# Speech Analysis Synthesis and Perception

Third Edition

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

2008

# Contents

# List of Figures

# List of Tables

# 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

Our primary method of communication is speech. Humans are unique in our ability to transmit information with his voice. 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. It is more to our credit that we have developed the facility from apparatus designed to subserve other, more vital purposes.

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 in 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. Little more than a hundred years have passed since the first practical telephone was put into operation; there are now, by some accounts, 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.

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 wilh 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. Waveform coding was the most widely used form of telephony until approximately the year 2000, when the number of digital cellular telephones began to outnumber the number of analog handsets. As we shall see, even digital telephony preserves the waveform, in the sense that only perceptually insignificant distortions are allowed.

Although the waveform principle is exceedingly satisfactory and has endured for almost a century, it 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 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.

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 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 create ideas larger than ourselves. The natural environment for speaking is noisy and complicated, with a continuously changing visual and auditory channel, as depicted, for example, in Fig. **??**. In this famous painting, a group of friends relaxes 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.

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

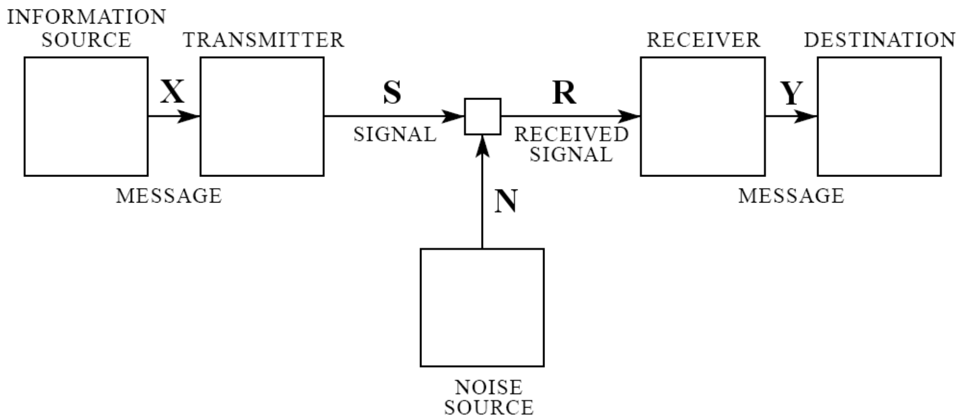


Figure 1.2: Schematic diagram of a general communication system. $X$ =source message, $Y$ =received message, $S$ =transmitted signal, $R$ =received signal, $N$ =noise. (After Shannon and Weaver, 1949)

Before speaking, every talker conceives a message: a sequence of words, possibly annotated with subtle hints of nuance and opinion (?, ?). The message is symbolic, and therefore digital: most of the content of a spoken message may be equivalently conveyed in an e-mail. In most cases, however, we find it pleasant to encode the message in an analog medium, by configuring the speech articulators (the lips, jaw, tongue, soft palate, larynx, and lungs) in order to generate an acoustic waveform. A listener measures the acoustic signal, and converts it into a neural code. The neural code passes through a series of neural circuits until, eventually, the listener has decoded the intended linguistic message–or something approximating the intended 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 performance of the channel.

The mathematical theory of information (?, ?) provides a useful mechanism for analyzing the end-to-end performance of any communications channel, independent of the details of its implementation. Fig. **??** 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. 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 the 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 are much less effective than machines. Second, low information rate allows speech transmission over extremely noisy acoustic channels. Human listeners (but not machines, yet) are able to correctly understand meaningful linguistic messages transmitted at signal to noise ratios (SNR) as low as -20dB; in this respect, humans are much 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 party in Fig. **??**.

## 1.2 Entropy of the Speech Source

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 $I = \log_2(1/P(x_i))$ bits[1]. The average information associated with selections from the set is the ensemble average

$$H(X) = \sum_i P(x_i) \log_2 \left( \frac{1}{P(x_i)} \right) = -\sum_i P(x_i) \log_2 P(x_i)$$

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 **??** shows a list of 42 English phonemes including vowels, diphthongs and consonants, and their relative frequencies of occurrence in prose (?, ?). 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 **??**, 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[2]. In conversational speech about 10 phonemes are uttered per second. The written equivalent of the information generated is therefore less than 50 bits/sec.

## 1.3 Conditional Entropy of Received Speech

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. Perceptual errors can be characterized by the conditional probability that the receiver decodes symbol $y_j$, given that the transmitter encoded symbol $x_i$. This probability may be writtten as $P_{AB\gamma}(y_j|x_i)$, 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

$$P_{AB\gamma}(y_j|x_i) = \delta_{ij} \equiv \begin{cases} 1 & y_j = x_i \\ 0 & \text{otherwise} \end{cases} \tag{1.1}$$

If $P_{AB\gamma}(y_j|x_i) \neq \delta_{ij}$, then one may say that the communication system is itself introducing "information" into the received signal. This is an undesirable behavior, because the "information" generated by the communication channel is independent of the information generated at the source; this extra "information" is usually called "error." The average rate at which the communication

[1] 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." All three units are commonly used in practice.

[2] 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 (?, ?).

Table 1.1: Relative frequencies of English speech sounds in standard prose. (After Dewey, 1923)

| Vowels and diphthongs | | | Consonants | | |
|---|---|---|---|---|---|
| Phoneme | relative frequency of occurence (%) | $-P(x_i)\log_2 P(x_i)$ | Phoneme | relative frequency of occurence (%) | $-P(x_i)\log_2 P(x_i)$ |
| TIPAI | 8.53 | 0.3029 | TIPAn | 7.24 | 0.2742 |
| TIPAA | 4.63 | 0.2052 | TIPAt | 7.13 | 0.2716 |
| TIPAæ | 3.95 | 0.1841 | TIPAr | 6.88 | 0.2657 |
| TIPAE | 3.44 | 0.1672 | TIPAs | 4.55 | 0.2028 |
| TIPA5 | 2.81 | 0.1448 | TIPAd | 4.31 | 0.1955 |
| TIPA2 | 2.33 | 0.1264 | TIPAl | 3.74 | 0.1773 |
| TIPAi | 2.12 | 0.1179 | TIPAT | 3.43 | 0.1669 |
| TIPAe, TIPAeI | 1.84 | 0.1061 | TIPAz | 2.97 | 0.1507 |
| TIPAu | 1.60 | 0.0955 | TIPAm | 2.78 | 0.1437 |
| TIPAAI | 1.59 | 0.0950 | TIPAk | 2.71 | 0.1411 |
| TIPAoU | 1.30 | 0.0815 | TIPAv | 2.28 | 0.1244 |
| TIPAO | 1.26 | 0.795 | TIPAw | 2.08 | 0.1162 |
| TIPAU | 0.69 | 0.0495 | TIPAp | 2.04 | 0.1146 |
| TIPAAU | 0.59 | 0.0437 | TIPAf | 1.84 | 0.1061 |
| TIPAA | 0.49 | 0.0376 | TIPAh | 1.81 | 0.1048 |
| TIPAo | 0.33 | 0.0272 | TIPAb | 1.81 | 0.1048 |
| TIPAju | 0.31 | 0.0258 | TIPAN | 0.96 | 0.0644 |
| TIPAOI | 0.09 | 0.0091 | TIPAS | 0.82 | 0.0568 |
| | | | TIPAg | 0.74 | 0.0524 |
| | | | TIPAj | 0.60 | 0.0443 |
| | | | TIPAtS | 0.52 | 0.0395 |
| | | | TIPAdZ | 0.44 | 0.0344 |
| | | | TIPAT | 0.37 | 0.0299 |
| | | | TIPAZ | 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 channel introduces errors into a transmitted signal is called the *equivocation* or *conditional entropy* of $Y$ given $X$, and is defined to be

$$H_{AB\gamma}(Y|X) = -\sum_i \sum_j P_{AB\gamma}(x_i, y_j) \log_2 P_{AB\gamma}(y_j|x_i) \tag{1.2}$$

$$= -\sum_i P(x_i) \sum_j P_{AB\gamma}(y_j|x_i) \log_2 P_{AB\gamma}(y_j|x_i)$$

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

$$I_{AB\gamma}(X,Y) = H(X) - H_{AB\gamma}(Y|X) \tag{1.3}$$

$$= \sum_i \sum_j P(x_i) P_{AB\gamma}(y_j|x_i) \left( \frac{P_{AB\gamma}(y_j|x_i)}{P(x_i)} \right)$$

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

$$H(Y|X) \approx 0.985 \log_2(1/0.985) + 0.015 \log_2(1/0.015) = 0.11 \text{ bits/symbol}^3. \tag{1.4}$$

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. (**??**) 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 (?) 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,,,s,z,,,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 /a/. 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. **??**. In a confusion matrix, entry $C(i, j)$ lists the number of times that phoneme $x_i$ was perceived as phoneme $y_j$. The conditional probability $P(y_j|x_i)$ may be estimated as

$$P(y_j|x_i) \approx \frac{C(i,j)}{\sum_j C(i,j)} \tag{1.5}$$

Using the approximation in **??**, 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. Fig. **??**(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. Fig. **??**(b) shows the information transmitted over the lowpass-filtered and highpass filtered channels, as a function of the cutoff frequency.

[3] This approximation results from the assumption that only two events matter: the phoneme is either correctly or incorrectly recognized. The actual equivocation of a 42-phoneme 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 confusion matrix (6300Hz bandwidth, -6dB SNR). Entry $(i, j)$ in the matrix lists the number of times that a talker said consonant $x_i$, and a listener heard consonant $y_j$. Each consonant was uttered as the first phoneme in a CV syllable; the vowel was always /a/. (After Miller and Nicely, 1955)

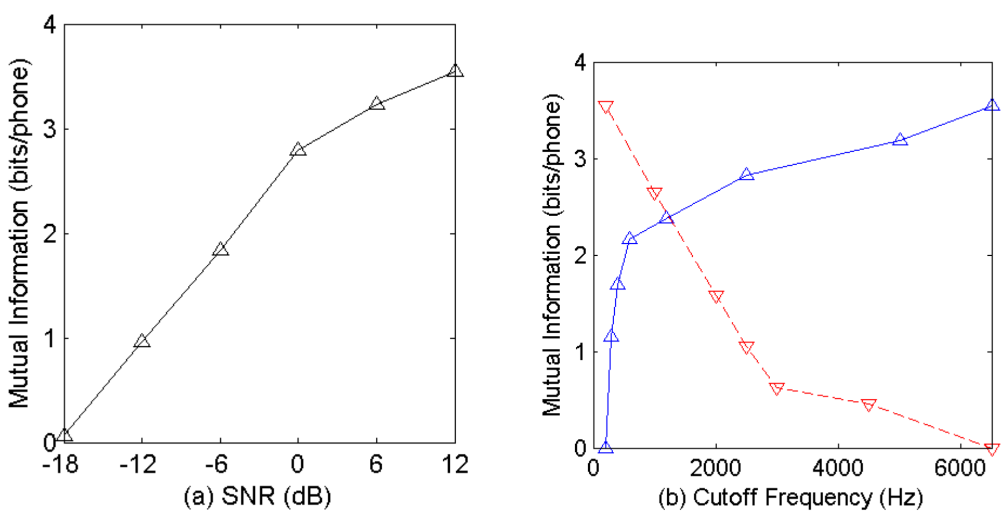


Figure 1.4: (a) Mutual information between spoken and perceived consonant labels, as a function of SNR, over an acoustic channel with 6300Hz bandwidth (200-6500Hz). (b) Mutual information between spoken and perceived consonant labels, at 12dB SNR, over lowpass and highpass acoustic channels with the specified cutoff frequencies. The lowpass channel contains information between 200Hz and the cutoff; bit rate is shown with a solid line. The highpass channel contains information between the cutoff and 6500Hz; bit rate is shown with a dashed line. (After Miller and Nicely, 1955)

## 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 (?) that no algorithm can transmit more information than

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

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

$$C(B, S/N) = \int_0^B \log_2\left(1 + \frac{S(f)}{N(f)}\right) df \quad \frac{\text{bits}}{\text{second}} \tag{1.7}$$

where $S(f)$ and $N(f)$ are the power spectra of the speech and noise, respectively. Speech is transmitted over an acoustic channel with bandwidths varying between about 3000Hz (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 $S/N = 0.0625$), 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 30dB, or $S/N \approx 1000$), the channel capacity of a 20kHz channel is 20,000 bits/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 **??** 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 (?, ?). 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 (?, ?). 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 (?, ?). 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 (?, ?). 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.

## 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 phonemes 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. After a reader has finished understanding chapter 6, the rest 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.

## References


Dewey, G. (1923). *Relative frequency of English speech sounds.* Cambridge, Massachusetts: Harvard University Press.

Fletcher, H. (1922). The nature of speech and its interpretation. *Bell System Technical Journal*, *1*, 129-144.

Levelt, W. J. M. (1989). *Speaking: from intention to articulation.* Cambridge, MA: MIT Press.

Licklider, J. C. R., Stevens, K. N., & Hayes, J. R. M. (1954). *Studies in speech, hearing and communication. final report, contract W-19122ac-1430.* Cambridge, Mass.

Miller, G. A., & Nicely, P. E. (1955). Analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.*, *27*, 338-352.

Pierce, J. R., & Karlin, J. E. (1957). Information rate of a human channel. *Proc. I.R.E.*, *45*, 368.

Pollack, I., & Ficks, L. (1954). Information of elementary multidimensional auditory displays. *J. Acoust. Soc. Am.*, *26*, 155-158.

Shannon, C., & Weaver, W. (1949). *The mathematical theory of communication.* Urbana: University of Illinois.

Webster, J. C. (1961). Information in simple multidimensional speech messages. *J. Acoust. Soc. Am.*, *33*, 940-944.