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 we 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 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., $\approx$ 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.

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.

Quantify the fo
ments:

[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).

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.

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.

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

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.

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 the 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.

Mark, I wrote thi
chance meeting ir
27, 2011 I hope yo
on it. Give it a try

1.1.1 Shannon's theory of information

The mathematical theory of information [Shannon, 1948, Shannon and Weaver, 1949] provides a mathematical framework for analyzing the end-to-end performance of any communications channel, *independent of the details of any system implementation*. Figure 1.2 shows the 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.

? A first important concept here is information as reciprocal probability; as the probability becomes small (the event becomes rare) and the amount of information increases. 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 $\sum_n p_n = 1$. This last constraint simply states that there must always be an outcome from every trial.

Key to Shannon's definition of information is that the *meaning* of the message plays no role. Rather information is strictly define in terms of the *entropy* $\mathcal{H}$, computed over N discrete probabilities $[p_n], n = 1, 2, \cdots, N$ defined by an *ensemble of discrete symbols*, represented using an equivilent set of *binary symbols* {0,1} (denoted bits), as used by Morses' code of dashes and dots. Morses' purpose in using bits to represent the letters of the English alphabet was *(i)* to be efficient [the most common vowel /e/ = *dot* is 0 and common consonant /t/ = *dash* is 1] while at the same time *(ii)* reducing the channel error rate to zero (i.e., an error rate so small as to be irrelivant), since a dash and a dot can be made arbritarilly distinct by using an arbritarilly small bandwidth.[2]

[2]It now seems obvious that Shannon's post WW-II work was inspired by Morses' code along with the Enigma effort at Benchly Park (Ref: *The Code Breakers*). The German's Enigma machine encryption code was first broken at Benchly

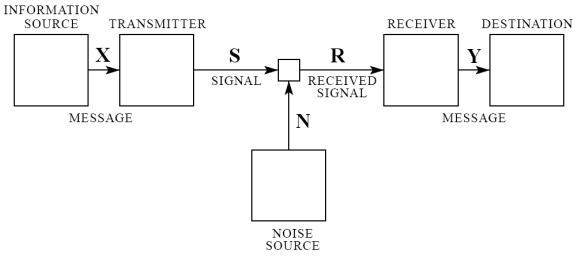

Figure 1.2: Schematic diagram of a Shannon's channel (communication model). The basic definitions are in terms of the siz boxes: X =source message, Y =received message, S =transmitted signal, R =received signal, N =noise (After Shannon and Weaver, 1949). Here it is assumed that the message X is a discrete 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.

The definition of *an ensemble of messages* is all possible unique sequences of symbols. Every ensemble, which is described by a probability distribution, has an entropy $\mathcal{H}$, defined as the average number of possible messages $N = 2^{\mathcal{H}}$ (measured in *bits/sequence*).

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 speech is remarkably low. There are apparently two reasons for this low information rate: *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, the human channel appears to be errorful compared to the modern electronic channel. The other side of the coin is that such low information rates allow 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, making humans *appear* to be more effective than machines in high noise. 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.

1.1.2 Entropy

Shannon's theory provided a paradigm shift [Pierce, 1973, Brillouin, 1962]. His elementary information theory relations between information $\mathcal{I}$ and entropy $\mathcal{H}$ help to quantify the information associated with the selection of a discrete message from a specified ensemble. If the messages x_n of the ensemble are independent, with probability of occurrence $p(x_n)$, the partial, or *specific information* (i.e., information density), associated with a selection is $\mathcal{I}_k = \log_2(1/p(x_k))$ bits.[3] The average information associated with selections from the set is the ensemble average $\mathcal{E}$ of density $\mathcal{I}_k$

? Is it 1:1 for

$$\mathcal{H}(X) \equiv \mathcal{E}\{\mathcal{I}\} = -\sum_n p(x_n)\log_2 p(x_n)$$

bits, or the source entropy. The log() transformation converts reciprocal probability into entropy in *bits*. 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 *volume* of the message space, which numerically is $N = 2^{\mathcal{H}}$. Note that entropy is a rate based on a *per symbol* transmission sampling rate. For the case of speech, this symbol (sampling) rate is difficult to identify with precision.

tion of prob-
sulting vari-
inimazation)
ntropy.

Park, thanks to one Polish spy (insert name), who worked in the factory where the Egnima was produced.

[3]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."

1.1.3 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., $P(x_n) = 1/42$] 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.[4] In conversational speech about 10 phonemes are uttered per second. The written equivalent of the information generated is therefore less than 50 bits/sec.

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

1.2 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

[4] 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 are used as estimates of $P(x_n)$, 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
			g	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	

$H(X) = -\sum_n P(x_n) \log_2 P(x_n) = 4.9$ 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.

may lead to errors in the received message. Perceptual 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 signment.

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[5]

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

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 $p_{AB\gamma}(y_j|x_n)$ for every (n, 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 $x_n \in \{$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

[5] 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.

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

$$p_{h|s} \approx \frac{C(s,h)}{\sum_h C(s,h)} \tag{1.5}$$

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

STIMULUS	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

UNVOICED VOICED NASAL

RESPONSE

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

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.

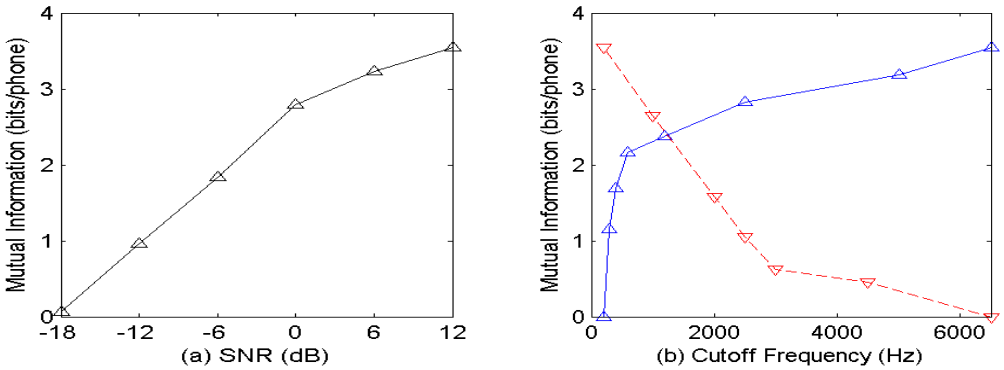

Figure 1.4: (a) Mutual information between spoken and perceived consonant labels, as a function of SNR, over an acoustic channel with 6.3 [kHz] bandwidth (0.2-6.5 [kHz]). (b) Mutual information between spoken and perceived consonant labels, at 12dB SNR, over lowpass and highpass acoustic channels with 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).

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 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.

1.3 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 and Weaver [1949] demonstrated 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 20 [kHz] (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.

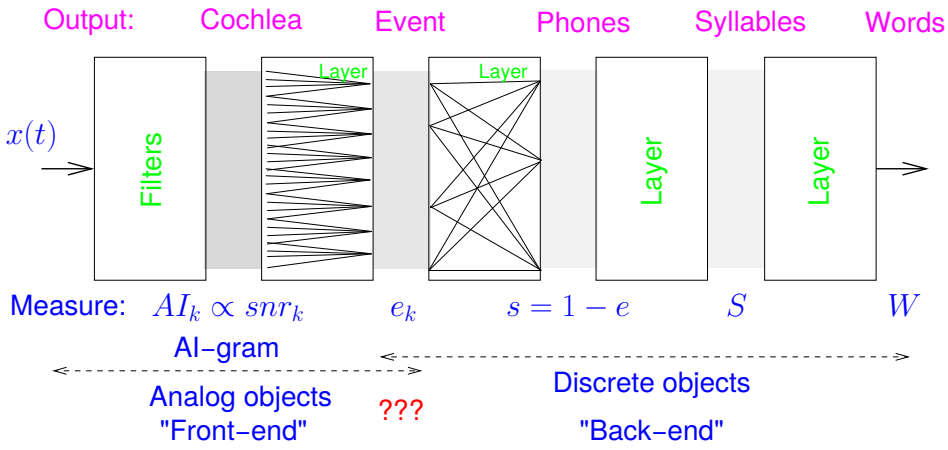

Figure 1.5: Model block diagram summary of speech recognition by humans. At the top of each block is a label that attempts to identify the physical operation, or a unit being recognized. The labels below the boxes indicate the probability measure defined at that level. See the text for the discussion 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.

1.3.1 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 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.

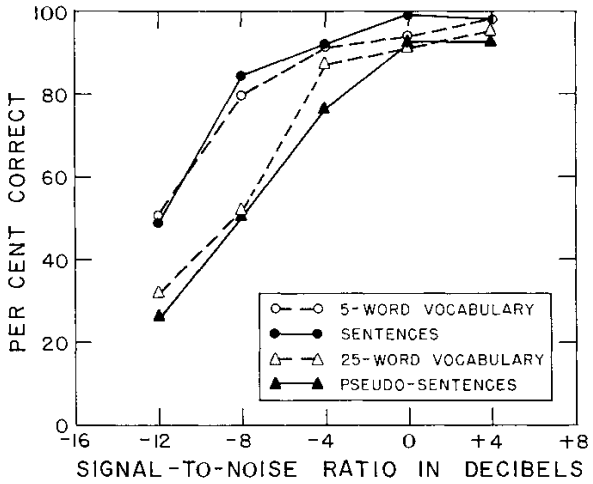

Figure 1.6: This figure, from Miller (1962), summarizes the results of a 4-way experiment, performed as a function of the signal to noise ratio. Test 1 (open circles, dashed line) shows $P_c(SNR)$ for 5 word vocabularies, with no context. In test 2 (closed circles, solid line) 5 word sentences were made 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.

1.3.2 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*[6] 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 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].

1.4 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

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

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.

is the infor-
ewave, or a
about a sine
dwidth mod-