

Of course, the elements of the speech chain have not changed since the original publication of our book. Indeed, the physics of sound and the physiology of speech and hearing remain unchanged since the earliest humans. On the other hand, great changes have occurred during the last three decades in our understanding of the subject and in the technology we use to perform research and to build commercial devices to transmit, produce, or recognize speech.

One of the driving forces behind these advances has been the availability of powerful yet affordable computer technology. Digital techniques have become pervasive in our society. Consumer products, medical equipment, and manufacturing automation often depend on digital technology, as do hearing aids, speech recognizers and synthesizers, and audiometers.

We would like to thank the many people who have shared their ideas with us and who have been kind enough to read drafts of part or all of this book, including Jont Allen, Bishnu Atal, David Berkley, Nikel Jayant, Peter Ladefoged, Ilse Lehiste, Steve Levinson, Mike Noll, David Pisoni, Larry Rabiner, Catherine Ringen, Alex Rudnicki, Manfred Schroeder, Juergen Schroeter, and Bruce Smith. They helped us greatly to focus on the improvements that we believe make this edition more informative than its predecessors. We would also like to thank Julia Hirschberg and Thrasos Pappas for their kind help in preparing some of the speech spectrograms included in the book.

*Peter Denes
Elliot Pinson*



The Speech Chain

We usually take for granted our ability to produce and understand speech and give little thought to its nature and function, just as we are not particularly aware of the action of our hearts, brains, or other essential organs. It is not surprising, therefore, that many people overlook the great influence of speech on the development and functioning of human society.

Wherever human beings live together, they develop a system of talking to each other; even people in the most isolated societies use speech. Speech, in fact, is one of the few basic abilities-tool making is another-that set us apart from other animals and are closely connected with our ability to think abstractly.

Why is speech so important? One reason is that the development of human culture is made possible- to a great extent-by our ability to share experiences, to exchange ideas and to transmit knowledge from one generation to another; in other words, our ability to communicate with others. We can communicate with each other in

many ways. The smoke signals of the Apache Indian, the starter's pistol in a 100-yard dash, the sign language used by deaf people, the Morse Code and various systems of writing are just a few examples of the many different systems of communication that have evolved to meet special needs. Unquestionably, however, speech is the system that human societies have found, under most circumstances, to be far more efficient and convenient than any other.

You may think that writing is a more important means of communication than speech. After all, the written word and the output of printing presses appear to be more efficient and more durable means of transmitting information. Yet, no matter how many books and newspapers are printed, the amount of information exchanged by speech is still greater. The use of books and printed matter has expanded greatly in our society, but so has the use of telephones, radio, and television.

In short, human society relies heavily on the free and easy interchange of ideas among its members and, for many reasons, we have found speech to be our most convenient form of communication.

Through its constant use as a tool essential to daily living, speech has developed into a highly efficient system for the exchange of even our most complex ideas. It is a system particularly suitable for widespread use under the ever changing and varied conditions of life. It is suitable because it remains functionally unaffected by the many different voices, speaking habits, dialects and accents of the millions who use a common language. And it is suitable for widespread use because speech—to a surprising extent—is invulnerable to severe noise, distortion and interference.

Speech is well worth careful study. It is worthwhile because the study of speech provides useful insights into the nature and history of human civilization. It is worthwhile for the communications engineer because a better understanding of the speech mechanism helps in developing better and more efficient communication systems. It is worthwhile for all of us because we depend on speech so heavily for communicating with others.

The study of speech is also important for the development of human communication with machines. We all use automatons, like push-button telephone-answering machines and automatic elevators, which either receive instructions from us or report back to us on their

operations. Frequently, they do both, like the computers used so extensively in our society; their operation increasingly relies on frequent, fast, and convenient exchanges of information with users. In designing communication systems or “languages” to link user and machine, it should prove worthwhile to have a firm understanding of speech, that system of person-to-person communication whose development is based on the experience of many generations.

When most people consider speech, they think only in terms of moving lips and tongue. A few others, who have found out about sound waves, perhaps in the course of building or using stereo systems, will also associate certain kinds of sound waves with speech. In reality, speech is a far more complex process, involving many more levels of human activity, than such a simple approach would suggest.

A convenient way of examining what happens during speech is to take the simple situation of two people talking to each other. For example, you as the speaker, want to transmit information to another person, the listener. The first thing you have to do is arrange your thoughts, decide what you want to say and then put what you want to say into *linguistic form*. The message is put into linguistic form by selecting the right words and phrases to express its meaning, and by placing these words in the order required by the grammatical rules of the language. This process is associated with activity in the speaker's brain, and it is from the brain that appropriate instructions, in the form of impulses along the motor nerves, are sent to the muscles that activate the vocal organs—the lungs, the vocal cords, the tongue, and the lips. The nerve impulses set the vocal muscles into movement which, in turn, produce minute pressure changes in the surrounding air. We call these pressure changes a *sound wave*. Sound waves are often called *acoustic waves*, because acoustics is the branch of physics concerned with sound.

The movements of the vocal organs generate a speech sound wave that travels through the air between speaker and listener. Pressure changes at the ear activate the listener's hearing mechanism and produce nerve impulses that travel along the acoustic nerve to the listener's brain. In the listener's brain, a considerable amount of nerve activity is already taking place, and this activity is modified by the nerve impulses arriving from the ear. This modification of brain activity, in ways that are not yet fully understood, brings about

recognition of the speaker's message. We see, therefore, that speech communication consists of a chain of events linking the speaker's brain with the listener's brain. We shall call this chain of events the *speech chain* (see Figure 1.1).

It might be worthwhile to mention at this point that the speech chain has an important side link. In the simple speaker-listener situation just described, there are really two listeners, not one, because speakers not only speak, but also listen to their own voice. In listening, they continuously compare the quality of the sounds they produce with the sound qualities they intended to produce and make the adjustments necessary to match the results with their intentions.

There are many ways to show that speakers are their own listeners. Perhaps the most amusing is to delay the sound "fed back" to the speaker. This can be done quite simply by recording the speaker's voice on a tape recorder and playing it back a fraction of a second later. The speaker listens to the delayed version over ear-phones. Under such circumstances, the unexpected delay in the feedback sound makes the speaker stammer and slur. This is the so-called *delayed speech feedback effect*. Another example of the importance of "feedback" is the general deterioration of the speech of people who have suffered prolonged deafness. Deafness, of course, deprives people of the speech chain's feedback link. To a limited extent, we can tell the kind of deafness from the type of speech deterioration it produces.

Let us go back now to the main speech chain, the links that connect speaker with listener. We have seen that the transmission of a message begins with the selection and ordering of suitable words and sentences. This can be called the *linguistic level* of the speech chain.

The speech event continues on the *physiological level*, with neural and muscular activity, and ends, on the speaker's side, with the generation and transmission of a sound wave, the *physical (acoustic) level* of the speech chain.

At the listener's end of the chain, the process is reversed. Events start on the physical level, when the incoming sound wave activates the hearing mechanism. They continue on the physiological level with neural activity in the hearing and perceptual mechanisms. The speech chain is completed on the linguistic level when the listener

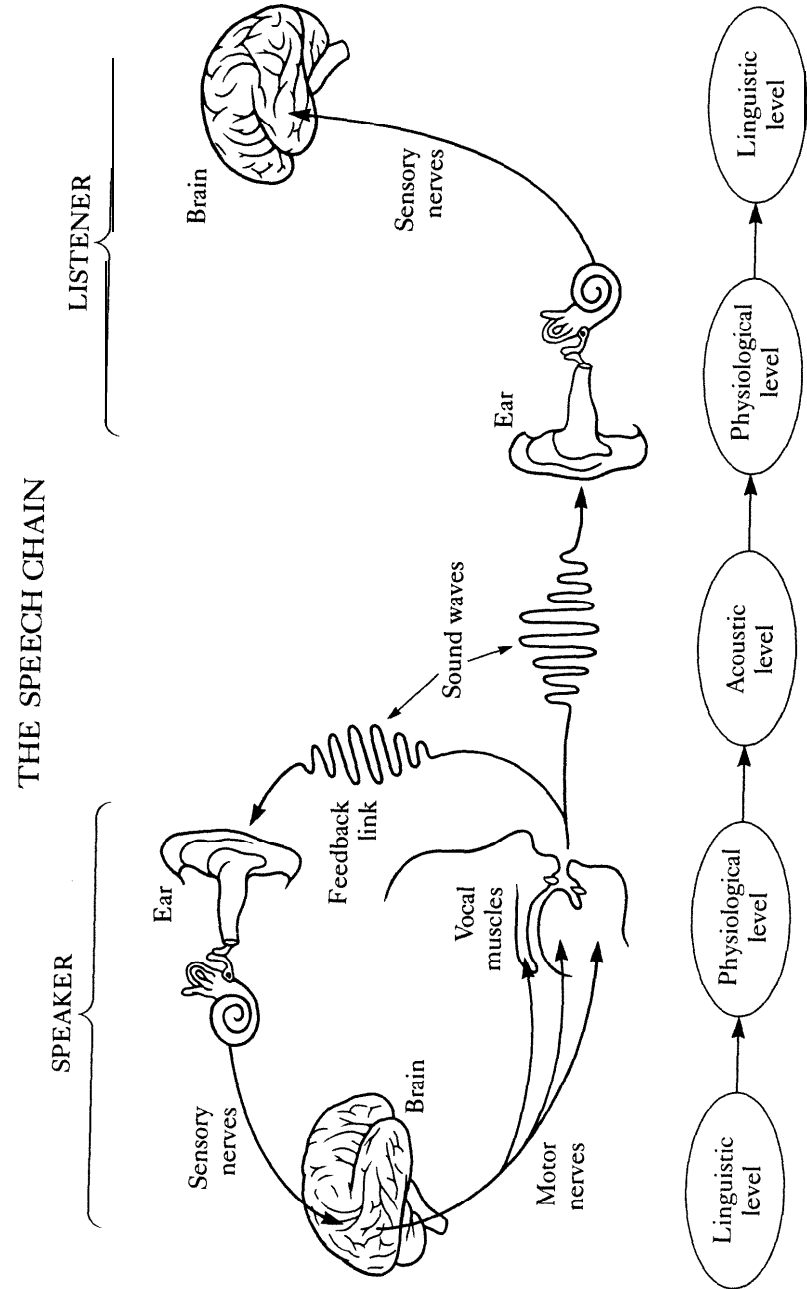


FIGURE 1.1 The speech chain: the different forms of a spoken message in its progress from the brain of the speaker to the brain of the listener.

recognizes the words and sentences transmitted by the speaker. The speech chain, therefore, involves activity on at least three levels—linguistic, physiological and physical—first on the speaker’s side and then at the listener’s end.

We may also think of the speech chain as a communication system in which ideas to be transmitted are represented by a code that undergoes transformations as speech events proceed from one level to another. We can draw an analogy here between speech and Morse code. In Morse code, certain patterns of dots and dashes stand for different letters of the alphabet; the dots and dashes are a code for the letters. This code can also be transformed from one form to another. For example, a series of dots and dashes on a piece of paper can be converted into an acoustic sequence, like “beep-hip-bip-beep.” In the same way, the words of our language are a code for concepts and material objects. The word “dog” is the code for a four-legged animal that wags its tail, just as “dash-dash-dash” is Morse code for the letter “o.” We learn the code words of a language—and the rules for combining them into sentences—when we learn to speak.

During speech transmission, the speaker’s linguistic code of words and sentences is transformed into physiological and physical codes—in other words, into corresponding sets of muscle movements and air vibrations—before being reconverted into a linguistic code at the listener’s end. This is analogous to translating the written “dash-dash-dash” of Morse code into the sounds, “beep-beep-beep.”

Although we can regard speech transmission as a chain of events in which a code for certain ideas is transformed from one level or medium to another, it would be a great mistake to think that corresponding events at the different levels are the same. There is some relationship, to be sure, but the events are far from identical. For example, there is no guarantee that people will produce identical sound waves when they pronounce the same word. In fact, they are more likely to produce different sound waves when they pronounce the same word. By the same token, they may very well generate similar sound waves when pronouncing different words.

This state of affairs was demonstrated experimentally. A group of people listened to the same sound wave, representing a word, on three occasions when the word was embedded in three different

sentences. The listeners agreed that the test word was heard either as “bit” or “bet” or “bat,” depending on which of the three sentences was used.

The experiment clearly shows that the general circumstances (context) under which we listen to speech profoundly affect the specific words we associate with particular sound waves. Put differently, the relationship between a word and a particular sound wave, or between a word and a particular muscle movement or pattern of nerve impulses, is not unique. There is no label on a speech sound wave that invariably associates it with a particular word. Depending on context, we recognize a particular sound wave as one word or another. A good example of this is reported by people who speak several languages fluently. They sometimes recognize indistinctly heard phrases as being spoken in one of their languages, but realize later that the conversation was in another of their languages.

Knowledge of the right context can even make the difference between understanding and not understanding a particular sound wave sequence. You may have listened to announcements made over a loudspeaker in an unfamiliar, noisy place like a bus or subway station. The chances are that many of the words were incomprehensible to you because of noise and distortion. Yet this same speech would be clearly intelligible to regular users of the station, simply because they have more knowledge of the context than you. In this case, the context is provided by their experience in listening under noisy conditions, and by their greater knowledge of the kind of messages to expect.

The strong influence of circumstance on what you recognize is not confined to speech. When you watch television or movies, you probably consider the scenes you see as quite life-like. But pictures on television are much smaller than life-size and those on a movie screen are much larger. Context will make the small television picture, the life-sized original, and the huge movie scene appear to be the same size. Black-and-white television and movies also appear quite life-like, despite their lack of true color. Once again, context makes the multicolored original and the black and white screen seem similar. In speech, as in these examples, we are usually quite unaware of our heavy reliance on context.

We can say, therefore, that speakers will not generally produce identical sound waves when they pronounce the same words on different occasions. Listeners, in recognizing speech, do not rely only on information derived from the speech wave they receive. They also rely on their knowledge of an intricate communication system, subject to the rules of language and speech, and on cues provided by the subject matter and the identity of the speaker.

In speech communication, then, we do not actually rely on precise knowledge of specific cues. Instead, we relate a great variety of ambiguous cues against the background of the complex system we call our common language. When you think about it, there is no other way speech could function efficiently. It does seem unlikely that millions of speakers, with all their different voice qualities, speaking habits and accents, would ever produce anything like identical sound waves when they say the same words. People engaged in speech research know this only too well, much to their regret. Even though our instruments for measuring the characteristics of sound waves are more accurate and flexible than the human ear, we are still unable to build a machine that can recognize speech nearly as effectively as a human being. We can measure characteristics of speech waves with great accuracy, but we do not know the nature and rules of the contextual system against which the results of our measurements must be related, as they are so successfully related in the brains of listeners.

In the following chapters, we will describe the speech chain—from speaker to listener—as fully as current knowledge and the scope of this book allow. What we have said so far should give you some clues as to why only a part of what follows is concerned with the laws governing events on any one level of the speech chain; in other words, with the physics of speech and the behavior of nerves and muscles. The rest of the book, in common with the dominant trends of modern speech research, deals with the relationship of events on different levels of the speech chain, and how the events are affected by context. It describes the kinds of sound waves produced when we speak the speech sounds and words of English; the relationship between the articulatory movements of our vocal organs and the speech wave produced; how our hearing mechanism transforms

sound waves into nerve impulses and sensations; how we perceive speech sound waves as words and sentences. There is also a chapter on the digital processing of speech—a technology used widely today for the study and practical applications of speech and language. The final two chapters deal with the generation of artificial speech and the recognition of speech by computer.

ADDITIONAL READING

G. A. Miller, *The Science of Words*, Scientific American Library, N.Y., 1991

W. S.-Y. Wang (Ed.), *The Emergence of Language: Development and Evolution*, W. H. Freeman, N.Y., 1991