carried forward from frame to frame, but not the state of the analysis filter? What happens if the state of the analysis filter is carried forward, but not the state of the synthesis filter? What happens if neither filter state is carried forward? (Hint: compare the LPC residual with and without filter state carry-over.)

b. Quantize the LPC coefficients using log area ratio quantization. Aim for an average of 4-5 bits per LPC coefficient. Verify that when you use the same quantized filter $\hat{A}(z)$ for both analysis and synthesis, LPC quantization does not introduce any errors into the reconstructed signal. How many bits per second are you using to quantize the LPC coefficients?

Problem 12.9

Write a program **[PF, QF] = tf2lsf(A)** which computes ordered line spectral frequencies corresponding to the LPC coefficients in **A**. The LSF matrices **PF** and **QF** should each contain as many rows as **A**, and $p/2$ columns.

a. Plot the analog P-frequencies, $p_n F_s/2\pi$, and compare to a spectrogram of the utterance. Do the P-frequencies track the formants during voiced segments? What happens if there are 5 P-frequencies, but only 4 trackable formants? What do the P-frequencies do during unvoiced segments? Do you think this behavior is likely to make the P-frequencies more or less quantizable than the LPC-based formant frequencies? Why?

   Plot the Q-frequencies $q_n F_s/2\pi$. How do the Q-frequencies relate to the P-frequencies? Is there ever a time when a formant frequency is tracked more closely by q_n than by p_n?

   How rapidly do the line spectral frequencies change as a function of time? What is the range of each individual line spectral frequency? What is the range of the difference between neighboring line spectral frequencies? Can you think of an efficient way of quantizing the line spectral frequencies?

b. Quantize the LPC coefficients using LSF quantization, then convert the quantized LSFs back into direct-form LPC coefficients, and synthesize speech. How does the speech sound? How can you guarantee that the quantized LPC synthesis filter will be stable?

Problem 12.10

Write a program **DHAT = vocode(N0, B, THETA)** which creates a simulated LPC excitation signal **DHAT**.

In frames for which **B(i)<THETA**, the excitation signal should be unit-variance uncorrelated Gaussian noise (use the **randn** function). In frames for which **B(i)>THETA**, the excitation signal should be an impulse train with a period of **N0** samples, starting at some sample number called **START**.

When two neighboring frames are voiced, make sure to set the variable **START** so that the spacing between the last impulse in one frame and the first impulse in the next frame is a valid pitch period — otherwise, you will hear a click at every frame boundary! Also, be careful to set the amplitude of the impulse train so that the average energy of the excitation signal in each frame is always

$$\frac{1}{\text{STEP}} \sum_{n=1}^{\text{STEP}} \hat{d}^2(n) = 1 \tag{12.322}$$

Create a signal **DHAT** using a value of **THETA** which you think will best divide voiced and unvoiced segments, and using values of **N0** and **B** calculated from a natural utterance. Pass the output of the **vocode** function through your **lpcsyn** function. Listen to the sentence. How does the coding quality compare to 3 and 4 bit linear PCM? How about companded PCM? Do certain parts of the sentence sound better than others?

Calculate the SNR and SEGSNR of your vocoded utterance. Is the SEGSNR of an LPC vocoder better or worse than the SEGSNR of a PCM coder with similar perceptual quality? Why?

Problem 12.11

Write a program **DHAT = ppv(N0, B)** that creates a synthesized LPC excitation signal by filtering white noise through a "pitch-prediction filter." Be careful to carry filter state forward from one frame to the next.

Use **ppv** to create an excitation signal **DHAT**, and pass it to **lpcsyn** to synthesize a copy of your input sentence. How does this version of the sentence sound? How does it compare to the regular vocoder? Do some parts of the sentence sound better? Do some parts sound worse?

Problem 12.12

Synthesize a sentence, using either the regular vocoder or the pitch-prediction vocoder, but use one of the following modified parameters in place of the correct parameter. What does the sentence sound like? Why?

N0_modified = median(N0) * ones(size(N0));
N0_modified = round(0.5*N0);
B_modified = zeros(size(B));

Problem 12.13

In Self-Excited LPC (SELP), the LPC excitation vector $u(n)$ is created by scaling and adding past samples of itself:

$$u(n) = \sum_{i=-I}^{I} b_i u(n-D-i) \quad \Leftrightarrow \quad U(z) = \frac{1}{P(z)} = \frac{1}{1-\sum_{i=-I}^{I} b_i z^{-(D+i)}} \tag{12.323}$$

In order to obtain useful LPC excitation values, the samples of $u(n)$ for $n<0$ (before the beginning of the sentence) are initialized using samples of Gaussian white noise.

a. **Pitch Prediction Coefficients for a Perfectly Periodic Signal**

Self-excited LPC implicitly assumes that the ideal continuous-time excitation signal $u_c(t)$ is perfectly periodic, but that the period may not be an exact multiple of the sampling period T:

$$U_c(s) = \frac{1}{P_c(s)} = \frac{1}{1-e^{-sT_0}}, \quad T_0 = DT + \epsilon \tag{12.324}$$

Quest. 1: Choose coefficients b_i such that $p(n) = p_a(nT)$, where $p_a(t)$ is obtained by lowpass-filtering $p_c(t)$ at the Nyquist rate. You may assume that $I=\infty$.

b. **Pitch Prediction Coefficients for a Real-Life Signal**

Equation 12.323 can be written in vector form as follows:

$$U = BX \tag{12.325}$$

$$X = \begin{bmatrix} u(n-(D-1)) & \dots & u(n-(D-1)+L-1) \\ u(n-D) & \dots & u(n-D+L-1) \\ u(n-(D+1)) & \dots & u(n-(D+1)+L-1) \end{bmatrix}, \qquad B = [b_{-1},\ b_0,\ b_1] \tag{12.326}$$

Quest. 2: Find the value of B which minimizes the squared quantization error $|E|^2 = |UH - \tilde{S}|^2$. Hint: The derivation follows Kondoz equations 6.21-6.28.

c. **Sub-Frames**

The pitch prediction delay D is chosen by searching over some range of possible pitch periods, $D_{min} \le D \le D_{max}$, and choosing the value of D which minimizes the squared quantization error $|E|^2$. If there is no overlap between the samples of X and U, that is, if $D > L$ for all possible D, then it is possible to compute the squared error $|E|^2$ using pure matrix computation. Matrix computation reduces the computational complexity, and (in matlab) it reduces the programming time by a lot; therefore all practical systems set the minimum value of D to $D_{min} = L + 1$.

LPC is almost never computed using frames of size $L < D$, so analysis by synthesis systems often break up each LPC frame into M sub-frames, where M is some integer:

$$L = \frac{L_{LPC}}{M} \tag{12.327}$$

LPC coefficients are thus calculated using frames of $\frac{L_{LPC}}{F_s} \approx 20-30$ms in length, but the excitation parameters D and B are calculated using smaller frames of length L.

Even if $L = L_{LPC}/M$, very short pitch periods may be shorter than one sub-frame in length. Fortunately, pitch prediction, as shown in equation 12.323, works pretty well if D is any small integer multiple of the pitch period, for example $D = 2T_0/T$. The accuracy of pitch prediction drops slightly, however, so it is best for D to be no more than 2-3 times the pitch period.

> Quest. 3: What is the largest sub-frame size L which will allow you to represent pitches up to $F_0 = 250$Hz with a delay D of no more than twice the pitch period?

d. Implementation

Implement self-excited LPC.

- For each sub-frame, calculate all of the possible values of U_D for $D_{min} \leq D \leq D_{max}$, where $D_{min} = L + 1$ and D_{max} is the pitch period of a low-pitched male speaker ($F_0 \approx 80Hz$).
- For each value of U_D, calculate the squared quantization error $|E|^2$.
- Choose the value of D which minimizes $|E|^2$.

Be sure to carry forward, from frame to frame, both the LPC filter states, and enough delayed samples of U to allow you to perform pitch prediction. The LPC filter states should be initialized to zero at the beginning of the sentence, but the pitch prediction samples should be initially set to unit-variance Gaussian white noise.

Quantize D and B. Examine the segmental SNR and sound quality of your coder using both quantized and unquantized D and B.

Quest. 4: What segmental SNR do you obtain using unquantized D and B? With quantized D and B? How many bits per second do you need to transmit $A(z)$, D, and B?

Problem 12.14

In this problem, you will design a multi-vector LPC analysis-by-synthesis coder. This coder is designed to correct some of the problems of self-excited LPC.

a. **Stochastic Codevectors**

We can represent aperiodic and partially periodic signals by extending the excitation matrix X as follows:

$$X_M = \begin{bmatrix} u(n-(D-1)) & \dots & u(n-(D-1)+L-1) \\ u(n-D) & \dots & u(n-D+L-1) \\ u(n-(D+1)) & \dots & u(n-(D+1)+L-1) \\ c_{k1}(0) & \dots & c_{k1}(L-1) \\ c_{k2}(0) & \dots & c_{k2}(L-1) \\ \vdots & \vdots & \vdots \\ c_{kM}(0) & \dots & c_{kM}(L-1) \end{bmatrix} \tag{12.328}$$

Where $c_{k1}(m)$ are the samples of a "code vector" C_{k1} which is chosen from a set of K possible code vectors in order to minimize the squared error,

$$|E|^2 = |\hat{S} - S|^2 = |BXH - \tilde{S}|^2 \tag{12.329}$$

In the original CELP algorithm, the codebook consists of K Gaussian random vectors. In MPLPC, the codevectors are impulses:

$$\begin{aligned} \text{CELP:} \quad c_k(m) &\sim \mathcal{N}(0,1), \quad 0 \le k \le K-1 && (12.330) \\ \text{MPLPC:} \quad c_k(m) &= \delta(k), \quad 0 \le k \le L-1 && (12.331) \\ & && (12.332) \end{aligned}$$

Quest. 1: Suppose you are creating an MPLPC coder in which the X matrix you used in your SELP coder will be augmented by 5 impulse vectors, numbered C_{k1} through C_{k5}. If you want to find the globally optimum combination of D, $k1$, $k2$, $k3$, $k4$, and $k5$, how many times do you need to evaluate equation 12.329?

b. **Iterative Optimization**

In order to avoid the impossible computational complexity of a global search, many CELP coders and all MPLPC coders perform an iterative search. In an iterative search, the best pitch delay D is calculated as in the SELP coder, resulting in a $3 \times L$ matrix X_0, and a 3-element gain vector B_0:

$$\tilde{S} \approx B_0 \hat{S}_0 = B_0 X_0 H \tag{12.333}$$

Given X_0 and B_0, the fourth excitation vector can be chosen by first creating a reference vector $\tilde{S}_1$,

$$\tilde{S}_1 = \tilde{S} - B_0 X_0 H \tag{12.334}$$

$\tilde{S}_1$ represents the part of $\tilde{S}$ which is not well represented by $B_0 \hat{S}_0$; in fact, $\tilde{S}_1$ is the part of $\tilde{S}$ which is orthogonal to $B_0 \hat{S}_0$. Therefore, the optimum fourth excitation vector is the one which minimizes

$$|E_1^k|^2 = |\tilde{S}_1 - g_1 C_{k1} H|^2 \tag{12.335}$$

Once the optimum value of $k1$ has been found, the gain vector B_1 must be recomputed, in order to minimize the total squared error

$$|E_1|^2 = |\tilde{S} - B_1 X_1 H|^2 \tag{12.336}$$

Quest. 2: Find the optimum value of g_1 as a function of the reference vector $\tilde{S}_1$ and the codebook vector C_{k1}. Find the optimum value of B_1 as a function of $\tilde{S}$ and X_1. Show that, in general, $B_1 \neq [B_0 g_1]$.

Once B_1 and X_1 have been computed, the procedure outlined above can be repeated, as often as necessary. Typically, the number of pulse vectors required to achieve good quality using MPLPC is a fraction of L. Classical CELP coders only use one stochastic codevector, but the VSELP algorithm used in most cellular telephone standards uses two.

c. **Implementation**

Add a stochastic codebook to your SELP coder, in order to create a generalized LPC-AS coder (be sure to save your SELP coder under a different name, so that you can go back to it if you need to!) Your generalized LPC-AS coder should accept as input a $K \times L$ codebook matrix C, each of whose rows contains one stochastic code vector C_k. You should also accept an input argument M which tells the function how many stochastic codevectors to find. The final command line might look something like this:

```
[YMAT(i,:), filter_states, ...] = myfunc(XMAT(i,:), filter_states, ..., C, M)
```

Create an MPLPC codebook (consisting of impulses) and a CELP codebook (consisting of about 1000 Gaussian random vectors). Test your coder using both CELP and MPLPC codebooks.

Quest. 3: Plot the segmental SNR of your coded speech as a function of the number of stochastic code vectors, M, for both MPLPC and CELP, with the codebook gain vector B still unquantized. Comment on any differences between MPLPC and CELP.

Quantize the codebook gain vector B.

Quest. 4: Quantize all of your gain terms, then choose coder configurations for both CELP and MPLPC which produce reasonable-sounding speech. For both of these two configurations, what is the total bit rate, in bits per second?

Problem 12.15

When humans listen to a coded speech signal, some components of the quantization noise are masked by peaks in the signal spectrum. Low bit rate LPC-AS systems therefore often weight the error, in order to reduce the importance of error components near a spectral peak. The most common error weighting filter is based on the LPC analysis filter $A(z)$:

$$E_w(z) = E(z)\frac{A(z)}{A(z/\alpha)} \tag{12.337}$$

Write a program that perceptually weights the error signal of a speech coder. Compare the perceptual qualities predicted by SNR, SEGSNR, and perceptually weighted SEGSNR.

Bibliography

Abate, J. E. (**1967**), "Linear and adaptive delta modulation," Proc. IEEE **55**, 298–308.

Adoul, J.-P., Mabilleau, P., Delprat, M., and Morisette, S. (**1987**), "Fast CELP coding based on algebraic codes," in *Proc. ICASSP*, pp. 1957–1960.

Ahlfors, L. V. (**1953**), *Complex analysis* (McGraw-Hill Book Co., New York).

Alipour, F. and Scherer, R. C. (**Feb. 1995**), "Pulsatile airflow during phonation: An excised larynx model," J. Acoust. Soc. Am. **97**(2), 1241–1248.

Alipour-Haghigi, F. and Titze, I. R. (**1983**), "Simulation of Particle Trajectories of Vocal Fold Tissue During Phonation," in *Vocal Fold Phys.: Biomech., Acoust., and Phonatory Control*, edited by I. R. Titze and R. C. Scherer (Denver, CO), pp. 183–190.

Alipour-Haghigi, F. and Titze, I. R. (**1985**), "Viscoelastic modeling of canine vocalis muscle in relaxation," J. Acoust. Soc. Am. **78**, 1939–1943.

Alipour-Haghigi, F. and Titze, I. R. (**1987**), "Twitch response in the canine vocalis muscle," J. Speech Hear. Res. **30**, 290–294.

Alipour-Haghigi, F. and Titze, I. R. (**1989**), "Tetanic contraction in vocal fold muscle," J. Speech Hear. Res. **32**, 226–231.

Alipour-Haghigi, F. and Titze, I. R. (**Sept. 1991**), "Elastic models of vocal fold tissues," J. Acoust. Soc. Am. **90**(3), 1326–1331.

Allen, J. (**March 1971**), "Speech synthesis from unrestricted text," in *IEEE Int. Conv. Digest* (New York).

Allen, J. B. (**1994**), "How do humans process and recognize speech?" IEEE Transactions on speech and audio **2**(4), 567–577.

Atal, B. S. (**1962**), "Generalized short-time power spectra and autocorrelation function," J. Acoust. Soc. Am. **34**, 1679–1683.

Atal, B. S. (**1982**), "Predictive coding of speech at low bit rates," IEEE Trans. Comm. **30**, 600–614.

Atal, B. S. (**1986**), "High-quality speech at low bit rates: multi-pulse and stochastically excited linear predictive coders," in *Proc. ICASSP*, pp. 1681–1684.

Atal, B. S. and Hanauer, S. L. (**1971**a), "Low-bit-rate speech transmission by linear prediction of speech signals," J. Acoust. Soc. Am. **49**, 133 (A).

Atal, B. S. and Hanauer, S. L. (**1971**b), "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Am. **50**, 637–655.

Atal, B. S. and Remde, J. R. (**1982**), "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Proc. ICASSP*, pp. 614–617.

Atal, B. S. and Schroeder, M. R. (**1956**), "On the separation and measurement of formant frequencies," J. Acoust. Soc. Am. **28**, 159 (A).

Ayers, E. W. (**August 1959**), "Speech synthesizers using formant principles," Tech. Rep. 20315, British Post Office Res. Station.

Barney, A., Shadle, C. H., and Davies, P. (**1999**), "Fluid flow in a dynamic mechanical model of the vocal folds and tract. I Measurements and theory," J. Acoust. Soc. Am. **105**, 444–455.

Barney, H. L. and Dunn, H. K. (**1957**), "Speech analysis; Speech synthesis; Chapters 12 and 13," in *Manual of phonetics*, edited by L. Kaiser (North-Holland Publ. Co., Amsterdam).

Bastide, R. P. and Smith, C. P. (**1955**), "Electrical synthesizer of continuous speech," J. Acoust. Soc. Am. **27**, 207 (A).

Bauer, B. (**apr 1944**), "Note on Radiation Impedance," J. Acoust. Soc. Am. **15**(4), 223–224.

Baumann, R. H., Licklider, J. C. R., and Howland, B. (**1954**), "Electronic word recognizer," J. Acoust. Soc. Am. **26**, 137 (A).

Bekesy, G. V. (**1943**), "Uber die Resonanzkurve und die Abklingzeit der verschiedenen Stellen der Schneckentrennwand," Akust. Z. **8**, 66–76.

Bekesy, G. V. (**1953**), "Shearing microphonics produced by vibrations near the inner and outer hairs cells," J. Acoust. Soc. Am. **25**, 786–790.

Bekesy, G. V. (**1960**), *Experiments in hearing* (McGraw-Hill Book Co., New York).

Bell, C. G., Fujisaki, H., Heinz, J. M., Stevens, K. N., and House, A. S. (**1961**), "Reduction of speech spectral by analysis-by-synthesis techniques," J. Acoust. Soc. Am. **33**, 1725–1736.

Bennett, W. R. (**1941**), "Time-division multiplex systems," Bell System Tech. J. **20**, 199–221.

Bennett, W. R. (**1953**), "The correlatograph," Bell System Tech. J. **32**, 1173–1185.

Beranek, L. L. (**1947**), "The design of speech communication systems," Proc. LR.E. **35**, 880–890.

Beranek, L. L. (**1954**), *Acoustics* (McGraw-Hill Book Co., New York).

Bernard, A., Liu, X., Wesel, R., and Alwan, A. (**1998**), "Channel adaptive joint-source channel coding of speech," in *Proceedings of the 32nd Asilomar conference on signals, systems, and computers*, vol. 1, pp. 357–361.

Bernard, A., Liu, X., Wesel, R., and Alwan, A. (**1999**), "Embedded joint-source channel coding of speech using Symbol Puncturing of Trellis codes," in *Proc. IEEE ICASSP*, vol. 5, pp. 2427–2430.

Biddulph, R. (**1954**), "Short-term autocorrelation analysis and correlatograms of spoken digits," J. Acoust. Soc. Am. **26**, 539–544.

Blackman, R. B. and Tukey, T. W. (**1959**), *The measurement of power spectra* (Dover Publications,, New York).

Bliss, J. C. (**1962**), "Kinesthetic-tractile communications," IRE Trans. Inform. Theory **IT-8**, 92–99.

Bloch, B. and Trager, G. L. (**1942**), *Outline of linguistic analysis* (Waverly Press, Baltimore).

Bogert, B. P. (**1951**), "Determination of the effects of dissipitation in the cochlear partition by means of a network representing the basilar membrane," J. Acoust. Soc. Am. **23**, 151–154.

Bogert, B. P. (**1956**), "The vobanc–a two-to-one spech bandwidth reduction system," J. Acoust. Soc. Am. **28**, 399–404.

Bogert, B. P., Healy, M. J. R., and Tukey, J. W. (**1962**), "The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking," in *Proc. Symposium Time Series Analysis* (Wiley and Sons, New York), pp. 209–243.

Bogert, B. P., Healy, M. J. R., and Tukey, J. W. (**1963**), "The frequency analysis of time-series for echoes," in *Proc. Symp. Time Series Analysis*, edited by M. Rosenblatt, pp. 209–243.

Bolt, R. H. (**1970**), "Speaker identification by speech spectrograms: A scientists' view of its reliability for legal purposes," J. Acoust. Soc. Am. **47**, 597–612.

Bondarko, L. U., Zagoruyko, N. G., Kozevnikov, V. A., Molchanov, A. P., and Chistovich, L. A. (**1968**), "A model of human speech perception," Tech. rep., Acad. Sci., U.S.S.R., Sibirsk, Nauka.

Born, M. and Wolf, E. (**1959**), *Principles of Optics: Electromagnetic Theory of Propagation, Int erference and Diffraction of Light* (Pergamon Press, Oxford).

Borst, J. M. (**1956**), "The use of spectrograms for speech analysis and synthesis," J. Audio Eng. Soc. **4**, 14–23.

Brady, P. T., House, A. S., and Stevens, K. N. (**1961**), "Perception of sounds characterized by a rapidly changing resonant frequency," J. Acoust. Soc. Am. **33**, 1357–1362.

Brandstein, M., Monta, P., Hardwick, J., and Lim, J. (**1990**), "A real-time implementation of the improved MBE speech coder," Proc. ICASSP **1**, 5–8.

Bricker, P. D. and Flanagan, J. L. (**1970**), "Subjective assessment of computer-simulated telephone calling signals," IEEE Trans. Audio and Electroacoust. **AU-18**, 19–25.

Brillouin, L. (**1953**), *Wave propagation in periodic structures* (Dover, London), 1946 edition, with corrections and new appendix.

Brillouin, L. (**1962**), *Science and Information Theory* (Academic Press Inc., London), first edition 1956.

Bullington, K. and Fraser, J. M. (**1959**), "Engineering aspects of TASL," Bell System Tech J. **38**, 353–364.

Campanella, S. J., Coulter, D. C., and Irons, R. (**September 1962**), "Influence of transmission error on formant coded compressed speech signals," in *Proc. Stockholm Speech Comm. Seminar, RLT* (Stockholm, Sweden).

Campbell, J. P. and Tremain, T. E. (**1986**), "Voiced/Unvoiced Classification of Speech with Applications to the U.S. Government LPC-10E Algorithm," in *Proc. ICASSP*, pp. 473–476.

Carlson, A. B. (**1968**), *Communication systems* (McGraw-Hill Book Co., New York).

Carroll, J. D. (**1972**), "Individual differences and multidimensional scaling," in *Multidimensional scaling: Theory and applications in the behavioral sciences*, edited by R. N. Shepard, A. K. Romney, and S. Nerlove (Seminar Press).

CDMA (**1992**), "Wideband Spread Spectrum Digital Cellular System Dual-Mode Mobile Station-Base Station Compatibility Standard," Tech. Rep. Proposed EIA/TIA Interim Standard, Telecommunications Industry Association TR45.5 Subcommittee.

Chang, S.-H. (**1956**), "Two schemes of speech compression system," J. Acoust. Soc. Am. **28**, 565–572.

Chen, J.-H., Cox, R. V., Lin, Y.-C., Jayant, N., and Melchner, M. J. (**1992**), "A low delay CELP coder for the CCITT 16 kb/s speech coding standard," IEEE J. Sel. Areas Commun. **10**, 830–849.

Chen, J.-H. and Gersho, A. (**1995**), "Adaptive postfiltering for quality enhancement of coded speech," IEEE Trans. Speech Audio Process. **3**(1), 59–71.

Cherry, C. (**1957**), *On human communication* (John Wiley & Sons, New York).

Cherry, L. (**1969**), "Excitation of vocal tract synthesizers," J. Acoust. Soc. Am. **45**, 764–769.

Chiba, T. and Kajiyama, M. (**1941**), *The vowel, its nature and structure* (TokyoKaiseikan Pub. Co., Tokyo).

Chistovich, L. A. (**1955**), "On the discrimination of complex audio signals, report 1," Problemy Fiziol. Akust. **3**, 18–26.

Chistovich, L. A. (**August 1962**), "Temporal course of speech sound perception," in *Proc. Internat. Congr. Acoust.* (Copenhagen, Denmark).

Chistovich, L. A., Klaas, A. Y., and Alekin, R. . (**1961**), "The importance of imitation in the recognition of sound sequences," Vopr. Psikhol. **5**, 173–182.

Chistovich, L. A., Kozhevnikov, V. A., and Alyakrinskü, V. V. (**1965**), "Speech, articulation and perception," Tech. rep., Acad. Sci. U.S.S.R., Nauka.

Cohen, A. and T'Hart, J. (**September 1962**), "Speech synthesis of steady-state segments," in *Proc. Stockholm Speech Comm. Seminar, R.I.T* (Stockholm, Sweden).

Coker, C. H. (**1965**), "Real-time formant vocoder, using a filter bank, a general, purpose digital computer, and an analog synthesizer," J. Acoust. Soc. Am. **38**, 940 (A).

Coker, C. H. (**1968**), “Speech synthesis with a parametric articulatory model,” in *Proc. Kyoto Speech Symposium* (Kyoto, Japan), vol. A-4-1-A-4-6.

Coker, C. H. (**April 1972**), “An experiment in computer communications through a data loop,” Bell System Tech. J. .

Coker, C. H. and Cummiskey, P. (**1965**), “On-line computer control of a formant synthesizer,” J. Acoust. Soc. Am. **38**, 940(A).

Coker, C. H., Umeda, N., and Browman, C. P. (**March 1971**), “Automatic synthesis from text,” in *IEEE Int. Conv. Digest.* (New York).

Cooley, J. W. and Tukey, J. W. (**1965**), “An algorithm for the machine calculation of complex Fourier series,” Math. Compo **19**, 297–301.

Cooper, F. S. (**1950**), “Spectrum analysis,” J. Acoust. Soc. Am. **22**, 761–762.

Cooper, F. S. (**1957**), “A bandwidth compression device,” J. Acoust. Soc. Am. **29**, 777 (A).

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., and Gerstman, L. J. (**1952**), “Some experiments on the perception of synthetic speech sounds,” J. Acoust. Soc. Am. **24**, 597–606.

Cover, T. M. and Thomas, J. A. (**1991**), *Elements of information theory* (John Wiley & Sons, Inc., New York).

Cox, R., Gay, S., Shoham, Y., Quackenbush, A., Seshadri, N., Jayant, N., Conway, J. H., and Sloane, N. J. A. (**February 1988**), “New Directions in Subband Coding,” IEEE JSAC **6**(2), 391–409.

Cox, R., Hagenauer, J., Seshadri, N., and Sundberg, C. (**August 1991**), “Subband Speech Coding and Matched Convolutional Coding for Mobile Radio Channels,” IEEE Tr. on Signal Processing **39**(8), 1717–1731.

Cranen, B. and Boves, L. (**March 1987**), “On Subglottal Formant Analysis,” J. Acoust. Soc. Am. **81**(3), 734–746.

Crowther, W. R. and Rader, C. M. (**1966**), “Efficient coding of vocoder channel signals using linear transformation,” Proc. IEEE **54**, 1594–1595.

Daguet, J. (**September 1962**), "“Codimex” speech compression system,” in *Proc. Stockholm Speech Comm. Seminar, R.I.T* (Stockholm, Sweden).

Das, A. and Gersho, A. (**1999**), “Low-rate multimode multiband spectral coding of speech,” International Journal of Speech Technology **2**(4), 317–327.

Das, S. K. and Mohn, W. S. (**1969**), "Pattern recognition in speaker verification," Proc. Fall Joint Computer Conference , 721–732.

Davidson, G. and Gersho, A. (**1986**), "Complexity reductio methods for vector excitation coding," in *Proc. ICASSP*, pp. 2055–2058.

Davis, H. (**1951**), "Chapter 28," in *Handbook of experimental psychology*, edited by S. S. STEVENS (John Wiley & Sons, New York).

Davis, H. (**1957**), "Chapter 4," in *Handbook of noise control*, edited by C. M. Harris (McGraw-Hill Book Co., New York).

Davis, H. (**1958**), "A mechano-electrical theory of cochlear action," Ann. Otol. Rhinol. Laryngol. **67**, 789–801.

Davis, H. (**1965**), "A model for transducer action in the cochlea," Cold Spring Harbor Symp. Quant. Biol. **30**, 181–190.

Davis, K. H., Biddulph, R., and Balashek, S. (**1952**), "Automatic recognition of spoken digits," J. Acoust. Soc. Am. **24**, 637–642.

Davis, S. and Mermelstein, P. (**August 1980**), "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," Trans. ASSP **ASSP-28**(4), 357–366.

DDVPC (**Nov. 1984**), "LPC-10e Speech Coding Standard," Tech. Rep. FS-1015, US Department of Defense Voice Processing Consortium.

DDVPC (**1989**), "CELP Speech Coding Standard," Tech. Rep. FS-1016, US Department of Defense Voice Processing Consortium.

de Jager, F. (**1952**), "Deltamodulation, A method of PCM transmission using the l-unit code," Philips Res. Rept. **7**, 442–466.

Denes, P. B. and Mathews, M. V. (**1960**), "Spoken digit recognition using time-frequency pattern matching," J. Acoust. Soc. Am. **32**, 1450–1455.

Dennis, J. B. (**September 1962**), "Computer control of an analog vocal tract," in *Proc. Stockholm Speech Comm. Seminar, R.I.T* (Stockholm, Sweden).

D'Eustachio, D. (**1960**), "Articulation testing in moderate sized rooms," J. Acoust. Soc. Am. **32**, 1525(A).

Dewey, G. (**1923**), *Relative frequency of English speech sounds* (Harvard University Press, Cambridge, Massachusetts).

Dimolitsas, S. (**1995**), "Evaluation of voice coded performance for the Inmarsat Mini-M system," in *Tenth Int. Conf. Digital Sat. Comm.*

Dixon, N. R. and Maxey, H. D. (**1968**), "Terminal analog synthesis of continuous speech using the diphone method of segment assembly," IEEE Trans. Audio and Electroacoust. **AU-16**, 40–50.

Doddington, G. R. (**1971**), "A method of speaker verification," J. Acoust. Soc. Am. **49**, 139 (A).

Dolansky, L. . (**1955**), "An instantaneous pitch-period indicator," J. Acoust. Soc. Am. **27**, 67–72.

Dolansky, L. . (**1960**), "Choice of base signals in speech signal analysis," IRE Trans. Audio **8**, 221–229.

Dreyfus-Graf, J. (**September 1962**), "Phonetograph und Schallwellcn-Quante1ung," in *Proc. Stockholm Speech Comm. Seminar, R.l.T* (, Stockholm, Sweden).

Dudgeon, D. E. (**1970**), "Two-mass model of the vocal cords," J. Acoust. Soc. Am. **48**, 118 (A).

Dudley, H. (**Dec. 1939**), "The Vocoder," Bell System Technical Journal , 122–126.

Dudley, H. (**1939a**), "Remaking speech," J. Acoust. Soc. Am. **11**, 169–177.

Dudley, H. (**1958**), "Phonetic pattern recognition vocoder for narrow-band speech transmission," J. Acoust. Soc. Am. **30**, 733–739.

Dudley, H. and Balashek, S. (**1958**), "Automatic recognition of phonetic patterns in speech," J. Acoust. Soc. Am. **30**, 721–732.

Dudley, H. and Jr., O. G. (**1946**), "Visible speech translators with external phosphors," J. Acoust. Soc. Am. **18**, 62–73.

Dunn, H. K. (**1950**), "The calculation of vowel resonances, and an electrical vocal tract," J. Acoust. Soc. Am. **22**, 740–753.

Dunn, H. K. (**1961**), "Methods of measuring vowel formant bandwidths," J. Acoust. Soc. Am. **33**, 1737–1746.

Dunn, H. K., Flanagan, J. L., and Gestrin, P. J. (**1962**), "Complex zeros of a triangular approximation to the glottal wave," J. Acoust. Soc. Am. **34**, 1977(A).

Dunn, H. K. and White, S. D. (**1940**), "Statistical measurements on conversational speech," J. Acoust. Soc. Am. **11**, 278–288.

E. E. David, J. (**1956**), "Naturalness and distortion in speech-processing devices," J. Acoust. Soc. Am. **28**, 586–589.

Egan, J. (**November 1944**), "Articulation testing methods, II," Tech. Rep. 38021944, OSRD, (U.S. Dept. of Commerce Report PB 22848).

Elias, P. (**1955**), "Articulation testing methods, II," IRE Trans. Information Theory **IT-1**, 16–33.

et al., M. H. (**March 1999**), "SIP: Session Initiation Protocol," IETF RFC , http://www.cs.columbia.edu/ hgs/sip/sip.html.

Fairbanks, G. (**1940**), *Voice and articulation drillbook, second ed* (Harper & Brothers, New York).

Fant, G. (**1956**), "On the predictability of formant levels and spectrum envelopes from formant frequencies," in *For Roman Jakobsen* (Mouton & Co., 's-Gravenhage).

Fant, G. (**1958**), "Modern instruments and methods for acoustic studies of speech," Acta Polytech. Scand **1**, 1–81.

Fant, G. (**1959**a), "Acoustic analysis and synthesis of speech with applications to Swedish," Ericsson Technics **15**, 3–108.

Fant, G. (**1959**b), "The acoustics of speech," in *Proc. Internat. Congr. Acoust.*

Fant, G. (**1960**), *Acoustic Theory of Speech Production* (Mouton and Co., The Hague).

Fant, G., Ishizaka, K., Lindqvist, J., and Sundberg, J. (**1972**), "Subglottal Formants," Speech Trans. Lab. Q. Prog. Stat. Rep. 1, Royal Institute of Technology, Stockholm.

Fant, G., Kruckenberg, A., and Bøavegøard, M. (**1994**), "Voice source parameters in continuous speech. Transformations of LF-parameters." in *Proc. ICSLP* (Yokohama), pp. 1451–1454.

Fant, G., Liljencrants, J., and Lin, Q. (**1986**), "A Four-Parameter Model of Glottal Flow," Speech Trans. Lab. Q. Prog. Stat. Rep. 4, Royal Institute of Technology, Stockholm.

Fant, G. and Lin, Q. (**1987**), "Glottal Source—Vocal Tract Acoustic Interaction," Speech Trans. Lab. Q. Prog. Stat. Rep. 1, Royal Institute of Technology, Stockholm.

Fant, G. and Lin, Q. (**1988**), "Comments on Glottal Flow Modelling and Analysis," Speech trans. lab. quart. prog. rep., Royal Institute of Technology, Stockholm.

Farnsworth, D. W. (**March 1940**), "High-Speed Motion Pictures of the Human Vocal Folds," Bell System Technical Journal **18**, 203–208.

Fettweis, A. (**1986**), "Wave digital filters: Theory and practice," Proc. IEEE **74**(2), 270–327, standard Reference.

Flanagan, J., Coker, C., Rabiner, L., Schafer, R., and Umeda, N. (**October 1970**), "Synthetic voices for computers," IEEE Spectrum **7**(10), 22–45.

Flanagan, J., Ishizaka, K., and Shipley, K. (**1975**), "Synthesis of speech from a dynamic model of the vocal cords and vocal tract," Bell System Technical Journal **54**, 485–506.

Flanagan, J. L. (**1955**), "Difference limen for the intensity of a vowel sound," J. Acoust. Soc. Am. **27**, 1223–1225.

Flanagan, J. L. (**1955b**), "A difference limen for vowel formant frequency," J. Acoust. Soc. Am. **27**, 613–617.

Flanagan, J. L. (**1956a**), "Automatic extraction of formant frequencies from continuous speech," J. Acoust. Soc. Am. **28**, 110–118.

Flanagan, J. L. (**1956b**), "Bandwidth and channel capacity necessary to transmit the formant information of speech," J. Acoust. Soc. Am. **28**, 592–596.

Flanagan, J. L. (**1957**a), "Difference limen for formant amplitude," J. Speech Hear. Dis. **22**, 205–212.

Flanagan, J. L. (**1957**b), "Estimates of the maximum precision necessary in quantizing certain "Dimensions" of vowel sounds," J. Acoust. Soc. Am. **29**, 533–534.

Flanagan, J. L. (**1957**c), "Note on the design of "Terminal-Analog" speech synthesizers," J. Acoust. Soc. Am. **29**, 306–310.

Flanagan, J. L. (**1958**), "Some properties of the glottal sound source," J. Speech Hear. Res. **1**, 99–116.

Flanagan, J. L. (**1960**a), "Analog measurements of sound radiation from the mouth," J. Acoust. Soc. Am. **32**, 1613–1620.

Flanagan, J. L. (**1960**b), "Resonance-vocoder and baseband complement," IRE Trans. Audio AU-S , 95–102.

Flanagan, J. L. (**1961**), "Audibility of periodic pulses and a model for the threshold," J. Acoust. Soc. Am. **33**, 1540–1549.

Flanagan, J. L. (**August 1962**a), "Computer simulation of basilar membrane displacement," in *Proc. Internat. Congr. Acoust.* (Copenhagen, Denmark).

Flanagan, J. L. (**1962**b), "Models for approximating basilar membrane displacement–part II," Bell System Tech. J. **41**, 959–1009.

Flanagan, J. L. (**September 1965**), "Recent studies in speech research at Bell Telephone Laboratories (II)," in *Proc. Internat. Congr. Acoust.* (Liege, Belgium).

Flanagan, J. L. (**1969**), "Use of an interactive laboratory computer to study an acoustic oscillator model of the vocal cords," IEEE Trans. Audio and Electroacoust. **AU-17**, 2–6.

Flanagan, J. L. (**December 1971**), "Focal points in speech communication research," IEEE Trans. Com. Tech **COM-19**, 1006–1015.

Flanagan, J. L., E. E. David, J., and Watson, B. J. (**August 1962**a), "Physiological correlates of binaural lateralization," in *Proc. Internat. Congr. Acoust.* (Copenhagen, Denmark).

Flanagan, J. L. and Guttman, N. (**1960**), "On the pitch of periodic pulses," J. Acoust. Soc. Am. **32**, 1308–1328.

Flanagan, J. L., Guttman, N., and Watson, B. J. (**1962**b), "Pitch of periodic pulses with nonuniform amplitudes," J. Acoust. Soc. Am. **34**, 738 (A).

Flanagan, J. L. and House, A. S. (**1956**), "Development and testing of a formant-coding speech compression system," J. Acoust. Soc. Am. **28**, 1099–1106.

Flanagan, J. L. and Landgraf, L. (**1968**), "Self-oscillating source for vocal-tract synthesizers," IEEE Trans. Audio and Electroacoust. **AU-16**, 57–64.

Flanagan, J. L. and Lundry, E. A. (**1967**), "Bandwidth compression of speech by analytic signal rooting," Proc. IEEE **55**, 396–401.

Flanagan, J. L. and Saslow, M. G. (**1958**), "Pitch discrimination for synthetic vowels," J. Acoust. Soc. Am. **30**, 435–442.

Fletcher, H. (**1922**), "The nature of speech and its interpretation," Bell System Technical Journal **1**, 129–144.

Fletcher, H. (**1953**), *Speech and Hearing in Communication* (van Nostrand, Princeton, NJ).

Fletcher, W. W. (**1950**), "A study of internal laryngeal activity in relation to vocal intensity," Ph.D. thesis, Northwestern Univ., Evanston, Ill.

Florencio, D. (**1993**), "Investigating the use of asymmetric windows in CELP vocoders," in *Proc. ICASSP*, vol. II, pp. 427–430.

Forgie, J. W. and Forgie, C. D. (**1962**), "Automatic method of plosive identification," J. Acoust. Soc. Am. **34**, 1979 (A).

Franke, E. K. (**April 1951**), "Mechanical impedance measurements of the human body surface," Tech. Rep. 64691951, U.S. Air Force, Wright Development Center, Wright-Patterson Air Force Base, Dayton, Ohio.

French, N. R. and Steinberg, J. C. (**1947**), "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.

Freudberg, R., Delellis, J., Howard, C., and Schaffer, H. (**November 1967**), "An all-digital pitch excited vocoder technique using the FFT algorithm," in *Proc. 1967 Conf. on Speech Communication and Processing* (Air Force Cambridge Research Labs and IEEE Audio and Electroacoustics Group).

Frick, F. C. (**1962**), "Degarble," J. Acoust. Soc. Am. **34**, 717 (A).

Fry, D. B. and Denes, P. (**1958**), "The solution of some fundamental problems in mechanical speech recognition," Language and Speech **1**, 35–58.

Fujimura, O. (**1961**), "The Nagoya group of research on speech communication," Phonetica **7**, 160–162.

Fujimura, O. (**September 1962**), "Formant-antiformant structure of nasal murmurs," in *Proc. Stockholm Speech Comm. Seminar* (Stockholm, Sweden).

Fujimura, O. and Lindquist, J. (**1971**), "Sweep-tone measurements of the vocal tract characteristics," J. Acoust. Soc. Am. **49**, 541–558.

Fujisaki, H. (**1960**), "Automatic extraction of fundamental period of speech by autocorrelation analysis and peak detection," J. Acoust. Soc. Am. **32**, 1518 (A).

Furui, S. (**1989**), *Digital Speech Processing, Synthesis, and Recognition* (Marcel Dekker, Inc., New York, NY).

Gabor, D. (**April 1952**), "Lectures on communication theory," Tech. Rep. 238, Research Laboratory of Electronics, MIT, Cambridge, Mass.

Galambos, R. (**1958**), "Neural mechanisms in audition," Laryngoscope **68**, 388–401.

Galunov, V. I. (**1966**), "Some features of speech perception," Akust. Zh, **12**, 422–427.

Gardner, W., Jacobs, P., and Lee, C. (**1993**), "QCELP: A variable rate speech coder for CDMA digital cellular," in *Speech and Audio Coding for Wireless and Network Applications*, edited by B. Atal, V. Cuperman, and A. Gersho (Kluwer Academic Press, Dordrecht, The Netherlands), pp. 85–93.

Gersho, A. and Paksoy, E. (**June 1999**), "An overview of variable rate speech coding for cellular networks," in *IEEE Int. Conf. on Selected Topics in Wireless Communications Proceedings*, pp. 172–175.

Gerson, I. and Jasiuk, M. (**1991**), "Vector sum excited linear prediction (VSELP)," in *Advances in Speech Coding*, edited by B. S. Atal, V. S. Cuperman, and A. Gersho (Kluwer, Dordrecht, The Netherlands), pp. 69–80.

Gill, J. S. (**September 1959**), “Automatic extraction of the excitation function of speech with particular reference to the use of correlation methods,” in *Proc. Internat. Congr. Acoust.* (Stuttgart, Germany).

Goeckel, D. (**1999**), “Adaptive coding for time-varying channels using outdated fading estimates,” IEEE Transactions on Communications **47**(6), 844–855.

Gold, B. (**1962**), “Computer program for pitch extraction,” J. Acoust. Soc. Am. **34**, 916–921.

Goldberg, R. and Riek, L. (**2000**), *A Practical Handbook of Speech Coders* (CRC Press, Boca Raton, FL).

Golden, R. M. (**1963**), “Digital computer simulation of a sampled-datavoice-excited vocoder,” J. Acoust. Soc. Am. **35**, 1358–1366.

Golden, R. M. (**1966**), “Phase vocoder,” Bell System Tech. J. **45**, 1493–1509.

Goldsmith, A. and Chua, S. G. (**1997**), “Variable-rate variable power MQAM for fading channels,” IEEE Transactions on Communications , 1218–1230.

Goodman, D. J. (**1969**), “The application of delta modulation to analog-to-PCM encoding,” Bell System Tech. J. **48**(2), 321–343.

Gopinath, B. and Sondhi, M. M. (**1970**), “Determination of the shape of the human vocal tract from acoustical measurements,” Bell System Tech. J. **49**(6), 1195–1214.

Gottesman, O. and Gersho, A. (**1999**), “Enhanced Waveform Interpolative Coding at 4 kbps,” in *Proc. ICASSP* (Phoenix, AZ).

Gould, G. T. (**1951**), “Design of a speech stretcher, FM-TV,” J. Rad, Comm. **11**, 30–36.

Gould, K., Cox, R., Jayant, N., and Melchner, M. (**1993**), “Robust Speech Coding for the Indoor Wireless Channel,” ATT Technical Journal , 64–73.

Greefkes, J. A. (**1957**), ““Frena,” a system of speech transmission at high noise levels,” Philips Tech. Rev. **19**, 73–108.

Greefkes, J. A. and de Jager, F. (**1968**), “Continuous delta modulation,” Philips Res. Rept. **23**, 233–246.

Griffin, D. and Lim, J. (**1988**), “Multi-band excitation vocoder,” IEEE Transactions on Acoustics, Speech, and Signal Processing **36**(8), 1223–1235.

Grutzmacher, M. and Lottermoser, W. (**1937**), “Door ein Verfahren zur tragheitsfreien Aufzeichnung von Melodiekurven,” Akust. Z. **2**, 242–248.

(GSM), S. M. G. (**1997**), “Digital cellular telecommunications system: Enhanced Full Rate (EFR) speech transcoding,” Tech. Rep. GSM 06.60, European Telecommunications Standards Institute (ETSI).

(GSM), S. M. G. (**1998**), “Digital cellular telecommunications system (Phase 2+): Adaptive Multi-Rate (AMR) speech transcoding,” Tech. Rep. GSM 06.90, European Telecommunications Standards Institute (ETSI).

Guild, S. R., Crowe, S. J., Bunch, C. C., and Polvogt, L. M. (**1931**), “Correlations of differences in the density of innervation of the organ of Corti with differences in the acuity of hearing,” Acta Oto-Laryngol. **15**, 269–308.

Guinan, J. J. and Peake, W. T. (**1967**), “Middle-ear characteristics of anesthetized cats,” J. Acoust. Soc. Am. **41**, 1237–1261.

Guttman, N. and Flanagan, J. L. (**1962**), “Pitch of nonuniformly spaced pulses in periodic trains,” J. Acoust. Soc. Am. **34**, 1994 (A).

Guttman, N. and Flanagan, J. L. (**1964**), “Pitch of high-pass filtered periodic pulses,” J. Acoust. Soc. Am. **36**, 757–765.

Guttman, N. and Nelson, J. R. (**1968**), “An instrument that creates some artificial speech spectra for the severely hard of hearing,” Am. Ann. Deaf. **113**, 295–302.

Hagenauer, J. (**1988**), “Rate-compatible punctured convolutional codes and their applications,” IEEE Tr. Comm. **36**(4), 389–400.

Halle, M., Hughes, G. W., and Radley, J.-P. A. (**Jan. 1957**), “Acoustic Properties of Stop Consonants,” J. Acoust. Soc. Am. **29**(1), 107–116.

Halsey, R. J. and Swaffield, J. (**1948**), “Analysis-synthesis telephony, with special reference to the vocoder,” Inst. Elec. Engrs **95**, 391–411 pt. III.

Hanauer, S. L. and Schroeder, M. R. (**1966**), “Non-linear time compression and time normalization of speech,” J. Acoust. Soc. Am. **40**, 1243 (A).

Hardwick, J. and Lim, J. (**1988**), “A 4.8 kbps multi-band excitation speech coder,” in *Proc. ICASSP*, vol. 1, pp. 374–377.

Harris, K. S., Hoffman, H. S., and Griffith, B. C. (**1957**), “The discrimination of speech sounds within and across phoneme boundaries,” J. Expt, Psychol. **54**, 358–368.

Hasegawa-Johnson, M. (**2000**), “Line spectral frequencies are the poles and zeros of a discrete matched-impedance vocal tract model,” J. Acoust. Soc. Am. **108**(1), 457–460.

Haskew, J. R. (**June 1969**), "A comparison between linear prediction and linear interpolation," Master's thesis, Brooklyn Polytechnic Institute, New York, New York.

Hecker, M. H. L. (**1962**), "Studies of nasal consonants with an articulatory Spech synthesizer," J. Acoust. Soc. Am. **34**, 179–188.

Heinz, J. and Stevens, K. (**1964**), "On the derivation of area functions and acoustic spectra from cineradiographic films of speech," J. Acoust. Soc. Am. **36**, 1037.

Heinz, J. M. (**July 1958**), "Model studies of the production of fricative consonants," Quart. Progr. Rept. .

Heinz, J. M. (**August 1962**), "Reduction of speech spectra to descriptions in terms of vocal tract area functions," Ph.D. thesis, Mass. Inst. of Tech.

Heinz, J. M. and Stevens, K. N. (**1961**), "On the properties of voiceless fricative consonants," J. Acoust. Soc. Am. **33**, 589–596.

Hellwig, K., Vary, P., Massaloux, D., Petit, J. P., Galand, C., and Rosso, M. (**1989**), "Speech codec for the European mobile radio system," in *IEEE Global Telecomm. Conf.*

Hermansky, H. (**1990**), "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am. **87**(4), 1738–1752.

Hermansky, H. and Morgan, N. (**1994**), "RASTA Processing of Speech," IEEE Trans. Speech Audio Proc. **2**(4), 587–589.

Hersent, O., Gurle, D., and Petit, J.-P. (**2000**), *IP Telephony* (Addison Wesley).

Hildebrand, F. B. (**1948**), *Advanced calculus for engineers* (Prentice-Hall Inc., New York).

Hildebrand, F. B. (**1952**), *Methods of applied mathematics* (Prentice-Hall, Inc., New York).

Holes, J. N. and Kelly, L. C. (**January 1960**), "Apparatus for segmenting the formant frequency regions of a speech signal," Tech. Rep. 20566, British Post Office Research Station, Dollis Hill, London.

Holmes, J. N. (**December 1958**), "A method of tracking formants which remains effective in the frequency regions common to two formants," Tech. Rep. JU 8-2, Joint Speech Res. Unit, British Post Office, Eastcote, England.

Holmes, J. N. (**July 1961**), "Research on speech synthesis," Tech. Rep. JU 11-4, Joint Speech Res. Unit, British Post Office, Eastcote, England.

Holmes, J. N. (**September 1962**), "An investigation of the volume velocity waveform at the larynx during speech by means of an inverse filter," in *Stockholm Speech Comm. Seminar, R.I.T* (Stockholm, Sweden).

Houde, R. A. (**1967**), "A study of tongue body motion during selected speech sounds," Ph.D. thesis, Univ. of Michigan,.

House, A. S. (**1955**), "Development of a quantitative description of vowel articulation," J. Acoust. Soc. Am. **27**, 484–493.

House, A. S. (**1956**), "Studies of formant transitions using a vocal tract analog," J. Acoust. Soc. Am. **28**, 578–585.

House, A. S. (**1957**), "Analog studies of nasal consonants," J. Speech Hear. Disorders **22**, 190–204.

House, A. S. and Stevens, K. N. (**1958**), "The Estimation of Formant Bandwidths from Measurements of the Transient Response of the Vocal Tract," J. Speech Hear. Res. **1**, 309–315.

House, A. S., Stevens, K. N., Sandel, T. T., and Arnold, J. B. (**1962**), "On the learning of speechlike vocabularies," J. Verbal Learn. and Verbal Behavior **1**, 133–143.

Howard, C. R. (**1956**), "Speech analysis-synthesis schemes using continuous parameters," J. Acoust. Soc. Am. **28**, 1091–1098.

Howell, A. S., Schneider, G. O. K., and Stump, T. M. (**1961**), "Analog multiplexing of a telephone semi-vocoder," J. Acoust. Soc. Am. **33**, 1663 (A).

Huang, J. and Levinson, S. (**1999**), "Estimation of articulatory movement and its application to speech synthesis," J. Acoust. Soc. Am. **106**, 2180.

Huggins, W. H. (**1954**), "A note on autocorrelation analysis of speech sounds," J. Acoust. Soc. Am. **26**, 790–792.

Huggins, W. H. (**September 1957**), "Representation and analysis of signals, part I; the use of orthogonalized exponentials," Tech. Rep. AF 19 (604)-1941, ASTIA No AD 133741, Johns Hopkins University.

Hughes, G. W. (**1958**), "A real-time input system for a digital computer," J. Acoust. Soc.Am. **30**, 668 (A).

Hughes, G. W. (**May 1961**), "The recognition of speech by machine," Tech. Rep. 395, Res. Lab. Elect., Mass. Inst. Tech., Cambridge, Mass.

Ingard, U. (**1953**), "On the theory and design of acoustic resonators," J. Acoustic. Soc. Am. **25**, 1037–1061.

Inomata, S. (**1960**), "A new method of pitch extraction using a digital computer," J. Acoust. Soc. Japan 16 (4) , 283–285.

(IPA), I. P. A. (**1993**), "International Phonetic Alphabet," .

Ishizaka, K. and Flanagan, J. L. (**1972**a), "Acoustic properties of a two-mass model of the vocal cords," J. Acoust. Soc. Am. **51**, 91 (A).

Ishizaka, K. and Flanagan, J. L. (**July-Aug. 1972**b), "Synthesis of voiced sounds from a two-mass model of the vocal cords," Bell System Technical Journal **51**(6), 1233–1268.

Ishizaka, K. and Matsudaira, M. (**August 1968**), "What makes the vocal cords vibrate," in *Proc. Int. Congr. Acoust.* (Tokyo, Japan), vol. B-1-3.

Ishizaka, K., Matsudaira, M., and Kaneko, T. (**July 1976**), "Input acoustic-impedance measurement of the subglottal system," J. Acoust. Soc. Am. **60**(1), 190–197.

ISO/IEC (**1998**a), "Information Technology—Coding of Audiovisual Objects, Part 3: Audio, Subpart 1: Overview," Tech. Rep. ISO/JTC 1/SC 29/N2203, ISO/IEC.

ISO/IEC (**1998**b), "Information Technology—Coding of Audiovisual Objects, Part 3: Audio, Subpart 3: CELP," Tech. Rep. ISO/JTC 1/SC 29/N2203CELP, ISO/IEC.

ISO/IEC (**1998**c), "Information Technology—Coding of Audiovisual Objects, Part 3: Audio, Subpart 4: Time/Frequency Coding," Tech. Rep. ISO/JTC 1/SC 29/N2203TF, ISO/IEC.

ISO/IEC (**1998**d), "Information Technology—Very Low Bitrate Audio-Visual Coding, Part 3: Audio, Subpart 2: Parametric Coding," Tech. Rep. ISO/JTC 1/SC 29/N2203PAR, ISO/IEC.

ISO/IEC (**Oct. 1998**e), "Report on the MPEG-4 speech codec verification tests," Tech. Rep. JTC1/SC29/WG11, ISO/IEC.

Itakura, F. and Saito, S. (**August 1968**), "An analysis-synthesis telephony based on maximum likelihood method," in *Proc. Int. Congr. Acoust.* (Tokyo, Japan), vol. C-5-5.

Itakura, F. and Saito, S. (**1970**), "A statistical method for estimation of speech spectral density and formant frequencies," Electronics and Communications in Japan 53A , 36–43.

Ito, H., Serizawa, M., Ozawa, K., and Nomura, T. (**1998**), "An adaptive multi-rate speech codec based on MP-CELP coding algorithm for ETSI AMR standard," in *Proc. ICASSP*, vol. 1, pp. 137–140.

ITU-T (**1990**a), "40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)," Tech. Rep. G.726, International Telecommunications Union, Geneva.

ITU-T (**1990**b), "5-, 4-, 3- and 2-bits per Sample Embedded Adaptive Differential Pulse Code Modulation (ADPCM)," Tech. Rep. G.727, International Telecommunications Union, Geneva.

ITU-T (**1992**), "Coding of Speech at 16 kbit/s Using Low-Delay Code Excited Linear Prediction," Tech. Rep. G.728, International Telecommunications Union, Geneva.

ITU-T (**1993**a), "7 kHz Audio Coding Within 64 kbit/s," Tech. Rep. G.722, International Telecommunications Union, Geneva.

ITU-T (**1993**b), "Pulse Code Modulation (PCM) of Voice Frequencies," Tech. Rep. G.711, International Telecommunications Union, Geneva.

ITU-T (**1996**a), "Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)," Tech. Rep. G.729, International Telecommunications Union, Geneva.

ITU-T (**1996**b), "Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s," Tech. Rep. G.723.1, International Telecommunications Union, Geneva.

ITU-T (**1998**a), "Objective quality measurement of telephone-band (300-3400 Hz) speech codecs," Tech. Rep. P.861, International Telecommunications Union, Geneva.

ITU-T (**1998**b), "Packet Based Multimedia Communications Systems," Tech. Rep. H.323, International Telecommunications Union, Geneva.

J. C. Hammett, J. (**1971**), "An adaptive spectrum analysis vocoder," Ph.D. thesis, Dept. Elec. Eng., Georgia Inst. Tech., Atlanta, Ga.

Jarvinen, K., Vainio, J., Kapanen, P., Honkanen, T., Haavisto, P., Salami, R., LaFlamme, C., and Adoul, J.-P. (**1997**), "GSM enhanced full rate speech codec," in *Proc. ICASSP*, pp. 771–774.

Jayant, N., Johnston, J., and Safranek, R. (**1993**), "Signal Compression Based on Models of Human Perception," Proceedings of the IEEE **81**(10), 1385–1421.

Jayant, N. S. (**1970**), "Adaptive delta modulation with a one-bit memory," Bell System Tech. J. **49**, 321–342.

Johnson, M. and Taniguchi, T. (**1991**), "Low-Complexity Multi-Mode VXC Using Multi-Stage Optimization and Mode Selection," in *Proc. ICASSP*, pp. 221–224.

Johnstone, B. M. and Boyle, A. J. F. (**1967**), "Basilar membrane vibration examined with the Mossbauer technique," Science **158**, 389–390.

Jr., J. C., Tremain, T., and Welch, V. (**1991**), "The DOD 4.8 KBPS Standard (Proposed Federal Standard 1016)," in *Advances in Speech Coding*, edited by B. S. Atal, V. C. Cuperman, and A. Gersho (Kluwer, Dordrecht, The Netherlands), pp. 121–133.

Jr., J. C., Welch, V., and Tremain, T. (**1989**), "An Expandable Error-Protected 4800 BPS CELP Coder (U.S. Federal Standard 4800 BPS Voice Coder)," in *Proc. ICASSP*, pp. 735–738.

Jr., J. L. K. and Gerstman, L. J. (**1961**), "An artificial talker driven from a phonetic input," J. Acoust, Soc. Am. **33**, 835 (A).

Jr., J. L. K. and Lochbaum, C. (**September 1962**a), "Speech synthesis," in *Proc. Stockholm Speech Comm. Seminar, R.I.T* (, Stockholm, Sweden).

Jr., J. L. K. and Lochbaum, C. C. (**1962**b), "Speech Synthesis," in *Proc. Internat. Congr. Acoust.*, pp. G42:1–4.

Jr., O. G. and Schott, L. . (**1949**), "Extraction and portrayal of pitch of speech sounds," J. Acoust. Soc. Am. **21**, 487–495.

Judson, L. S. and Weaver, A. T. (**1942**), *Voice science* (F. S. Crofts & Co., New York).

Kabal, P. and Ramachandran, R. (**1986**), "The computation of line spectral frequencies using Chebyshev polynomials," IEEE Trans. Acoust., Speech, Signal Processing **ASSP-34**, 1419–1426.

Katsuki, Y. (**1960**), "Neural mechanism of hearing in cats and insects," in *Electrical activity of single cells* (Igakushoin, Hongo, Tokyo).

Kautz, W. H. (**1954**), "Transient synthesis in the time domain," IRE Trans. Circuit Theory **CT-1**, 29–39.

Keefe, D. H. (**1984**), "Acoustical wave propagation in cylindrical ducts: Transmission line parameter approximations for isothermal and nonisothermal boundary conditions," J. Acoust. Soc. Am. **75**, 1386–1391.

Kelly, J. M. and Miller, R. L. (**November 1967**), "Recent improvements in 4800 BPS voice-excited vocoders," in *Proc. Conf. on Speech Communication and Processing* (A.F. Cambridge Res. Labs. and IEEE Group on Audio and Electroacoust.).

Kempelen, W. V. (**1791**), *Le mechanisme de la parole, suivi de la Description d'une machine parlante* (J. V. Degen, Vienna).

Kersta, L. G. (**1948**), "Amplitude cross-section representation with the sound spectrograph," J. Acoust. Soc. Am. **20**, 796–801.

Kersta, L. G. (**1962a**), "Voiceprint identification," Nature **196**, 1253–1257.

Kiang, N. Y. S. and Peake, W. T. (**1960**), "Components of electrical responses recorded from the cochlea," Ann. Otol. Rhinol. Laryngol, **69**, 448–458.

Kinsler, L. F. and Frey, A. R. (**1962**), *Fundamentals of Acoustics, 2d edition* (John Wiley & Sons, New York).

Kleijn, W. (**1991**), "Speech coding below 4 kb/s using waveform interpolation," in *Proc. GLOBECOM*, vol. 3, pp. 1879–1883.

Kleijn, W. and Granzow, W. (**1991**), "Methods for waveform interpolation in speech coding," Digital Signal Processing , 215–230.

Kleijn, W. and Haagen, J. (**1995**), "A speech coder based on decomposition of characteristic waveforms," in *Proc. ICASSP*, pp. 508–511.

Kleijn, W., Shoham, Y., Sen, D., and Hagen, R. (**1996**), "A low-complexity waveform interpolation coder," in *Proc. ICASSP*, pp. 212–215.

Kleijn, W. B., Krasinski, D. J., and Ketchum, R. H. (**1988**), "Improved speech quality and efficient vector quantization in SELP," in *Proc. ICASSP*, pp. 155–158.

Kock, W. E. (**1956**), "Speech bandwidth compression," Bell Lab. Record **34**, 81–85.

Kock, W. E. (**June 1959**), "Narrowband transmission of speech," Tech. Rep. 2,890,285, U.S. Patent.

Kock, W. E. (**1962**), "Speech communication systems," Proc, I.R.E. **50**, 769–776.

Koenig, R. (**1946**), "The sound spectrograph," J. Acoust. Soc. Am. **18**, 19–49.

Kohler, M. (**1997**), "A comparison of the new 2400bps MELP Federal Standard with other standard coders," in *Proc. ICASSP*, pp. 1587–1590.

Kraft, L. G. (**1950**), "Correlation function analysis," J. Acoust. Soc. Am. **22**, 762–764.

Kramer, H. P. and Mathews, M. V. (**1956**), "A linear coding for transmitting a set of correlated signals," IRE Trans. Inform. Theory **IT-2**, 41–46.

Kringlebotn, M. (**1968**), "Experiments with some vibrotactile and visual aids for the deaf," Proc. Conf. on Speech-Analyzing Aids for the Deaf, Amer. Ann. Deaf **113**, 311–317.

Kroon, P., Deprettere, E. F., and Sluyter, R. J. (**1986**), "Regular-pulse excitation: A novel approach to effective and efficient multi-pulse coding of speech," IEEE Trans. ASSP **34**, 1054–1063.

Kruskal, J. (**1964**), "Nonmetric multidimensional scaling," Psychometrika **29**, 115–129.

Kryter, K. D. (**1962**), "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. **34**, 1689–1697.

Kulya, V. I. (**1962**a), "Application of Laguerre Functions to parametric coding of speech signals," Elektrosvyaz **7**, 33–39.

Kulya, V. I. (**1962**b), "Application of Laguerre Functions to parametric coding of speech signals," Telecommunications and Radio Engineering, part I. Telecommunications **7**, 34–41.

Kulya, V. I. (**March 1963**), "Analysis of a Chebyshev-type vocoder," Telecomm. and Radio Engng., Part 1 **3**, 23–32.

L. A. Yaggi, J. (**June 1962**), "Full-duplex digital vocoder," Tech. Rep. SP 14-A62, Texas Inst. Inc, Dallas.

L. A. Yaggi, J. and Mason, A. E. (**1963**), "Polymodal vocoder; a new approach to versatile and reliable voice communications," J. Acoust, Soc. Am. **35**, 806 (A).

Ladefoged, P. (**Nov. 1958**), "The perception of speech," in *Proc. Symp. on Mechanization of Thought Processes* (National Physical Laboratory Teddington, England).

Ladefoged, P. and Broadbent, D. E. (**1957**), "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 98–104.

Lawrence, W. (**1953**), "The synthesis of speech from signals which have a low information rate," in *Communication theory*, edited by W. Jackson (Butterworths Sci. Pub., London).

Leach, W. (**1996**), "A 2-port analogous circuit and SPICE model for Salmon's family of acoustic horns," J. Acoust. Soc. Am. **99**(3), 1459–1464, mar.

LeBlanc, W., Bhattacharya, B., Mahmoud, S., and Cuperman, V. (**1993**), "Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4kb/s speech coding," IEEE Trans. Speech and Audio Processing **1**(4), 373–385.

Lee, F. F. (**1969**), “Reading machine: From text to speech,” IEEE Trans. Audio and Electroacoust. **AU-17**, 275–282.

Lee, Y. W. (**1960**), *Statistical theory of communication* (John Wiley & Sons, New York).

Lehiste, I. and Peterson, G. E. (**1961**), “Transitions, Glides, and Diphthongs,” J. Acoust. Soc. Am. **33**(3), 268–277.

Levelt, W. J. M. (**1989**), *Speaking: from Intention to Articulation* (MIT Press, Cambridge, MA).

Levitt, H. and Nelson, J. R. (**1970**), “Experimental communication aids for the deaf,” IEEE Trans. Audio and Electroacoust. **AU-18**, 2–6.

Li, F. and Allen, J. B. (**2008**), “Additivity law of frequency integration for consonant identification in white noise,” J. Acoust. Soc. Am. Submitted Jul 30, 2008.

Liberman, A. M. and Borst, J. M. (**1951**), “The inter-conversion of audible and visible patterns as a basis for research in the perception of speech,” Proc. Nat. Acad. Sci. U.S. **37**, 318–325.

Liberman, A. M., Cooper, F. S., Harris, K. S., and Macneilage, P. F. (**September 1962**), “A motor theory of speech perception,” in *Proc. Stockholm Speech Comm. Seminar, R.I.T* (Stockholm, Sweden).

Liberman, A. M., Delattre, P. D., Cooper, F. S., and Gerstman, L. (**1954**), “The role of consonant-vowel transitions in the stop and nasal consonants,” Psychol. Monographs **68**.

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (**1957**), “The discrimination of speech sounds within and across phoneme boundaries,” J. Exp. Psychol. **54**, 358–368.

Licklider, J. C. R., Stevens, K. N., and Hayes, J. R. M. (**September 1954**), “Studies in speech, hearing and communication. Final Report,” Tech. Rep. W-19122ac-1430, Acoustics Lab. Mass Inst. of Tech., Cambridge, Mass.

Lieberman, P. (**1961**), “Perturbations in vocal pitch,” J. Acoust. Soc. Am. **33**, 597–603.

Lighthill, S. M. J. (**1978**), *Waves in fluids* (Cambridge University Press, England).

Lin, D. (**1986**), “New approaches to stochastic coding of speech sources at very low bit rates,” in *Signal Processing III: Theories and Applications*, edited by I. T. Y. et al. (Elsevier, Amsterdam), pp. 445–447.

Lin, Q. (**1990**), "Speech Production Theory and Articulatory Speech Synthesis," Ph.D. thesis, Royal Institute of Technology (KTH), Stockholm.

Lin, W.-C., Liang, C.-C., and Chen, C.-T. (**Sep. 1988**), "Dynamic Elastic Interpolation for 3-D Medical Image Reconstruction from Serial Cross Sections," IEEE Trans. on Medical Imaging **7**(3), 225–232.

Lindgren, N. (**March 1965**a), "Automatic speech recognition, part (I)," IEEE Spectrum **2**, 114–136.

Lindgren, N. (**April 1965**b), "Automatic speech recognition, part (II)," IEEE Spectrum **2**, 45–59.

Lindgren, N. (**May 1965**c), "Automatic speech recognition, part (III)," IEEE Spectrum **2**, 104–116.

Linvill, J. (**1969**), "Development progress on a microelectronic tactile facsimile reading aid for the blind," IEEE Trans. Audio and Electroacoust. **AU-17**, 271–274.

Lummis, R. C. (**1971**), "Real time technique for speaker verification by computer," J. Acoust. Soc. Am. **50**, 106 (A).

Malecot, A. (**1956**), "Acoustic cues for nasal consonants," Language **32**, 274–284.

Malme, C. I. (**January 1959**), "Detectability of small irregularities in a broadband noise spectrum," Quarterly Rept. .

Manley, H. J. (**1962**), "Fourier coefficients of speech power spectra as measured by Autocorrelation Analysis," J. Acoust. Soc. Am. **34**, 1143–1145.

Marcou, P. and Daguet, J. (**1956**a), "New methods of speech transmission," Ann. Telecommun. **11**, 118–126.

Marcou, P. and Daguet, J. (**1956**b), "New methods of speech transmission," in *Information Theory: Proc. of 3rd Symp. on Info. Theory, London*, edited by C. Cherry (Butterworths Sci. Pub., London), pp. 231–244.

Markel, J. D. (**1971**), "The Prony method and its application to speech analysis," J. Acoust. Soc. Am. **49**, 105 (A).

Markel, J. D. (**June 1972**), "Digital inverse filtering–a new tool for formant trajectory estimation," IEEE Trans. Audio and ElectroAcoust. **AU-20**.

Martin, T. B., Nelson, A. L., and Zadell, A. J. (**1964**), "Speech recognition by feature abstraction techniques," Tech. Rep. AL-TDR, Wright-Patterson AFB, Avionics Labs.

Massey, N. S. (**May 1994**), "Transients at Stop-Consonant Releases," Master's thesis, MIT, Cambridge, MA.

Mathews, M. V. (**1959**), "External coding for speech transmission," IRE Trans. Inform. Theory **IT-5**, 129–136.

Mathews, M. V. and Walker, P. (**1962**), "Program to compute vocal-tract poles and zeros," J. Acoust. Soc. Am. **34**, 1977 (A).

Mattingly, I. G. and Shearme, J. N. (**1964**), "Speech synthesis by rule," Language and Speech **7**, 127–143.

McCree, A. and Barnwell III, T. (**1995**), "A mixed excitation LPC vocoder model for low bit rate speech coding," IEEE Trans. Speech Audio Processing **3**(4), 242–50.

McCree, A. and Martin, J. D. (**1998**), "A 1.7 kb/s MELP coder with improved analysis and quantization," in *Proc. ICASSP*, vol. 2, pp. 593–596.

McCree, A., Truong, K., George, B., Barnwell, T., and Viswanathan, V. (**1996**), "A 2.4 kbps MELP coder candidate for the new U.S. Federal Standard," in *Proc. ICASSP*, vol. 1, pp. 200–203.

McDonald, R. (**1966**), "Signal-to-noise and idle channel performance of differential pulse code modulation sytems–particular application to voice signals," Bell System Tech. J. **45**, 1123–1151.

Mermelstein, P. (**1967**), "Determination of the Vocal-Tract Shape from Measured Formant Frequencies," J. Acoust. Soc. Am. **41**(5), 1283–1294.

Mermelstein, P. (**1969**), "Computer simulation of articulatory activity in speech production," in *Proc. Int. Joint Conf. on Artificial Intelligence* (Washington, D.C.).

Meyer-Eppler, W. (**1953**), "Zum Erzeugungsmeehanismus der Gerauschlaute," Z. Phonetik **7**, 196–212.

Miller, G. A. (**1947**), "Sensitivity to changes in the intensity of white noise and its relation to masking and loudness," J. Acoust. Soc. Am. **19**, 609–619.

Miller, G. A. (**1956**), "The magical number seven, plus or minus two: Some limits in our capacity for processing information," Psychol. Rev. **63**, 81–97.

Miller, G. A. (**1962**), "Decision units in the perception of speech," I.R.E. Trans. Inform. Theory **IT-8**, 81–83.

Miller, G. A., Heise, G. A., and Lichten, W. (**1951**), "The Intelligibility of Speech as a Function of the Context of the Test Materials," Journal of Experimental Psychology **41**, 329–335.

Miller, G. A. and Isard, S. (**1963**), “Some perceptual consequences of linguistic rules,” Jol. of Verbal Learning and Verbal Behavior **2**, 217–228.

Miller, G. A. and Nicely, P. E. (**1955**), “Analysis of Perceptual Confusions Among Some English Consonants,” J. Acoust. Soc. Am. **27**, 338–352.

Miller, R. L. (**1953**), “Improvements in the vocoder,” J. Acoust. Soc. Am. **25**, 832 (A).

Miller, R. L. (**1959**), “Nature of the vocal cord wave,” J. Acoust. Soc. Am. **31**, 667–677.

Moncur, J. P. and Dirks, D. (**1967**), “Binaural and monaural speech intelligibility in reverberation,” J. Speech Hear. Res. **10**, 186–195.

Moore, B. C. (**1997**), *An Introduction to the Psychology of Hearing* (Academic Press, San Diego, CA).

Morse, P. M. (**1948**), *Vibration and sound* (McGraw-Hill Book Co., New York).

Müller, A. R. (**1961**), “Network model of the middle ear,” J. Acoust. Soc. Am. **33**, 168–176.

Müller, A. R. (**August 1962**), “On the transmission characteristic of the middle ear,” in *Proc, IV Int. Congr. Acoust* (Copenhagen, Denmark).

Munson, W. A. and Montgomery, H. C. (**1950**), “A speech analyzer and synthesizer,” J. Acoust. Soc. Am. **22**, 678 (A).

Nakata, K. (**1960**), “Synthesis and perception of Japanese fricative sounds,” J. Radio Res. Lab. **7**, 319–333.

Nakata, K. (**1961**), “Recognition of Japanese vowels,” J. Radio Res. Lab. **8**, 193–212.

Nakata, K. and Suzuki, J. (**1959**), “Synthesis and perception of Japanese vowels and vowel-like sounds,” J. Radio Res. Lab. **6**, 617–634.

Nakatani, L. H. (**1971**), “Measuring the ease of comprehending speech,” Proc. 7th Int. Congr. Acoust. .

Netter, F. (**1962**), “Anatomical drawings of the ear,” Clinical Symposia **14**, 39–73.

Nevelskü, P. B. (**1966**), “Comparative study of the volume of the short-term and long-term memory,” Proc. 18th Inter. Psychol. Congr. Symp. , 21–26.

Noll, A. M. (**1967**), “Cepstrum pitch determination,” J. Acoust. Soc. Am. **41**, 293–309.

Noll, P. (**1997**), “MPEG Digital Audio Coding,” IEEE Signal Processing Magazine , 59–81.

Novorita, B. (**1999**), “Incorporation of temporal masking effects into bark spectral distortion measure,” in *Proc. ICASSP* (Phoenix, AZ), pp. 665–668.

Ochiai, Y. (**1958**), “Fondamentales des qualites phonemique et vocalique des paroles par rapport au timbre, obtenues en ernployant des voyelles japonais vocalisees par des sinets japonais,” Mem. Fac. Eng., Nagoya Univ. **10**, 197–201.

Ochiai, Y. and Kato, H. (**1949**), “Sur la nettete et la naturalité de la voix humaine reflechies du point de vue de la qualite de transmission,” Mem. Fac. Eng., Nagoya Univ. **1**, 105–115.

Oetinger, R. and Hauser, H. (**1961**), “An electrical network for the investigation of the mechanical vibrations of the inner ear,” Acustica **11**(3), 161–177.

Olive, J. P. (**1971**), “Automatic formant tracking by a Newton-Raphson technique,” J. Acoust. Soc. Am. **50**, 661–670.

Olson, H. F. and Belar, H. (**1961**), “Phonetic typewriter, III,” J. Acoust. Soc. Am. **33**, 1610–1615.

O’Neil, E. F. (**1959**), “TASI,” Bell Lab Record **37**, 83–87.

Oppenheim, A., Schafer, R., and Stockham, T. G. (**1968**), “Nonlinear filtering of multiplied and convolved signals,” Proc. IEEE **56**, 1264–1291.

Oppenheim, A. V. (**1969**), “Speech analysis-synthesis system based on homomorphic filtering,” J. Acount. Soc. Am. **45**, 459–462.

Oppenheim, A. V. (**1971**), “Predictive coding in a homomorphic vocoder,” IEEE Trans. Aud. Electroacoust. **AV-19**, 243–248.

Paget, S. R. (**1930**), *Human speech* (Harcourt, London and New York).

Paksoy, E., Chan, W.-Y., and Gersho, A. (**1992**), “Vector quantization of speech LSF parameters with generalized product codes,” in *Proc. ICASSP*, pp. 33–36.

Paksoy, E., de Martin, J. C., McCree, A., Gerlach, C., Anandakumar, A., Lai, M., and Viswanathan, V. (**1999**), “An adaptive multi-rate speech coder for digital cellular telephony,” in *Proc. of ICASSP*, vol. 1, pp. 193–196.

Paliwal, K. K. and Atal, B. S. (**1993**), “Efficient vector quantization of LPC parameters at 24 bits/frame,” IEEE Trans. Speech Audio Processing **1**, 3–14.

Papoulis, A. (**1965**), *Probability, Random Variables, and Stochastic Processes* (McGraw–Hill Book Co., New York).

Peake, W. T., M. H. Goldstein, J., and Kiang, N.-S. (**1962**), "Responses of the auditory nerve to repetitive acoustic stimuli," J. Acoust. Soc. Am. **34**, 562–570.

Pelorson, X., Hirschberg, A., van Hassel, R. R., and Wijnands, A. P. J. (**Dec. 1994**), "Theoretical and experimental study of quasisteady-flow separation within the glottis during phonation. Application to a modified two-mass model," J. Acoust. Soc. Am. **96**(6), 3416–3431.

Perkell, J. S. (**September 1965**), "Cineradiographic studies of speech: Implications of certain articulatory movements," in *Proc. 5th Int. Congr. Acoust.* (Liege, Belgium).

Perlman, A. L. (**1985**), "A technique for measuring the elastic properties of vocal fold tissue," Ph.D. thesis, University of Iowa, Iowa City, Iowa.

Perlman, A. L. and Titze, I. R. (**1988**), "Development of an *in vitro* technique for measuring elastic properties of vocal fold tissue," J. Speech Hear. Res. **31**, 288–298.

Perlman, A. L., Titze, I. R., and Cooper, D. S. (**1984**), "Elasticity of Canine Vocal Fold Tissue," J. Speech Hear. Res. **27**, 212–219.

Peterson, E. (**1951**), "Frequency detection and speech formants," J. Acoust. Soc. Am. **23**, 668–674.

Peterson, G. E. and Barney, H. L. (**March 1952**), "Control Methods Used in a Study of Vowels," J. Acoust. Soc. Am. **24**(2), 175–184.

Peterson, L. C. and Bogert, B. P. (**1950**), "A dynamical theory of the cochlea," J. Acoust. Soc. Am. **22**, 369–381.

Pickett, J. M. (**1969**), "Some applications of speech analysis to communication aids for the deaf," IEEE Trans. Audio and Electroacoust. **AU-17**, 283–289.

Pierce, A. D. (**1981**), *Acoustics: An Introduction to its Physical Principles and Applications* (McGraw-Hill, New York).

Pierce, J. R. (**1969**), "Whither speech recognition," J. Acoust. Soc. Am. **46**, 1049–1051(L).

Pierce, J. R. and Karlin, J. E. (**1957**), "Information rate of a human channel," Proc. I.R.E. **45**, 368.

Pimonow, L. (**September 1962**), "Coded speech and its application in aids for the deaf," in *Proc. Stockholm Speech Comm. Seminar, R.I.T* (Stockholm, Sweden).

Pirogov, A. A. (**1959**a), "A harmonic system for compressing speech-spectra," Telecommunications **3**, 229–242.

Pirogov, A. A. (**1959**b), “A harmonic system for compressing speech-spectra,” Elektrosviaz No. **3**, 8–17.

Pollack, I. (**1952**), “The information of elementary auditory displays,” J. Acoust. Soc. Am. **24**, 745–749.

Pollack, I. and Ficks, L. (**1954**), “Information of elementary multidimensional auditory displays,” J. Acoust. Soc. Am. **26**, 155–158.

Potter, R. K., Korr, G. A., and c. Green, H. (**1947**), *Visible speech* (D. van Nostrand Co., New York).

Potter, R. K. and Steinberg, J. C. (**1950**), “Toward the specification of speech,” J. Acoust. Soc. Am. **22**, 807–820.

Prestigiacomo, A. J. (**1957**), “Plastic tape sound spectrograph,” J. Speech Hear. Disorders **22**, 321–327.

Pruzansky, S. (**1963**), “Pattern-matching procedure for automatic talker recognition,” J. Acoust. Soc. Am. **35**, 354–358.

Purves, B., Blackett, K., and Strong, W. (**1970**), “Speech synthesis with a vocal tract synthesizer,” J. Acoust. Soc. Am. **47**, 93(A).

Rabiner, L. and Juang, B.-H. (**1993**), *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ).

Rabiner, L. and Schafer, R. (**1978**), *Digital Processing of Speech Signals* (Prentice-Hall Inc., New Jersey).

Rabiner, L. R. (**1968**a), “Digital-formant synthesizer for speech synthesis studies,” J. Acoust. Soc. Am. **43**, 822–828.

Rabiner, L. R. (**1968**b), “Speech synthesis by rule: An acoustic domain approach,” Bell System Tech. J. **47**, 17–37.

Rabiner, L. R., Schafer, R. W., and Rader, C. M. (**1969**), “The chirp z-transform algorithm and its application,” Bell System Tech. J. **48**, 1249–1292.

Rader, C. (**1967**), “Systems for compressing the bandwidth of speech,” IEEE Trans. Audio and Electroacoust. **AU-15**(3).

Ragazzini, J. R. and Franklin, G. F. (**1958**), *Sampled-data control systems* (McGraw-Gill, New York).

Ramachandran, R. P. and Kabal, P. (**1987**), “Stability and performance analysis of pitch filters in speech coders,” IEEE Trans. ASSP **35**(7), 937–946.

Ramamoorthy, V. and Jayant, N. (**1984**), "Enhancement of ADPCM speech by adaptive post-filtering," AT&T Bell Labs. Tech. J. , 1465–1475.

Ranke, O. F. (**1942**), "Das Massenverhaltnis zwischen Membran und Fliissigkeit im Innenohr," Akust. Z. **7**, 1–11.

Reddy, D. R. (**1967**), "Computer recognition of connected speech," J. Acoust. Soc. Am. **42**, 329–347.

Reddy, D. R. (**1969**), "Segment-synchronization problem in speech recognition," J. Acoust. Soc. Am. **46**, 89 (A).

Regnier, M. S. and Allen, J. B. (**2008**), "A method to identify noise-robust perceptual features: application for consonant /t/," J. Acoust. Soc. Am. **123**(5), 2801–2814.

Rhode, W. S. (**1971**), "Observations of the vibration of the basilar membrane in squirrel monkeys using the Mossbauer technique," J. Acoust. Soc. Am. **49**, 1218–1231.

Richardson, E. G. (**1953**), *Technical aspects of sound* (Elsevier Publ. Co., Amsterdam).

Riesz, R. R. (**1928**), "Differential intensity sensitivity of the ear for pure tones," Phys. Rev. **31**, 867–875.

Riesz, R. R. and Schott, L. (**1946**), "Visible speech cathode-ray translator," J. Acoust. Soc. Am. **18**, 50–61.

Riesz, R. R. and Watkins, S. A. (**1939**), "A synthetic speaker," J. Franklin Inst. **227**, 739–764.

Risberg, A. (**1959**), "A new coding amplifier system for the severely hard of hearing," in *Proc. 3rd Inter. Congr. on Acoust.* (Stuttgart, Germany).

Ritsma, R. (**1967**), "Frequencies dominant in the perception of the pitch of complex sounds," J. Acoust. Soc. Am., **42**, 191–198.

Rix, A., Beerends, J., Hollier, M., and Hekstra, A. (**Sep 2000**), "PESQ-the new ITU standard for end-to-end speech quality assessment," in *AES 109th Convention* (Los Angeles, CA).

Rose, J. E., Galambos, R., and Hughes, J. R. (**1959**), "Microelectrode studies of the cochlear nuclei of the cat," Bull. Johns Hopkins Hosp. **104**, 211–251.

Rose, R. C. and Barnwell III, T. P. (**1986**), "The self-excited vocoder—an alternate approach to toll quality at 4800 bps," in *Proc. ICASSP*.

Rosen, G. (**1958**), "Dynamic analog speech synthesizer," J. Acoust. Soc. Am. **30**, 201–209.

Rosenberg, A. E. (**1965**), "Effect of masking on the pitch of periodic pulses," J. Acoust. Soc. Am. **38**, 747–758.

Rosenberg, A. E. (**1971**a), "Effect of glottal pulse shape on the quality of natural vowels," J. Acoust. Soc. Am. **49**, 583–590.

Rosenberg, A. E. (**1971**b), "The preference of slope overload to granularity in the delta modulation of speech," J. Acoust. Soc. Am. **49**, 133 (A).

Rosenberg, A. E. (**1971a**), "Listener performance in a speaker verification task," J. Acoust. Soc. Am. **50**, 106 (A).

Rosenblith, W. A. and Stevens, K. N. (**1953**), "On the DL for frequency," J. Acoust. Soc. Am. **25**, 980–985.

Rydbeck, N. and Sundberg, C. E. (**1976**), "Analysis of digital errors in non-linear PCM systems," IEEE Trans. Communications **COM-24**, 59–65.

Salami, R., Laflamme, C., Adoul, J.-P., Kataoka, A., Hayashi, S., Moriya, T., Lamblin, C., Massaloux, D., Proust, S., Kroon, P., and Shoham, Y. (**1998**), "Design and Description of CS-ACELP: A Toll Quality 8 kb/s Speech Coder," IEEE Trans. Speech and Audio Processing **6**(2), 116–130.

Salmon, V. (**1946**a), "Generalized plane wave horn theory," J. Acoust. Soc. Am. **17**(3), 199–211, jan.

Salmon, V. (**1946**b), "A new family of horns," J. Acoust. Soc. Am. **17**(3), 212–218, jan.

Sawashima, M. (**August 1968**), "Observation of the glottal movements," in *Proc. Speech Symp* (Kyoto, Japan), pp. C–2–1.

Schafer, R. W. and Flanagan, J. L. (**June 1971**), "Speech synthesis by concatenation of formant-coded words," in *Bell System Tech. J*, vol. 50.

Schafer, R. W. and Rabiner, L. R. (**1970**), "System for automatic formant analysis of voiced speech," J. Acoust. Soc. Am. 47, pt. **2**, 634–648.

Schott, L. O. (**1948**), "A playback for visible speech," Bell Lab. Record **26**, 333–339.

Schroeder, M. (**1967**), "Determination of the Geometry of the Human Vocal Tract by Acoustic Measurements," J. Acoust. Soc. Am. **41**(4), 1002–1010.

Schroeder, M. and Atal, B. (**1985**), "Code-Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates," in *Proc. ICASSP*, pp. 937–940.

Schroeder, M. R. (**1959**), "Recent progress in speech coding at Bell Telephone Laboratories," in *Proc. Internat. Congr. Acoust.* (Stuttgart, Germany).

Schroeder, M. R. (**1962**), "Correlation techniques for speech bandwidth compression," J. Audio Eng. Soc. **10**, 163–166.

Schroeder, M. R. (**August 1968**), "Predictive coding of speech signals," in *Proc. Int. Congr. Acoust* (Tokyo, Japan), pp. C–5–4.

Schroeder, M. R. and Bird, C. M. (**1962**), "Single channel speech interpolator for 2:1 bandwidth reduction," J. Acoust. Soc. Am. **34**, 2003 (A).

Schroeder, M. R., Logan, B. F., and Prestigiacomo, A. J. (**September 1962**), "New applications of voice-excitation to vocoders," in *Stockholm Speech Comm. Seminar, R.I.T* (Stockholm, Sweden).

Seki, H. (**1958**), "A new method of speech transmission by frequency division and multiplication," J. Acoust. Soc. Japan **14**, 138–142.

Shadle, C. H. (**March 1985**), "The Acoustics of Fricative Consonants," Ph.D. thesis, MIT, Cambridge, MA.

Shadle, C. H., Barney, A., and Davies, P. (**1999**), "Fluid flow in a dynamic mechanical model of the vocal folds and tract. II Implications for speech production studies," J. Acoust. Soc. Am. **105**, 456–466.

Shannon, C. and Weaver, W. (**1949**), *The mathematical theory of communication* (University of Illinois, Urbana).

Shearme, J. N. (**1959**), "A simple maximum selecting circuit," Electronic Eng. **31**, 353–354.

Shearme, J. N. (**September 1962**), "Analysis of the performance of an automatic formant measuring system," in *Proc. Stockholm Speech Comm. Seminar, R.I.T* (Stockholm, Sweden).

Shearme, J. N., Smith, G. F., and Kelly, L. C. (**1962**), "A formant tracking system for speech measurements," Tech. Rep. JU 7-2, British post office. Joint Speech Research Unit, Eastcote, England.

Shepard, R. (**1962**), "The analysis of proximities: Multidimensional scaling with an unknown distance function (I and II)," Psychometrika **27**, 125–140,219–246.

Shipley, K. (**1971**), "Digital conversion of adaptive delta modulation to linear delta modulation," J. Acoust. Soc. Am. **50**, 107 (A).

Shoham, Y. (**1997**), “Very low complexity interpolative speech coding at 1.2 to 2.4 kbp,” in *Proc. ICASSP*, pp. 1599–1602.

Singhal, S. and Atal, B. S. (**1984**), “Improving performance of multi-pulse LPC coders at low bit rates,” in *Proc. ICASSP*, pp. 1.3.1–1.3.4.

Sinha, D. and Sundberg, C.-E. (**1999**), “Unequal error protection methods for perceptual audio coders,” in *Proc. ICASSP*, vol. 5, pp. 2423–2426.

Sivian, L. J. (**1929**), “Speech power and its measurement,” Bell System Tech. J. **8**, 646–661.

Slaymaker, F. H. (**1960**), “Bandwidth compression by means of vocoders,” IRE Trans. Audio **AU-8**, 20–26.

Smith, C. P. (**1951**), “A phoneme detector,” J. Acoust. Soc. Am. **23**, 446–451.

Smith, C. P. (**May 1957**), “Speech data reduction,” Tech. Rep. TR-57-111, Astia No. AD 117290, Air Force Cambridge Research Center, Bedford, Mass.

Smith, C. P. (**1963**), “Voice-communications method using pattern matching for data compression,” J. Acoust. Soc. Am. **35**, 805 (A).

Smith, S. (**1958**), “Diphlophonie und Luft-Schall-Explosionen,” Arch. Ohren-, Nasen- u. Kehlkopfheilk. ver. Z. **173**, 504–508.

Sondhi, M. M. and Schroeter, J. (**July 1987**), “A hybrid time-frequency domain articulatory speech synthesizer,” Trans. ASSP **ASSP-35**(7), 955–967.

Soong, F. and Juang, B.-H. (**1984**), “Line Spectral Pair (LSP) and Speech Data Compression,” in *Proc. ICASSP*, pp. 1.10.1–1.10.4.

Stachurski, J., McCree, A., and Viswanathan, V. (**1999**), “High quality MELP coding at bit rates around 4 kb/s,” in *Proc. ICASSP*, vol. 1, pp. 485–488.

Stark, R. E., Cullen, J. K., and Chase, R. (**1968**), “Preliminary work with the new Bell telephone visible speech translator,” Proc. Conf. on Speech-Analyzing Aids for the Deaf, Amer. Ann. Deaf, **113**, 205–214.

Stead, L. G. and Jones, E. T. (**March 1961**), “The SR.D.E. speech bandwidth compression project,” Tech. Rep. 1133, Signals Research and Development Establishment, Christchurch, England.

Steele, R. W. and Cassel, L. E. (**1963**a), “Dynamic encoding as applied to a channel vocoder,” J. Acoust. Soc. Am. **35**, 789 (A).

Steele, R. W. and Cassel, L. E. (**1963**b), “Effect of transmission errors on the intelligibility of vocoded speech,” IEEE Trans. Comm. Sys, **11**, 118–123.

Stevens, K. (**December 1956**), "Stop consonants," in *Quart. Rept., Acoustics Laboratory, Mass. Inst. Tech.* (Stockholm, Sweden).

Stevens, K. N. (**1950**), "Autocorrelation analysis of speech sounds," J. Acoust. Soc. Am. **22**, 769–771.

Stevens, K. N. (**1952**), "The perception of sounds shaped by resonant circuits," Ph.D. thesis, MIT, Cambridge, MA.

Stevens, K. N. (**1955**), "Auditory testing of a simplified description of vowel articulation," J. Acoust. Soc. Am. **27**, 882–887.

Stevens, K. N. (**May 1971**), "Airflow and Turbulence Noise for Fricative and Stop Consonants: Static Considerations," J. Acoust. Soc. Am. **50**(4), 1180–1192.

Stevens, K. N. (**1999**), *Acoustic Phonetics* (MIT Press, Cambridge, MA).

Stevens, K. N. and House, A. S. (**1955**), "Development of a Quantitative Description of Vowel Articulation," J. Acoust. Soc. Am. **27**(3), 401–493.

Stevens, K. N., Kasowski, S., and Fant, C. G. M. (**1953**), "An electrical analog of the vocal tract," J. Acoust. Soc. Am. **25**, 734–742.

Stevens, S. S. and Davis, H. (**1938**), *Hearing* (John Wiley & Sons, New York).

Sugamura, N. and Itakura, F. (**1981**), "Speech data compression by LSP speech analysis-synthesis technique," Trans. IECE **J 64-A**(8), 599–606, (in Japanese).

Sugimoto, T. and Hashimoto, S. (**September 1962**), "The voice fundamental pitch and formant tracking computer program by short-term autocorrelation function," in *Proc. Stockholm Speech Cornm. Seminar. R.I.T* (Stockholm, Sweden).

Supplee, L., Cohn, R., Collura, J., and McCree, A. (**1997**), "MELP: The New Federal Standard at 2400 BPS," in *Prof. ICASSP*, pp. 1591–1594.

Suzuki, J., Kadokawa, Y., and Nakata, K. (**1963**), "Formant frequency extration by the method of moment calculations," J. Acoust. Soc. Am. **35**, 1345–1353.

Tang, B., Shen, A., Alwan, A., and Pottie, G. (**March 1997**), "A Perceptually-Based Embedded Subband Speech Coder," IEEE Transactions on Speech and Audio Processing **5**(2), 131–140.

Taniguchi, T. (**1988**), "ADPCM with a Multiquantizer for Speech Coding," IEEE Journal Sel. Areas Communications **6**(2), 410–424.

Taniguchi, T., Amano, F., and Unagami, S. (**1990**), "Combined source and channel coding based on multimode coding," in *Proc. ICASSP*, pp. 477–480.

Tardelli, J. and Kreamer, E. (**1996**), "Vocoder intelligibility and quality test methods," in *Proc. ICASSP*, pp. 1145–1148.

Tarnoczy, T. H. (**1950**), "The speaking machine of Wolfgang von Kempelen," J. Acoust. Soc. Am. **22**, 151–166.

Tasaki, I., Davis, H., and Eldredge, D. H. (**1954**), "Exploration of cochlear potentials in guinea pig with a microelectrode," J. Acoust. Soc. Am. **26**, 765–773.

Teas, D. C., Eldredge, D. H., and Davis, H. (**1962**), "Cochlear responses to acoustic transients," J. Acoust. Soc. Am. **34**, 1438–1459.

Teranishi, R. and Umeda, N. (**August 1968**), "Use of pronouncing dictionary in speech synthesis experiments," in *Proc. Int. Congr, Acoust* (Tokyo, Japan), pp. B–5–2.

Tierney, J. (**1965**), "Digitalized voice-excited vocoder for telephone quality inputs using bandpass sampling of the baseband signal," J. Acoust. Soc. Am. **37**, 753–754.

Titchmarsh, E. C. (**1932**), *The theory of functions* (Oxford University Press, London).

Tomozawa, A. and Kaneko, H. (**1968**), "Companded delta modulation for telephone transmission," IEEE Trans. Comm. Tech. **COM-16**, 149–157.

Trancoso, I. and Atal, B. (**1986**), "Efficient procedures for finding the optimum innovation in stochastic coders," in *Proc. ICASSP*, pp. 2379–2382.

Tunturi, A. R. (**1955**), "Analysis of cortical auditory responses with the probability pulse," Am. J. Physiol. **181**, 630–638.

Umeda, N. (**March 1970**), "Text-to-speech conversion," in *IEEE Int. Conv. Digest.* (New York), pp. 216–217.

Upton, H. (**1968**), "Wearable eyeglass speech-reading aid," Proc. Conf. on Speech-Analyzing Aids for the Deaf, Amer. Ann. Deaf **113**, 222–229.

Uvliden, A., Bruhn, S., and Hagen, R. (**1998**), "Adaptive multi-rate. A speech service adapted to cellular radio network quality," in *Proc. Thirty-second Asilomar Conference*, vol. 1, pp. 343–347.

Vainio, J., Mikkola, H., Jarvinen, K., and Haavisto, P. (**1998**), "GSM EFR based multi-rate codec family," in *Proc. ICASSP*, vol. 1, pp. 141–144.

van Bergeijk, W. A. (**1961**), "Studies with artificial neurons," II. Analog of the external spiral innervation of the cochlea. Kybernetik **1**, 102–107.

van den Berg, J. W. (**1955**), "Transmission of the vocal cavities," J. Acoust. Soc. Am. **27**, 161–168.

van den Berg, J. W. (**1960**), "An electrical analogue of the trachea, lungs and tissues," Acta Physiol. Pharmacol. Neerl. **9**, 361–385.

van den Berg, J. W., Zantema, J. T., and Jr., P. D. (**May 1957**), "On the Air Resistance and the Bernoulli Effect of the Human Larynx," J. Acoust. Soc. Am. **29**(5), 626–631.

Velichko, V. M. and Zagoruyko, N. G. (**1970**), "Automatic recognition of 200 words," Int. J. Man-Machine Studies **2**, 223–234.

Vilbig, F. (**1950**), "An apparatus for speech compression and expansion and for replaying visible speech records," J. Acoust. Soc. Am. **22**, 754–761.

Vilbig, F. (**1952**), "Frequency band multiplication or division and time expansion or compression by means of a string filter," J. Acoust. Soc. Am. **24**, 33–39.

Vilbig, F. and Haase, K. (**1956**a), "Some systems for speech-band compression," J. Acoust. Soc. Am. **28**, 573–577.

Vilbig, F. and Haase, K. (**1956**b), "Uber einige Systeme fur Sprachbandkompression," Nachr. Stechn. Fachber. **3**, 81–92.

Voiers, W. D. (**1977**), "Diagnostic acceptability measure for speech communication systems," in *Proc. ICASSP*, pp. 204–207.

Voiers, W. D. (**1983**), "Evaluating processed speech using the Diagnostic Rhyme Test," Speech Technol. **1**(4), 30–39.

Voiers, W. D. (**1991**), "Effects of noise on the discriminability of distinctive features in normal and whispered speech," J. Acoust. Soc. Am. **90**, 2327.

Vyssotsky, V. A. (**1961**), "A block diagram compiler," Bell System Tech. J. **40**, 669–676.

Wang, S., Sekey, A., and Gersho, A. (**1992**), "An objective measure for predicting subjective quality of speech coders," IEEE J. Select. Areas in Comm. , 819–829.

Webster, A. G. (**1919**), "Acoustical impedance and the theory of horns," Proc. Nat. Acad. Sci. V.S. **5**, 275–282.

Webster, J. C. (**1961**), "Information in simple multidimensional speech messages," J. Acoust. Soc. Am. **33**, 940–944.

Wegel, R. L. (**1930**), "Theory of vibration of the larynx," Bell System Tech. J. **9**, 207–227.

Weinstein, C. J. (**1966**), "Short-time Fourier analysis and its inverse," Master's thesis, M.I.T. Dept. of Electrical Engineering, Mass. Cambridge.

Werner, P. A. and Danielsson, K. (**1958**), "17 kanals vocoder i laboratorientforande FOA3," Tech. Rep. A345, Laboratory for National Defense, Stockholm.

Weston, R. C. (**September 1962**), "Sampling and quantizing the parameters of a formant tracking vocoder system," in *Proc. Stockholm Speech Comm. Seminar, R.I.T* (Stockholm, Sweden).

Wickelgren, W. A. (**1965**), "Distinctive features and errors in short term memory for English vowels," J. Acoust. Soc. Am. **38**, 583–588.

Wickelgren, W. A. (**1966**), "Distinctive features and errors in short term memory for English consonants," J. Acoust. Soc. Am. **38**, 388.

Wiener, F. M. and Ross, D. A. (**1946**), "The pressure distribution in the auditory canal in a progressive sound field," J. Acoust. Soc. Am. **18**, 401–408.

Wiener, N. (**1949**), *Extrapolation, Interpolation and Smoothing of stationary time series, with Engineering Applications* (MIT Press, Cambridge MA).

Wong, S. W. (**1991**), "An evaluation of 6.4kbps speech codecs for Inmarsat-M system," in *Proc. ICASSP*.

Yang, W., Benbouchta, M., and Yantorno, R. (**1998**), "Performance of the modified bark spectral distortion measure as an objective speech quality measure," in *Proc. ICASSP*, pp. 541–544.

Yang, W. and Yantorno, R. (**1999**), "Improvement of MBSD by scaling noise masking threshold and correlation analysis with MOS difference instead of MOS," in *Proc. ICASSP* (Phoenix, AZ), pp. 673–676.

Yeldener, S. (**1999**), "A 4kbps toll quality harmonic excitation linear predictive speech coder," in *Proc. ICASSP*, pp. 481–484.

Young, M. A. and Campbell, R. A. (**1967**), "Effects of context on talker identification," J. Acoust. Soc. Am. **42**, 1250–1254.

Zantema, J. T. and P. Doornenbal, j. (**1957**), "On the air resistance and the Bernoulli effect of the human larynx," J. Acoust. Soc. Am. **29**, 626–631.

Zue, V., Seneff, S., and Glass, J. (**1990**), "Speech database development at MIT: TIMIT and beyond," Speech Communication **9**, 351–356.

Zwislocki, J. (**1948**), "Theorie der Schneckenmechanik," Ph.D. thesis, Tech. Hochschule, Zurich.

Zwislocki, J. (**1957**), "Some impedance measurements on normal and pathological ears," J. Acoust. Soc. Am. **29**, 1312–1317.

Zwislocki, J. (**1959**), "Electrical model of the middle ear," J. Acoust. Soc. Am. **31**, 841(A).

List of Figures

List of Tables