I. 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*

Man's primary method of communication is speech. He is unique in his ability to transmit information with his voice. Of the myriad varieties of life sharing our world, only man has developed the vocal means for coding and conveying information beyond a rudimentary stage. It is more to his credit that he has developed the facility from apparatus designed to subserve other, more vital purposes.

Because man was constructed to live in an atmosphere, it is not unnatural that he 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 in several miles of such transmission.)

Despite the desires and motivations to accomplish communication at distances, it was not until man 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. Less than a hundred years have passed since the first practical telephone was put into operation. Today it is estimated that more than 200 million telephones are in use.

1.1. The Advent of Telephony

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 his well-equipped laboratory. Among these men, ALEXANDER GRAHAM BELL is considered by many not only to have produced and demonstrated the first practicable telephone, but also have made its first commercial application. Some contend that ELISHA GRAY was at least concomitant with BELL in his developments. Others claim PHILIPP REIS anticipated both BELL and GRAY by several years. Depending upon the country,

one can find factions in support of still other inventors. Regardless of when, and through whose efforts electrical telephony ceased to be a dream and yielded to practice, there is enough praise and admiration for all these early scientists.

Among the pioneers BELL was somewhat unique for his background in physiology and phonetics. His comprehension of the mechanisms of speech and hearing was undoubtedly valuable, if not crucial, in his electrical experimentation. Similar understanding is equally important with 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." This is what he set out to do and is what he accomplished. BELL'S basic notion—namely, preservation of acoustic waveform—clearly proved to be an effective means for speech transmission. To the present day most telephone systems operate on this principle.

Although the waveform principle is exceedingly satisfactory and has endured for almost a century, it probably is not the most efficient means for voice transmission. Communication engineers have recognized for many years that a substantial mismatch exists between the information capacity of the human source-sink 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.

Recent 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 facility. 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.

1.2. Efficient Transmission of Speech

The 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 information associated with a selection is $I=-\log_2 P(x_i)$ bits. The average information associated with selections from the set is the ensemble average

$$H(X) = -\sum_i P(x_i) \log_2 P(x_i)$$

bits, or the source entropy.

Consider, in these terms, a phonemic transcription of speech; that is, the written equivalent of the meaningfully distinctive sounds of speech. Take English for example. Table 1.1 shows a list of 42 English phonemes including vowels, diphthongs and consonants, and their relative frequencies of occurrence in prose (DEWEY). If the phonemes are selected for utterance with equal probability [i.e., $P(x_i) = \frac{1}{42}$] 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 1.1, then $H(X)$ falls to 4.9 bits. The sequential contraints imposed upon the selection of speech sounds by a given language reduce this average information still further[1]. In conversational speech about 10 phonemes are uttered per second. The written equivalent of the information generated is therefore less than 50 bits/sec.

The conventional voice link is of course not a discrete channel but a continuous one. For a continuous channel, an existence proof can be given for the maximum error-free rate of information transmission (SHANNON and WEAVER). If the channel has bandwidth BW cps and signal and noise powers S and N, respectively, a method of coding exists such that information can be signalled, with arbitrarily small error, at a rate $C = BW \log_2 [1 + (S/N)]$ bits/sec. A conventional (waveform) voice channel has a bandwidth typically around 3000 cps, or more, and a signal-to-noise ratio of about 30 db. The formula therefore indicates that such a channel has the capacity to transmit information at rates on the order or 30000 bits/sec.

Similar bit rates are encountered in conventional digital transmission of speech waveforms (without further encoding). In PCM transmission for example, the signal is sampled at the Nyquist rate (2 BW) and, to maintain tolerable distortion, the amplitude is commonly quantized to an accuracy of one or two per cent. For a 64 level (6 bit) quantization, therefore, a typical bit rate is $2(3000) \log_2 64 = 36000$ bits/sec.

Table 1.1. *Relative frequencies of English speech sounds in standard prose.* (After DEWEY)

Vowels and diphthongs			Consonants		
Phoneme	relative frequency of occurrence %	$-P(x_i) \log_2 P(x_i)$	Phoneme	relative frequency of occurrence %	$-P(x_i) \log_2 P(x_i)$
ɪ	8.53	0.3029	n	7.24	0.2742
ə	4.63	0.2052	t	7.13	0.2716
æ	3.95	0.1841	r	6.88	0.2657
ɛ	3.44	0.1672	s	4.55	0.2028
ɒ	2.81	0.1448	d	4.31	0.1955
ʌ	2.33	0.1264	l	3.74	0.1773
i	2.12	0.1179	ð	3.43	0.1669
e, eɪ	1.84	0.1061	z	2.97	0.1507
u	1.60	0.0955	m	2.78	0.1437
aɪ	1.59	0.0950	k	2.71	0.1411
oʊ	1.30	0.0815	v	2.28	0.1244
ɔ	1.26	0.795	w	2.08	0.1162
ʊ	0.69	0.0495	p	2.04	0.1146
aʊ	0.59	0.0437	f	1.84	0.1061
ɑ	0.49	0.0376	h	1.81	0.1048
o	0.33	0.0272	b	1.81	0.1048
ju	0.31	0.0258	ŋ	0.96	0.0644
ɔɪ	0.09	0.0091	ʃ	0.82	0.0568
			g	0.74	0.0524
			j	0.60	0.0443
			tʃ	0.52	0.0395
			dʒ	0.44	0.0344
			θ	0.37	0.0299
			ʒ	0.05	0.0055
Totals 38				62	

$H(X) = -\sum_i P(x_i) \log_2 P(x_i) = 4.9$ bits. If all phonemes were equiprobable, then $H(X) = \log_2 42 = 5.4$ bits.

These capacities are on the order of six or seven hundred times greater than that apparently required for the written equivalent. The latter presumably should require a bandwidth of only about 5 cps for the 30 db S/N channel. Does this mean that the acoustic speech signal contains 600 times more information than its discretely transcribed equivalent? Or does it suggest that the acoustic time-waveform is an

[1] 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 $P(x_i)$, 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 units such as paragraphs or chapters are considered, the redundancy may be still higher (SHANNON).

inefficient code for the speech information? Does it imply that the human is capable of processing information at 30000 bits/sec? Or does it mean that the receiver discards much of the transmitted information?

Intuitively it is clear that the acoustic signal contains more information than the written equivalent. How much more is not clear. Indeed it is not clear that such a measure can be made. The information rate of a continuous source can be defined only after a fidelity criterion is established for representing the signal in terms of a specific code. The criterion for defining the source entropy might be either subjective or objective. In speech communication the perceptual ability of the receiver usually dictates the necessary precision. Certain of these abilities can be established by psychoacoustic measurement. Different rates are ascribed to the source depending upon the coded form of the information and upon the perceptual criteria which apply. For example, if intelligibility and quality were the criteria, the source rate and channel capacity would be expected to be greater than if intelligibility alone were the criterion.

Although it may not be possible to answer the question "How much information is in the speech wave?", one can show from synthesis experiments that speech, closely equivalent perceptually to a specific waveform coding, can be transmitted over channels with capacities appreciably less than 30000 bits/sec. Present indications are that these capacities might eventually be made as low as one thousand to two thousand bits/sec. More will be said about such possibilities in a later chapter.

1.3. Capacity of the Human Channel

As just suggested, it is the fidelity criterion which establishes the information rate of a source. The criterion is determined by the ability of the receiver to discriminate differences in the received signal. Psychoacoustic experimentation with auditory limens often provides an upper bound to this ability. More basic, perhaps, but also more difficult to measure and apply in transmission system design, is the ability of man to assimilate and process information.

A number of experimental efforts have been made to assess man's informational capacity. The experiments necessarily concern specific, idealized perceptual tasks. They consequently yield information measures which are strictly interpretable only within the framework of the particular experiment. 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, Stevens and Hayes). 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 pointing to targets by hand, the rates 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). 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). 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, J. C.). 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 phonemes, 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.

In protracted conversation the constraints of the language and the reasonably efficient human memory usually enable a listener to switch between a decoding of phonemic content and an observation of personal traits. Prosodic information can be assimilated along with phonemic features which relate directly to the written equivalent of the spoken information. The latter are customarily identified with speech intelligibility, while the former is loosely associated with speech quality. Intelligibility is conventionally quantified in terms of articulation scores and rates of receiving the written-equivalent information. Speech quality, as yet, has little basis for quantification. Until both intelligibility and quality can be suitably defined, the fidelity criteria for estimating speech information rates will not be firmly established.

1.4. Analysis-Synthesis Telephony: An Approach to Improved Efficiency

Despite the equivocal aspects surrounding estimates of human channel capacity and speech information rates, it is clear that a mismatch exists between the capacity of the conventional voice channel and the information rate of the source feeding it. One approach toward improving the match is to incorporate into the transmission system as many as possible of the constraints characterizing the production and perception of speech. This information, built into the communication link, is information that need not be transmitted. Another way of viewing the situation is that the channel, so constrained, confines the message ensemble to the sounds of speech. In general, no other sounds will be transmitted with acceptable fidelity.

Having built into the system constraints appropriate to production and perception, communication is effected by signalling certain parameters of the constraints. The nature of the incorporated constraints therefore influences the form of coding for the speech information. Assume, for example, that the limitations on the mechanical movements of the vocal tract are to be taken into account in the transmission system. One receiving device for realizing such restrictions might be a mechanical or electrical analog of the vocal mechanism. Speech information might then be coded and transmitted in terms of tract dimensions, deformations, and properties of vocal excitation.

Voice communication systems in which a conscious effort is made to improve efficiency by constraining the facility according to the characteristics of speech and hearing are customarily referred to as *analysis-synthesis systems*. The term is often taken as synonymous with speech-compression or band-saving systems. A main purpose of this monograph is to set forth the fundamental properties of speech and hearing which relate to communication systems of the analysis-synthesis type. A further purpose is to outline techniques for utilizing the properties of speech and hearing in practical transmission systems.

In attending to these objectives, the physiological and acoustical properties of the human vocal apparatus will be considered first. Next, the fundamental principles of the hearing mechanism will be examined. These basic expositions will then be followed by topics in speech analysis, speech synthesis, and speech perception. The final discussion will center on application of the preceding results to realizable speech-coding systems.

II. 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 man discovered he could supplement his communicative hand signals with related "gestures" of his vocal tract. Sir Richard Paget sums up this speculation quite neatly. "What drove man to the

[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.