BELL SYSTEM SCIENCE EXPERIMENT No. 3

speech/synthesis

AN EXPERIMENT IN ELECTRONIC SPEECH PRODUCTION

CECIL H. COKER
PETER B. DENES
ELLIOT N. PINSON

Bell Telephone Laboratories

MADE AVAILABLE BY YOUR TELEPHONE COMPANY

SPEECH SYNTHESIS

Composed, Printed, and Bound by
Waverly Press, Inc., Baltimore, Md.

PREPARED BY BELL TELEPHONE LABORATORIES FOR EDUCATIONAL USE

SPEECH SYNTHESIS

An Experiment
in Electronic
Speech
Production

Cecil H. Coker
Peter B. Denes
Elliot N. Pinson

Bell Telephone Laboratories

Preface

The "Speech Synthesis" experiment, of which this book is a part, is one of several educational aids on speech and hearing, made available through the Bell System's High School Science program. Other material includes a classroom-oriented text entitled, *The Speech Chain*, a 19-minute color, sound film of the same title, and four classroom demonstration experiments.

Speech is an important subject in its own right; in addition, it is an excellent example of a topic that can be thoroughly understood only by using the points of view and methods of investigation of several disciplines, such as anatomy, physiology, acoustics, psychology and linguistics.

The "Speech Synthesis" experiment is intended to advance the student's understanding of speech production and recognition. The electronic circuit, if assembled properly, can produce a variety of vowel sounds. Experiments are suggested that demonstrate some of the acoustic and psychological factors involved in speech perception. The text also explains the importance of speech synthesizers as research tools for learning more about the processes of speech and hearing. Finally, we discuss some of the possible uses of speech synthesizers in future communication systems.

The "Speech Synthesis" material is intended primarily for capable biology and physics students at the secondary school level. Since no knowledge of electronics is required for building the synthesizer, the biology student should not be unduly handicapped. The chief requirements are patience and care, for it is not a short task to mount and connect the many components that make up the synthesizer.

An appendix on constructing the synthesizer, written by George R. Frost (who is also responsible for getting the circuit into its present form), provides some useful hints for the student. The authors appreciate his efforts and feel that the appendix is a valuable addition to the book.

The authors are grateful to those who helped in the preparation of this book. We owe a special debt to D. H. Van Lenten, whose fine editing improved its readability. His efforts to coordinate the work of the authors also were invaluable.

Cecil H. Coker
Peter B. Denes
Elliot N. Pinson

Table of Contents

CHAPTER 1 Spoken Communications, Speech Synthesis, and 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 normal functioning of human society.

Wherever human beings live together, they develop a system of talking to each other; even people in the most primitive societies use speech. Speech, in fact, is one of those few, basic abilities—tool making is another—that sets us apart from animals and is closely connected with our faculty for abstract thinking.

Why is speech so important? One reason is that the development of human civilization is made possible, to a great extent, by man's ability to share experiences, to exchange ideas and to transmit knowledge from one generation to another; in other words, his ability to communicate with other men. 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 finger signing 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 developed by man. Unquestionably, however, speech is the system that man found to be far more efficient and convenient than any other.

You may think that writing is a more important means of communication. After all, the development of civilization and the output of printing presses seem to parallel each other, and the written word appears to be a more efficient and more durable means of transmitting intelligence. It must be remembered, however, that no matter how many books and newspapers are printed, the amount of intelligence exchanged by speech is still vastly greater. The widespread use of books and printed matter may very well be an indication of a highly developed civilization, but so is the greater use of telephone systems. Those areas of the world where civilization is most highly developed are also the areas with the greatest density of telephones; and countries bound by social and political ties are usually connected by a well developed telephone system.

We can further bolster our argument that speech has a more fundamental influence than writing on the development of civilization by citing the many human societies that have developed and flourished without evolving a system of reading and writing. We know of no civilization, however, where speech was not available.

Perhaps the best example of the overwhelming importance of speech in human society is a comparison of the social attitudes of the blind to those of the deaf. Generally, blind people tend to get along with their fellow human beings despite their handicap. But the deaf, who can still read and write, often feel cut off from society. A deaf person, deprived of his primary means of communication, tends to withdraw from the world and live within himself.

In short, human society relies heavily on the free and easy interchange of ideas among its members and, for one reason or another, man has found speech to be his 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 constantly 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 enables him to exploit its built-in features 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 in the just-emerging field of man-to-machine communication. We all use automatons, like the dial telephone and automatic elevator, which either get their instructions from us or report back to us on their operations. Frequently, they do both, like the highly complex digital computers used in scientific laboratories. In designing communication systems or "languages" to link man and machine, it may prove especially worthwhile to have a firm understanding of speech, that system of man-to-man communication whose development is based on the experience of many generations.

SPEECH SYNTHESIS

Spoken communication has been studied in a variety of ways. Among these, the use of artificial speech has always held a special place. The production of artificial speech—or *speech synthesis*, as it is often called—is the process of generating speech-like sound waves by mechanical, electronic or other means not involving the use of the human vocal organs. Synthesized speech has been used in experiments to explore the fundamental nature of speech, and in devices developed for more practical ends.

Interest in artificial speech dates back to Ancient Greece. The Greeks endowed their gods with many human characteristics, including the ability to speak. It was only natural that the priests would try to give statues of their gods the human faculty of speech. Since the priests could not produce genuine synthesized speech, they used a concealed speaker and piped his voice to the mouths of the statues through speaking tubes.

There were no significant developments in speech synthesis until the 18th century. At that time, there was widespread interest in spring-operated automatons that simulated the action of living beings. Many automatons were developed: man-like figures that played the flute, a mechanical child that could write, and even an automated duck that drank water, took corn from your hand and "digested" its food by means of hidden chemicals.

Wolfgang von Kempelen, a Hungarian government official of high rank, lived during this age. He built an automaton that played an almost unbeatable game of chess; there was a deception involved, however, because the wonderful contraption concealed a chess-playing midget.

But Kempelen was also an accomplished development engineer. He built an "artificial talker," for example, that was a genuinely non-human, artificial, speech-producing device, not a clever deception like his chess machine.

His artificial talker was based on a surprisingly good understanding of the human speech-producing mechanism. It used bellows (to take the place of the lungs), a vibrating reed (the "vocal cords") and a tube whose shape could be altered by hand, just as the movements of the tongue change the shape of the mouth during speech. His machine could make most speech sounds and utter short phrases.

Interest in speaking machines went beyond stunt-making, however, and there is evidence of a more strictly scientific curiosity in speech production. In all ages, after all, there were people who had to overcome speech defects and people who wanted to speak foreign languages without pronounced accents. To be effective, their teachers had to know what factors were essential for producing different speech sounds. A good way to find out about these essentials was (and is) to build a model of the human vocal mechanism and see how changes in the model affected the sounds produced.

In 1779, the Imperial Russian Academy of St. Peterburg offered its annual competitive prize for the best explanation of the physiological differences involved in producing five vowel sounds; an additional requirement called for the design of an apparatus for producing the same vowels artificially. The prize was won by Christian Gottlieb Kratzenstein. Like Kempelen, Kratzenstein used tubes that approximated the shape and size of the mouth; but unlike Kempelen, Kratzenstein was an inferior showman and his machine failed to attract the attention it deserved.

Half a century later, a Viennese professor, Joseph Faber, built the machine shown in Fig. 1.1. Exhibited first in London, the machine could hold a conversation in normal speech, it

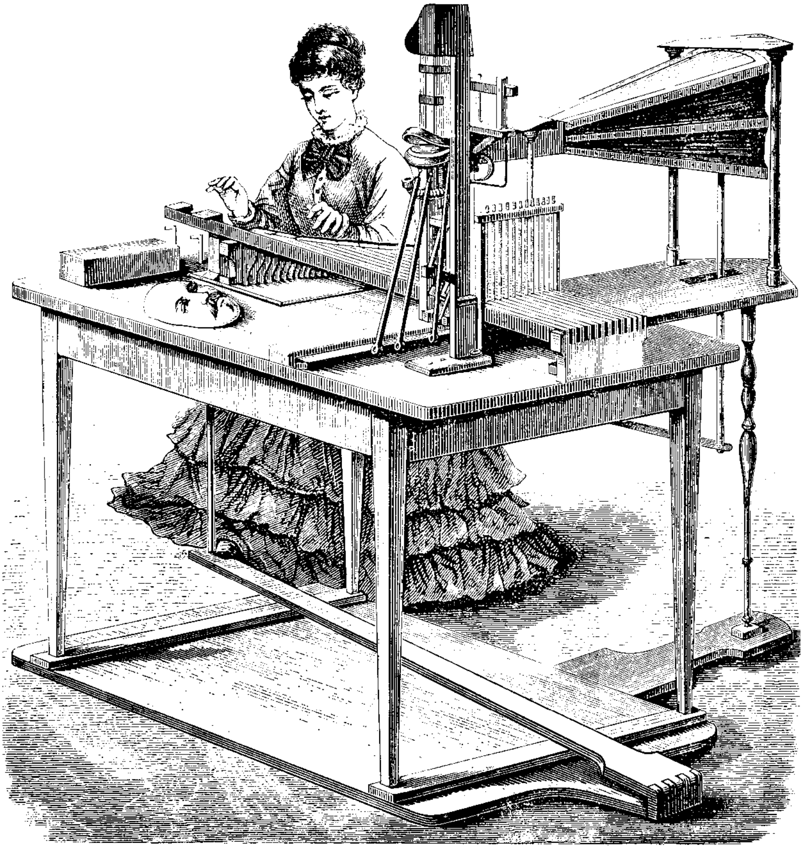


Fig. 1.1 Faber's talking machine, developed around 1850, could hold a conversation in normal speech. A foot-operated lever pumped the bellows at the top right. Air streaming from the bellows activated a vibrating reed whose buzz was modified by the resonances of the tubes on the operator's left side and by the lip-like opening directly above the tubes. The keyboard played by the young lady determined which resonant tubes were in use. The mask face-up on the front of the machine was included to complete the illusion.

could whisper and it could even sing a few tunes, including (Faber had some flair for showmanship) "God Save the Queen."

Sir Richard Paget synthesized intelligible speech in the 1920's. One of his "talking machines" consisted of a simple buzzer and a set of artificial "vocal cavities" he formed by cupping his hands. He could produce a variety of speech sounds by altering the size of the "cups" and by moving some of his fingers to simulate movements of the tongue. There is a story that Paget, sitting in a dentist's chair and unable to speak normally, artificially produced the phrase—"Easy there, you're on a nerve!"—to check his tormentor's enthusiasm for drilling.

Between 1850 and the 1930's, many other synthesizers were produced. They all used the "resonances" of air-filled tubes or mechanical resonators like tuning forks to simulate the resonant characteristics of the human vocal tract.

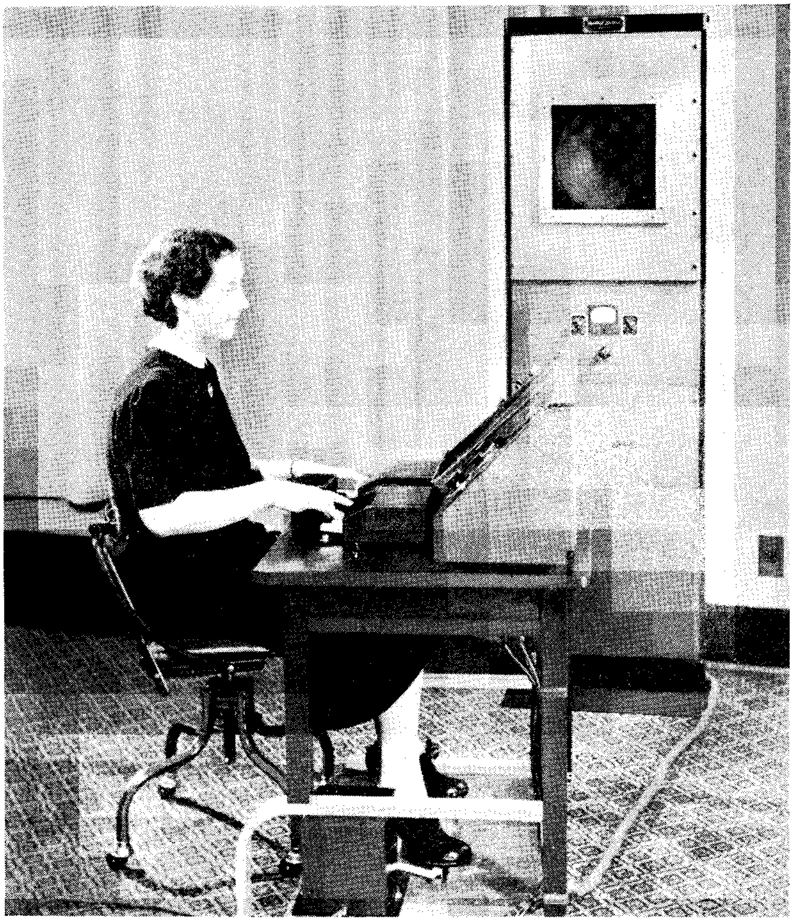


Fig. 1.2 The Voder, an electronic synthesizer, was exhibited at the 1939 World's Fair in New York. The young lady in the photograph could carry on a conversation by pressing keys; 13 keys controlled the speech sound produced and another key controlled volume. Three other keys were used to make the "stop" consonants. A foot-operated pedal gave rising and falling inflections.

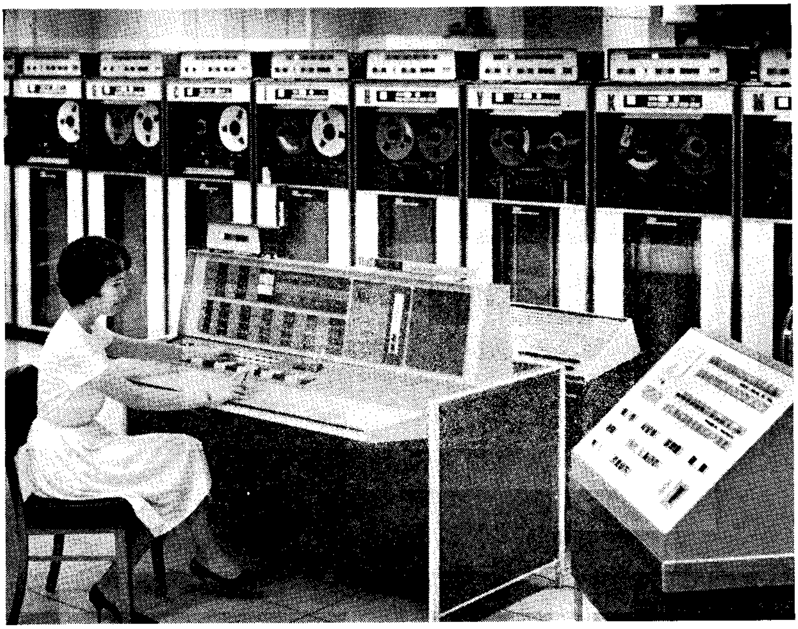


Fig. 1.3 Modern digital computers are used today to synthesize speech. Punched cards are fed to the computer; the computer reads and performs the instructions on these cards and produces a speech-like output on magnetic tape. The sounds are heard when the tape is played over a special tape recorder.

The advent of the vacuum tube brought about other possibilities: electrical circuits could be designed to simulate vocal tract resonances. The first really significant electrical synthesizer was Homer Dudley's *Voder* (shown in Fig. 1.2), an elaborate apparatus controlled by a piano-style key board. The Voder was exhibited at the 1939 New York World's Fair; today, it forms an important part of another device, the *vocoder*, which may one day go into service as a significant advance in electronic speech transmission. (One type of vocoder is described in Chapter 9.)

More recently, digital computers (see Fig. 1.3) have been used to synthesize speech. Computers can be programmed (instructed) to generate synthetic speech in several different ways. We will say more about computer-synthesized speech in Chapter 8. Briefly, punched cards containing instructions are fed to the computer; the computer reads these instructions, makes the necessary calculations and produces a special magnetic tape on which details of the synthesized speech waves are recorded. The sounds are heard when the tape is played back over a special tape recorder.

Electrical circuits and high speed computers offer powerful means for generating speech-like sounds. But what is the value of speech synthesizers? How can they be used to explain the nature of spoken communication, and what practical purpose can they serve? Before answering these questions, we should remember that artificial speech simulates the *sounds* produced by the human vocal apparatus. But there is much more to language than just sound. We can only explain the uses of synthesized speech by considering it in relation to the other events that take place during spoken communication. One more detour is necessary, then, before we get back to the subject of speech production, both normal and synthetic. The detour will consist of a look at the entire process of communication by speech.

THE SPEECH CHAIN

A convenient way of examining what happens during speech is to take the simple situation of two people talking to each other; one of them, the speaker, transmits information to the other, the listener. The first thing the speaker has to do is arrange his thoughts, decide what he wants to say and put what he wants 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 correct 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 of the vocal organs—the tongue, the lips and the vocal cords. The nerve impulses set the vocal muscles into movement which, in turn, produces minute pressure changes in the surrounding air. We call these pressure changes a sound wave.

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, the considerable amount of nervous activity already taking place is modified by the nerve impulses arriving from the ear. This modification of brain activity, in ways we do not fully understand, 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 Fig. 1.4. on page 10.)

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 the speaker not only speaks, he also listens to his own voice. In listening, he continuously compares the quality of the sounds he produces with the sound qualities he intended to produce and makes the adjustments necessary to match the results with his intentions.

There are many ways to show that a speaker is his own listener. 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 earphones. Under such circumstances, the unexpected delay in the fed-back sound makes the speaker stammer and slur. This is the so-called delayed speech feed-back effect. Another example of the importance of feed-back is the general deterioration of the speech of people who have suffered prolonged deafness. Deafness, of course, deprives these people of the speech chain's feed-back link. To some 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 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 level* of the speech chain.

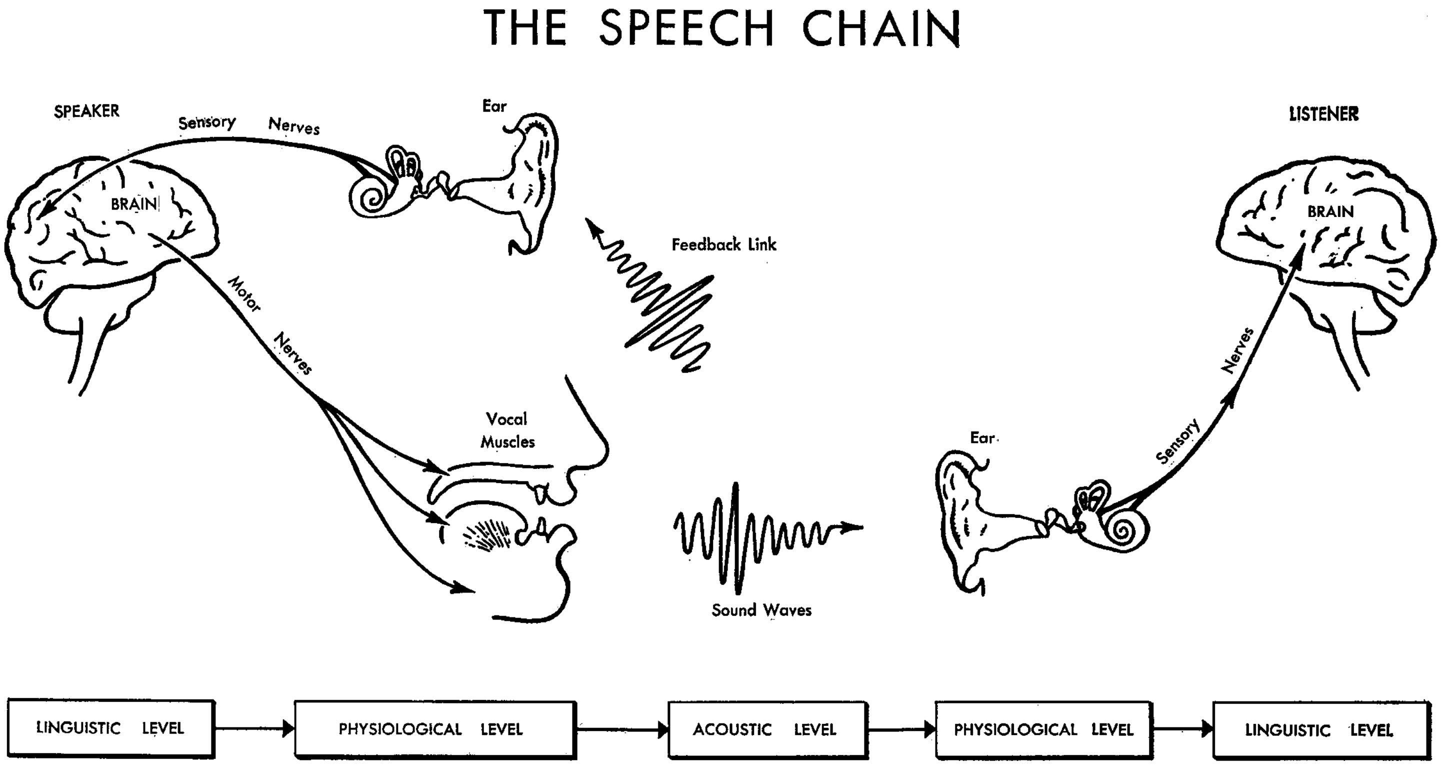


Fig. 1.4 The Speech Chain: the different forms in which a spoken message exists in its progress from the mind of the speaker to the mind of the listener.

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 recognizes the words and sentences transmitted by the speaker. The speech chain, therefore, involves activity on at least three different levels, the 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-bip-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 sound waves with identical characteristics when they pronounce the same word. In fact, they are more likely to produce sound waves of different characteristics 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 clearly demonstrated in an experiment carried out a few years ago. A group of people listened to the same sound wave, representing a word, on three occasions when the word was used in three different-sounding sentences. The listeners agreed that the test word was either "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 kind of words we associate with particular sound waves. In other words, 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 probably know that at some airports you can pay a dime and listen in on the conversations between pilots and the control tower. The chances are that many of the sentences would be incomprehensible to you because of noise and distortion. Yet this same speech wave would be clearly intelligible to the pilots simply because they have more knowledge of context than you. In this case, the context is provided by their experience in listening under conditions of distortion, and by their greater knowledge of the kind of message 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 much larger on a movie screen. 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 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. The listener, in recognizing speech, does not rely solely on information derived from the speech wave he receives. He also relies on his 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, we do not actually rely on a 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 really 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 considerably more accurate and flexible than the human ear, we are still unable to build a machine that will recognize speech. 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.

What we said in the last few pages gave you some insight into the kind of events that affect the operation of the speech chain. You saw that events on the linguistic level form the first and last links of the chain. We will want to know more about events on this level; the next chapter gives a brief description of *linguis-*

tic organization. The chapters after that deal with various aspects of speech production—both natural speech produced by human beings and synthesized speech produced by machines. First, we discuss those physical principles necessary for understanding speech sound waves and their generation. Then, we describe how the human vocal apparatus produces speech sounds, and how the speech synthesizer you will build produces vowel sounds; we also suggest a few experiments you can perform with the synthesizer. After that, we describe more elaborate synthesizers and how they have been used to increase our understanding of spoken communication. Finally, we discuss some applications to which increased knowledge of the speech process may lead.

There are also two useful appendices. The first offers several helpful hints on building your synthesizer; you should read it carefully before you start construction. The second appendix lists some circuit modifications you can make to your synthesizer, either to improve its performance or to make it more convenient to use.

CHAPTER 2 Linguistic Organization

In our discussion of the nature of speech, we explained that the message to be transmitted from speaker to listener is first arranged in linguistic form; the speaker chooses the right words and sentences to express what he wants to say. The information then goes through a series of transformations into physiological and acoustic forms, and is finally reconverted into linguistic form at the listener's end. The listener fits his auditory sensations into a sequence of words and sentences; the process is completed when he understands what the speaker said.

Throughout the rest of this book, we will concern ourselves with relating events on the physiological and acoustic levels with events on the linguistic level. When describing speech production, we will give an account of the type of vocal organ movements associated with speech sounds and words. When describing speech recognition and speech synthesis, we will discuss the kinds of speech sounds and words perceived when we hear sound waves with particular acoustic features. In this chapter, we will concentrate on what happens on the linguistic level itself; we will concentrate, in other words, on describing the units of language and how they function.

The units of language are symbols. Many of these symbols stand for objects around us and for familiar concepts and ideas. Words, for example, are symbols. The word "table" is the symbol for an object we use in our homes, the word "happy" represents a certain state of mind, and so on. Language is a system consisting of these symbols and the rules for combining them into sequences that express our thoughts, our intentions and our experiences. Learning to speak and understand a language involves learning these symbols, together with the rules for assembling them in the right order. We spend much of the first few years of our lives learning the rules of our native language. Through practice, they become habitual and we can apply them without being conscious of their existence.

The most familiar language units are words. Words, however, can be thought of as sequences of smaller linguistic units, the *speech sounds* or *phonemes*. The easiest way to understand the nature of phonemes is to consider a group of words like "heed," "hid," "head" and "had." We instinctively regard such words as being made up of an initial, a middle and a final element. In our four examples, the initial and final elements are identical, but the middle elements are different; it is the difference in these middle elements that distinguishes the four words. Similarly, we can compare all the words of a language and find those sounds that differentiate one word from another. Such distinguishing sounds are called phonemes and they are the basic linguistic units from which words and sentences are put together. Phonemes on their own do not symbolize any concept or object; only in relation to other phonemes do they distinguish one word from another. The phoneme "p," for example, has no independent meaning but, in combination with other phonemes, it can distinguish "heat" from "heap," "peel" from "keel," and so forth.

We can divide phonemes into two groups, vowels and consonants. There are 16 vowels and 22 consonants in English, as listed in Table 2.1.

TABLE 2.1—THE PHONEMES OF GENERAL AMERICAN ENGLISH

General American is the dialect of English spoken in midwestern and western areas of the United States and influences an increasing number of Americans. Certain phonemes of other regional dialects (e.g., Southern, British, etc.) can be different.

Vowels	Consonants	
ee as in h*e*at	*t* as in *t*ee	*s* as in *s*ee
ɪ as in h*i*t	*p* as in *p*ea	*sh* as in *sh*ell
e as in h*e*ad	*k* as in *k*ey	*h* as in *h*e
ae as in h*a*d	*b* as in *b*ee	*v* as in *v*iew
ah as in f*a*ther	*d* as in *d*awn	*th* as in *th*en
aw as in c*a*ll	*g* as in *g*o	*z* as in *z*oo
U as in p*u*t	*m* as in *m*e	*zh* as in mea*s*ure
oo as in c*oo*l	*n* as in *n*o	*l* as in *l*aw
ʌ as in *u*p	*ng* as in si*ng*	*r* as in *r*ed
uh as in th*e*	*f* as in *f*ee	*y* as in *y*ou
er as in b*ir*d	*θ* as in *th*in	*w* as in *w*e
oi as in t*oi*l		
au as in sh*ou*t		
ei as in t*a*ke		
ou as in l*o*ad		
ai as in m*i*ght		

English is the native language of hundreds of millions of people, in many parts of the world. But the English spoken in England is different from that spoken in Australia, which is different from the English spoken in the United States. In the United States alone, many different kinds of English are spoken. The speech of people who live in the South, for example, does not sound the same as the English spoken in New England.

Nevertheless, these different "forms" of English are basically so similar that they can all be called the same language—English, understandable to everyone of us. Of course, the kinds of English spoken in various parts of the world are different: we have no trouble spotting natives of England and Georgia just from the way they talk. We say that they speak different *dialects* of English. The vowels and consonants of different dialects can vary markedly.

The phonemes shown in Table 2.1 refer to the so-called *General American* dialect spoken in midwestern and western areas of the United States. It has an increasing influence on the speaking habits of large numbers of Americans. At the same time, you may find that your own dialect is somewhat different. No single dialect is fundamentally more "English" than any other. Moreover, not all dialects are regional; some are determined by social or cultural factors. For example, in one district of London, the Cockney dialect is spoken, while the more educated classes speak an entirely different "brand" of English which, incidentally, is spoken by the majority of educated people in most parts of Britain. In many other areas of the world, there are popular dialects that are distinguishably different from the dialects spoken by educated men and women. These "educated" dialects are usually taught in schools and spoken by radio and television announcers.

Phonemes can be combined into larger units called *syllables.* Although linguists do not always agree on the definition of a syllable, most native speakers of English have an instinctive feeling for its nature. A syllable usually has a vowel for a central phoneme, surrounded by one or more consonants. In most languages, there are restrictions on the way phonemes may be combined into larger units. In English, for example, we never find syllables that start with an "ng" phoneme: syllables like "ngees" or "ngoot" are impossible. Of course, such rules reduce the variety of syllables used in a language; the total number of English syllables is between only one and two thousand.

An even larger linguistic unit is the *word*, which normally consists of a sequence of several phonemes and one or more syllables. The most frequently used English words are sequences of between two and five phonemes. There are some words, like "awe" and "a," which have only one phoneme, and others that are made up of 10 or more.

Table 2.2—The Ten Most Frequently Used Words in English

I
the
a
it
to
you
of
and
in
he

The most frequently used words are, on the whole, short words with just a few phonemes. This suggests that economy of speaking effort may have an influence on the way language develops. Table 2.2 shows the 10 most frequently used English words.

Only a very small fraction of possible phoneme combinations are used as words in English. Even so, there are several hundred thousand English words, and new ones are being added every day. Although the total number of words is very large, only a few thousand are frequently used. Various language surveys indicate that—95 per cent of the time—we choose words from a library of only 5000 to 10,000 words. The vast number of other words are rarely used.

Words are combined into still longer linguistic units called *sentences*. The rules that outline the way sequences of words can be combined to form acceptable sentences are called the *grammar* of a language. Grammar tells us that the string of words, "the plants are green," is acceptable, but the sequence, "plants green are the," is not.

Grammar alone, however, does not determine word order. Sentences must make sense as well as satisfy the rules of grammar. For example, a sentence like "the horse jumped over the fence" is both grammatically acceptable and sensible. But the sequence, "the strength jumped over the fence," although grammatically correct, is meaningless and does not occur in normal use. The study of word meanings is called *semantics*, and we can see from our two examples that word order is influenced both by grammatical and semantic considerations.

Stress and *intonation* are also part of linguistic organization. They are used to express such things as the speaker's emotional attitude, to make distinctions between questions, statements and doubt, etc., and to indicate the relative importance attached to different words in a sentence. We can, for example, alter the sense of identical sentences simply by using stress and intonation. We can say, "*I* will be the judge of that" or, "I will be the judge of *that*," and although the same words appear in the two sequences, the meanings of the sentences are dissimilar. Stress and intonation are used extensively during speech, but there is really no adequate method of representing them in written material. We can use different types of punctuation, but this is only a partial solution of the problem. In fact, the trouble we occasionally have—when writing—to indicate distinctions quite easy to make in speech by stress and intonation, is a good example of their importance.

We have now seen that the fundamental units of our linguistic system are phonemes, syllables and words. In addition, we have the grammatical and semantic rules for combining these units into longer sequences. Stress and intonation are also important aspects of language. Together, they form the linguistic basis of speech, our most commonly used communication system.

CHAPTER 3 The Physics of Sound

Before we can discuss the nature of speech sound waves—how they are produced and perceived—we must understand a certain amount about sound waves in general. Sound waves in air are the principal subject of this chapter. The subject forms part of the field of *acoustics.* Since our book is concerned with the broad topic of speech synthesis, we will present only a brief introduction to the physics of sound, with emphasis on those aspects that are necessary for understanding the material in following chapters.

Sound waves in air are just one example of a large class of physical phenomena that involve *wave motion.* Surface waves in water and electromagnetic radiations, like radio waves and light, are other examples. All wave motion is produced by—and consists of—the vibration of certain quantities. In the case of sound waves, air particles are set into vibration; in the case of surface waves in water, water particles; and in the case of electromagnetic waves, the electrical and magnetic fields associated with the wave oscillate rapidly. Since vibrations play such an important part in wave motion, we will begin by explaining a few elementary facts about them.

VIBRATION

Perhaps the best way to approach the subject of vibration is in terms of a simple example. There are many to choose from, such as the vibrating prongs of a tuning fork, an oscillating piano string, a pendulum, or a spring and mass.

Let us examine the spring and mass arrangement shown in Fig. 3.1. One end of the spring is rigidly fixed and cannot move; the other end is attached to the mass, say a metal block. The mass rests on a surface it can easily slide along. When the mass is in its normal resting position, the pointer attached to it is at position *B* on the ruler.

If the mass is moved toward point A, the spring will be compressed and will exert a force on the mass that tends to move it back toward its rest position. If the mass is moved in the other direction, toward point C, the spring will be stretched; again, a force will act on the mass, tending to make it move toward its rest position, B. We see, then, that the spring always exerts a "restoring force" that tends to move the mass toward its rest position.

Suppose we displace the mass, say to point A, and release it. The spring force will make the mass move toward B. It will gain speed until it reaches point B and, because of its *inertia,* will pass through its rest position. Inertia, a property common to all matter, causes a body in motion to remain in motion (or a body at rest to remain at rest) in the absence of external forces. Once the mass is on the right-hand side of its rest position, the spring's restoring force opposes its motion and, eventually, brings it to a stop. The mass is again set in motion by the spring force acting in the direction of the rest position; it will pass through its rest position and continue to move back and forth. This to and fro motion of a body about its rest position is called *oscillation* or *vibration.*

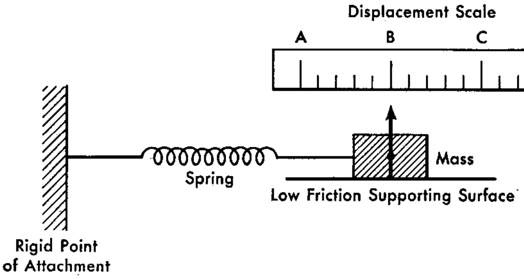


Fig. 3.1 A simple spring-mass oscillator.

Vibrations are likely to occur whenever the properties of *mass* and *elasticity* ("springiness") are present together. In air, the individual molecules are the masses of the system. The forces that act between these molecules behave very much like spring forces. For example, if we try to pack an excess number of molecules into a limited volume, a force arises that tends to resist the compression. This is the force that keeps a balloon or tire inflated and opposes our efforts to inflate a bicycle tire with a hand pump. These forces resemble spring behavior.

PROPERTIES OF VIBRATING SYSTEMS

All types of vibration have certain basic properties in common. We will define these properties, using as our example the spring-mass system of Fig. 3.1. These definitions apply to all vibratory motions and will be extensively used later in connection with sound waves. First, we will describe what we mean by the *amplitude*, *frequency* and *period* of a vibration.

If the mass is displaced from its rest position and allowed to vibrate, it moves back and forth between two positions that mark the extreme limits of its motion. The distance of the mass from point *B* at any instant is called its displacement. The maximum displacement is called the *ampli ude* of the vibration. If there are no energy losses during the motion—due to friction, for example—the maximum displacement of the mass will be the same on both sides of its rest position. Furthermore, the size of the displacement will be the same each successive time the mass moves out to the extremes of its motion.

The movement of the mass from *A* to *C* and back to *A* is called one *cycle* of oscillation. The number of complete cycles that take place in one second is called the *frequency* of the oscillation. If 15 complete cycles occur in one second, we say that the vibration has a frequency of 15 cycles per second (abbreviated, cps). The sound waves we will be interested in have frequencies ranging from tens to thousands of cycles per second.

The time taken to complete one cycle of vibration is called the *period* of the vibration. There is a simple relationship between the frequency of an oscillation and its period. The frequency is simply *1 divided by the period*; for example, if the period is 1/20 second, the frequency is 20 cps.

So far, we have more or less assumed that, once set into motion, the spring-mass combination would continue to vibrate indefinitely with the same amplitude. This type of motion is displayed graphically in Fig. 3.2(a). Here, we show the motion of a spring-mass system that vibrates with a period of two seconds. Initially, the mass is displaced a distance of one inch and released. After its initial displacement, the mass continues to move back and forth between the extremes of its displacement, one inch on either side of its rest position. Consequently, the amplitude of vibration is one inch.

In actual fact, the amplitude of vibration will steadily decrease because of energy losses in the system (due to friction, etc.). Vibrations whose amplitudes decay slowly are said to be lightly "damped," while those whose amplitudes decay rapidly are heavily "damped." Figs. 3.2(b) and 3.2(c) show damped oscillations; the damping is greater in Fig. 3.2(c).

We will find that the pressure variations that correspond to many interesting acoustic signals—speech waves, for example—are much more complex than the simple shape shown in Fig. 3.2(a). Nonetheless, we frequently find it convenient to discuss vibrations of this particular form; they are called *sinusoidal*

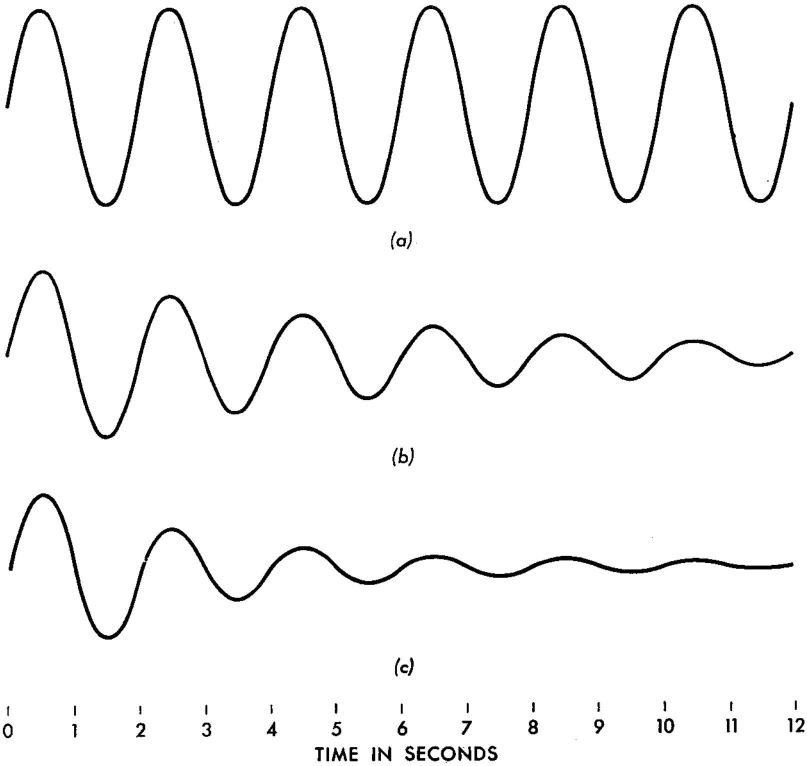


Fig. 3.2 Displacements of the vibrating mass with and without damping: (a) no damping; (b) lightly damped; (c) more heavily damped.

vibrations. The displacement of the mass in our spring-mass system is one example of sinusoidal motion; the movement of a simple pendulum is another. This sort of variation of a quantity with time has important mathematical properties that entitle it to special consideration, as we will see later in this chapter.

FREE AND FORCED VIBRATIONS

So far, we have considered only one way of setting our spring-mass system into vibration: displacing it from its rest position and leaving it free to oscillate without any outside influence. This type of motion is called a *free vibration.* Another way of setting the mass in motion is shown in Fig. 3.3. Here, instead of keeping the left end of the spring fixed, we move it backwards and forwards by using an external force. The mass will now move in a *forced vibration.*

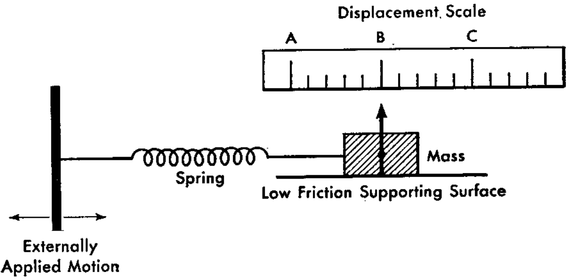


Fig. 3.3 Forced vibration of the spring-mass oscillator.

In free vibration, for a given mass and spring, the mass will always vibrate sinusoidally (with some damping), and the frequency of the oscillation will always be the same. This characteristic frequency is called its *natural* or *resonant* frequency.

The movement of the mass during a forced vibration depends upon the particular way we move the left-hand end of the spring. In what follows, we will assume that the "driving" motion is a sinusoidal displacement. In this case, the motion of the mass is also sinusoidal. Furthermore, the frequency of vibration of the mass is the same as the frequency of the driving motion.

RESONANCE AND FREQUENCY RESPONSE

If the mass is set into free vibration, in the way previously discussed, the amplitude of the oscillation is determined by the size of the initial displacement. It can be no larger than the initial displacement, and it will decay slowly because of losses in the system. In forced vibration, for a given spring-mass combination, the amplitude of the vibration depends on both the *amplitude* and the *frequency* of the motion impressed on the free end of the spring. For a given amplitude of forcing motion, the vibration of the mass is largest when the driving frequency equals the natural frequency of the system. This phenomenon, whereby a body undergoing forced vibration oscillates with greatest amplitude for applied frequencies near its own natural frequency, is called *resonance.* The frequency at which the maximum response occurs is called the *resonant* frequency, and it is the same as the system's natural frequency.

We can show graphically the amplitude with which the mass oscillates in response to a driving motion of any frequency. Such a graph is called a *frequency response* curve. Two frequency response curves are shown in Fig. 3.4. The horizontal axis shows the frequency of the driving motion. The vertical axis shows the amplitude of the response (the motion of the mass) for a constant amplitude of applied motion. At one cps—the natural

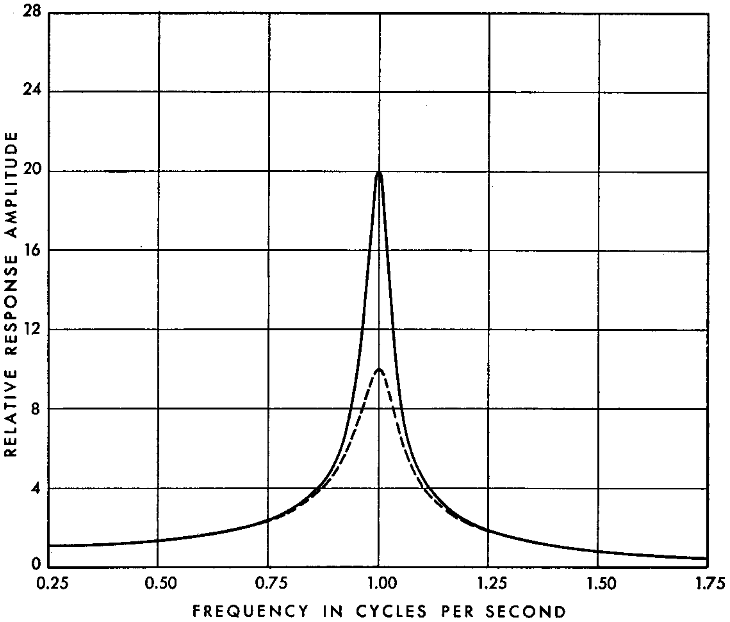


Fig. 3.4 Two frequency response curves with one cps resonant frequency. Dashed line shows oscillation with greater damping.

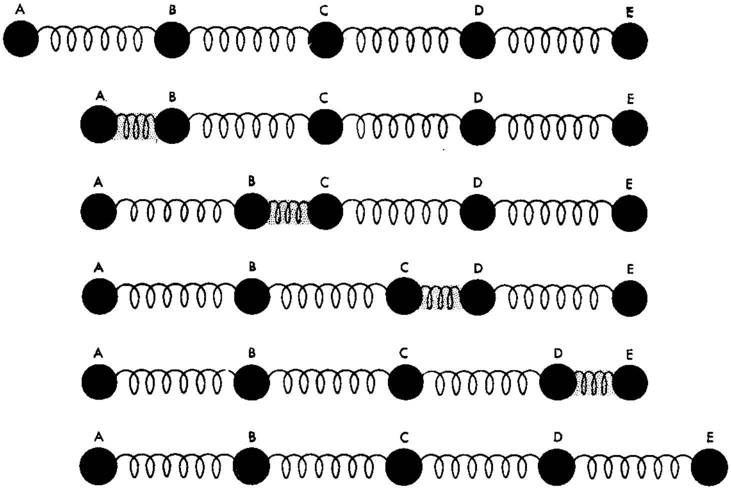


Fig. 3.5 The p1opagation of a compression along the particles of a medium.

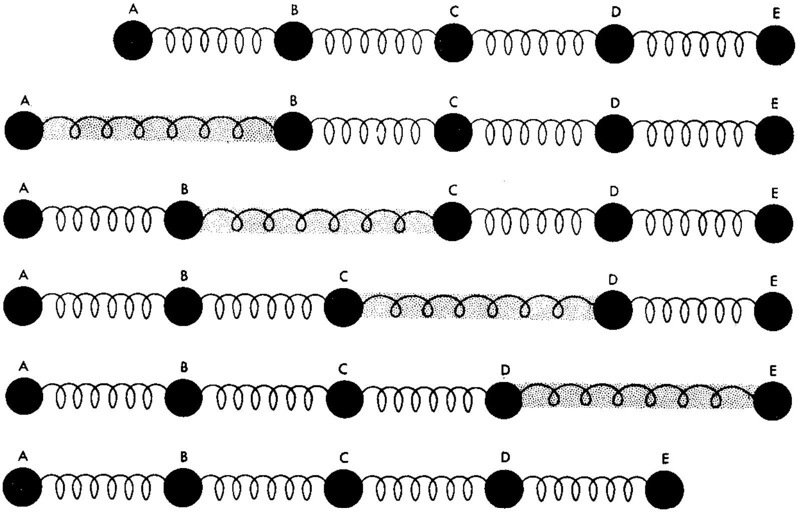


Fig. 3.6 The propagation of a rarefaction along the particles of a medium.

frequency of the vibrating body in our example—the response is much larger than the applied motion. This is due to the phenomenon of resonance. The curves in the figure show the behavior of two oscillators having the same natural frequency, but different damping (different amounts of energy loss). The smaller the energy losses, the greater the increase in movement produced by resonance.

SOUND WAVES IN AIR

All objects on earth are surrounded by air. Air consists of many small particles, more than 400 billion billion in every cubic inch. These particles move about rapidly in random directions. We can explain the generation and propagation of most sound waves without considering such random motions. It is sufficient to assume that each particle has some average "stable" position from which it is displaced by the passage of a sound wave.

If one particle is disturbed—moved nearer some of the others—a force develops that tends to push it back to its original position. Thus, when air is compressed, pushing the particles closer together, a force develops that tends to push them apart. By the same token, when air particles are separated by more than the usual distance, a force develops that tends to push them back into the emptier, rarefied space.

The air particles, in fact, behave just as though they were small masses of matter connected by springs. A line of such particles is shown in the top row of Fig. 3.5. If we push particle *A* toward the right, the "spring" between particles *A* and *B* is compressed. The spring's increased force will move particle *B* to the right, in turn increasing the force on the spring between *B* and *C*, and so forth. Whenever particles near a certain point are closer together than normal, we say that a state of *compression* exists at that point. The positions of the particles at successive instants of time are shown in the successive rows of Fig. 3.5. We see that the compression, which started at the left, moves along the line of particles toward the right. Similarly, if we push particle *A* to the left, we stretch the spring between *A* and *B*; the spring tension will move particle *B* to the left, stretching the spring between *B* and *C*, and so forth. Whenever particles are forced further apart than normal, a state of

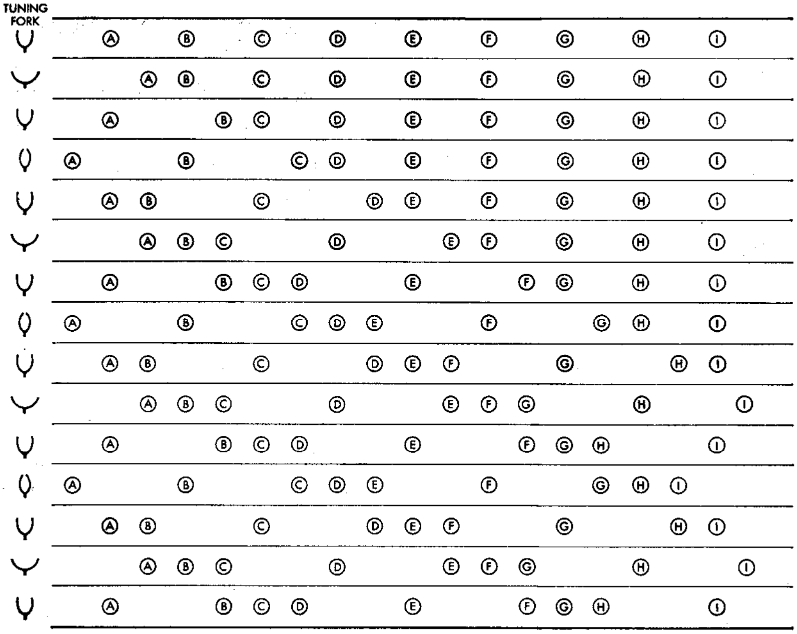


Fig. 3.7 The propagation of a wave along the particles of a medium.

rarefaction is said to exist in their vicinity. Fig. 3.6 shows that once particle *A* has been moved to the left, the resulting rarefaction moves toward the right, from particle to particle.

Suppose we have an oscillating tuning fork near particle *A*, as shown in Fig. 3.7. Consider what happens when the prong nearest particle *A* alternately moves it right and left. Each time the prong moves to the right, a compression wave is sent along the particle line; whenever the prong moves to the left, a rarefaction follows the compression wave. The prong moves once to the right and once to the left during each cycle of vibration; consequently, we get a compression followed by a rarefaction along the particle line for every cycle of vibration.

We see that all the particles go through the same back and forth motion as the tuning fork, but that the movement of each particle lags slightly behind the movement of the preceding particle. We also see that only the disturbance itself (the vibration) moves along the line of particles, and that the air particles move back and forth only about their fixed resting positions. A *sound wave* is the movement (propagation) of a disturbance through a material medium such as air, without permanent displacement of the particles themselves.

A simple demonstration can be set up to show that a sound vibration cannot be transmitted in the absence of a material medium. An electric buzzer is placed under a glass jar, as shown in Fig. 3.8. We can *see* and *hear* the buzzer vibrating. If we now pump the air out of the glass jar, we can *see* that the buzzer continues to vibrate, but the sound we *hear* becomes weaker and weaker as more and more air is removed until, finally, it is inaudible. The sound will be heard again when air is readmitted to the jar.

Surface waves on water exhibit some of the characteristic features of sound waves. Water waves are vibrations of water particles, much as sound waves are vibrations of air particles. The chief difference between the two is that in sound waves, the air particles vibrate in the direction of wave movement, while in surface waves, the water particles principally move up and down, at right angles to the direction of wave movement. Instead of the compressions and rarefactions peculiar to sound waves, water waves appear as crests and troughs on the surface of the water.

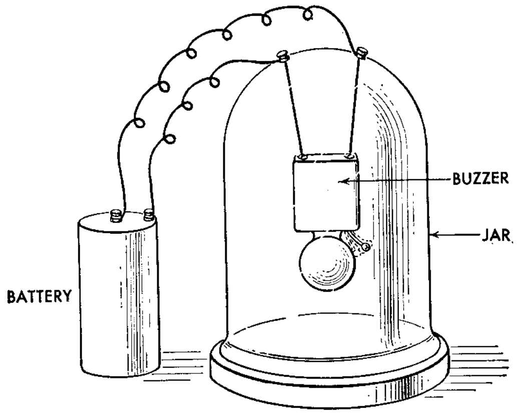


Fig. 3.8 A demonstration showing that sound cannot be transmitted in the absence of a material medium.

THE FREQUENCY AND VELOCITY OF A SOUND WAVE

The frequency at which air particles vibrate (the same as the frequency of the sound source) is called the frequency of the sound wave. We can normally hear sound waves whose frequencies lie between 20 and 20,000 cps. Sound waves at much higher frequencies do exist, but they are inaudible to man. Bats, for instance, use very high frequency sound waves to locate their prey, much as we use radar to pick up targets.

The speed at which the vibrations propagate through the medium is called the *velocity* of the wave. We can determine this velocity in water surface waves by observing the movement of a wave crest. Water waves move slowly, only a few miles an hour. Sound waves in air travel much faster, about 1130 feet per second at sea level; this corresponds to some 770 miles per hour.

How far does the wave travel during one cycle of vibration? We can turn to our tuning fork again. As the fork vibrates, it sends compression after compression along the air particles. The first compression is generated and travels away from the tuning fork; one cycle of vibration later, the fork generates a second compression. By the time the second compression is generated, the first compression has moved further away; the distance between the two compressions is the distance the wave has traveled during one cycle of vibration. The distance between two successive compressions (or between two water wave crests) is called one *wavelength.* A wavelength is also the distance the wave travels in one cycle of vibration of the air particles. If there are f cycles in one second, the wave will travel a distance of f wavelengths in one second. Since the distance traveled in one second is the velocity, it follows that the velocity is equal to the product of the frequency and the wavelength. The wavelength of a sound wave whose frequency is 20 cps is about 56 feet. If the frequency is increased to 1000 cps, the wavelength becomes shorter —about fourteen inches; at 20,000 cps, the wavelength is slightly less than three-quarters of an inch.

We have already said that every air particle in a sound wave vibrates the same way, except for the time lag between the movements of successive particles. The way a particle vibrates, then, is an important characteristic feature of a sound wave. We can plot the displacement of a particle from its rest position, instant by instant. In sound wave measurement, however, it is usually convenient to measure and plot the sound pressure variations associated with the wave, and not the particle displacement itself. The form of such a curve is called the *waveshape.*

THE SPECTRUM

So far, we have considered only sound waves generated by tuning forks. Tuning forks vibrate sinusoidally and, consequently, the waveshape of the corresponding sound wave is also sinusoidal. Sound waves generated by our vocal organs, however, are almost never sinusoidal. In later chapters, we will see several examples of speech waveshapes; in this chapter, we give only two examples. Fig. 3.9 shows the typical waveshape of the sound "ah," and Fig. 3.10 shows the waveshape of the sound "sh." Although the waveshape in Fig. 3.9 is complicated, it clearly consists of repetitions of the same basic shape. In Fig. 3.10, on the other hand, there are no such repetitions. The repetitive wave of Fig. 3.9 is called a *periodic wave*; Fig. 3.10 shows an *aperiodic* wave. Strictly speaking, only waves with an infinite number of repetitions are periodic. But, in practice, many speech sound waves have enough repetitions to be regarded as periodic.

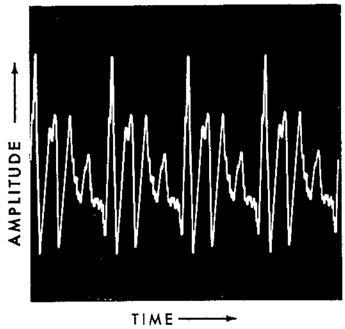


Fig. 3.9 An example of a periodic wave (a typical waveshape of the speech sound "ah").

The waveshapes of Figs. 3.9 and 3.10 are extremely complicated and seem difficult to describe. Fortunately, Joseph Fourier, a French mathematician of the 19th century, showed that any non-sinusoidal wave, no matter how complicated, can be

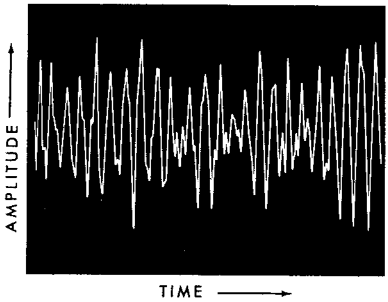


Fig. 3.10 An example of an aperiodic wave (a typical waveshape of the speech sound "sh").

represented as the sum of a number of sinusoidal waves of different frequencies, amplitudes and phases. (The phases of the sinusoidal waves refer to their relative timing—whether they reach the peaks of their vibrations at the same time, for example.) Each of these simple sinusoidal waves is called a *component.*

Fourier's results have been of great importance in analyzing many physical phenomena, not only sound; in fact, they were originally derived in connection with problems about heat flow in material bodies.

The *spectrum* of the speech wave specifies the amplitudes, frequencies and phases of the wave's sinusoidal components.

The illustrations in Fig. 3.11 show that the sum ot many sinusoidal waves is the equivalent of a wave with a non-sinusoidal shape. The frequencies of the sinusoidal waves in Figs. 3.11(a) and (b) are, respectively, five and three times the frequency of the wave in Fig. 3.11(c). When these three waves are added together—just by adding the displacements of all three, instant after instant—we get the clearly non-sinusoidal wave of Fig. 3.11(d). Notice that the basic pattern of the non-sinusoidal wave repeats with the same periodicity as the lowest frequency component (Fig. 3.11(c)) of all components added.

Figs. 3.12(a) to (c) show the same sinusoidal components as Figs. 3.11(a) to (c), but the phase of the component in Fig. 3.12(c) is different from the phase of the component in Fig. 3.11(c). The sum of the three components is shown in Fig. 3.12(d). We notice that the phase-change alters the waveshape of the resulting wave. This shows that we can get a variety of wave-

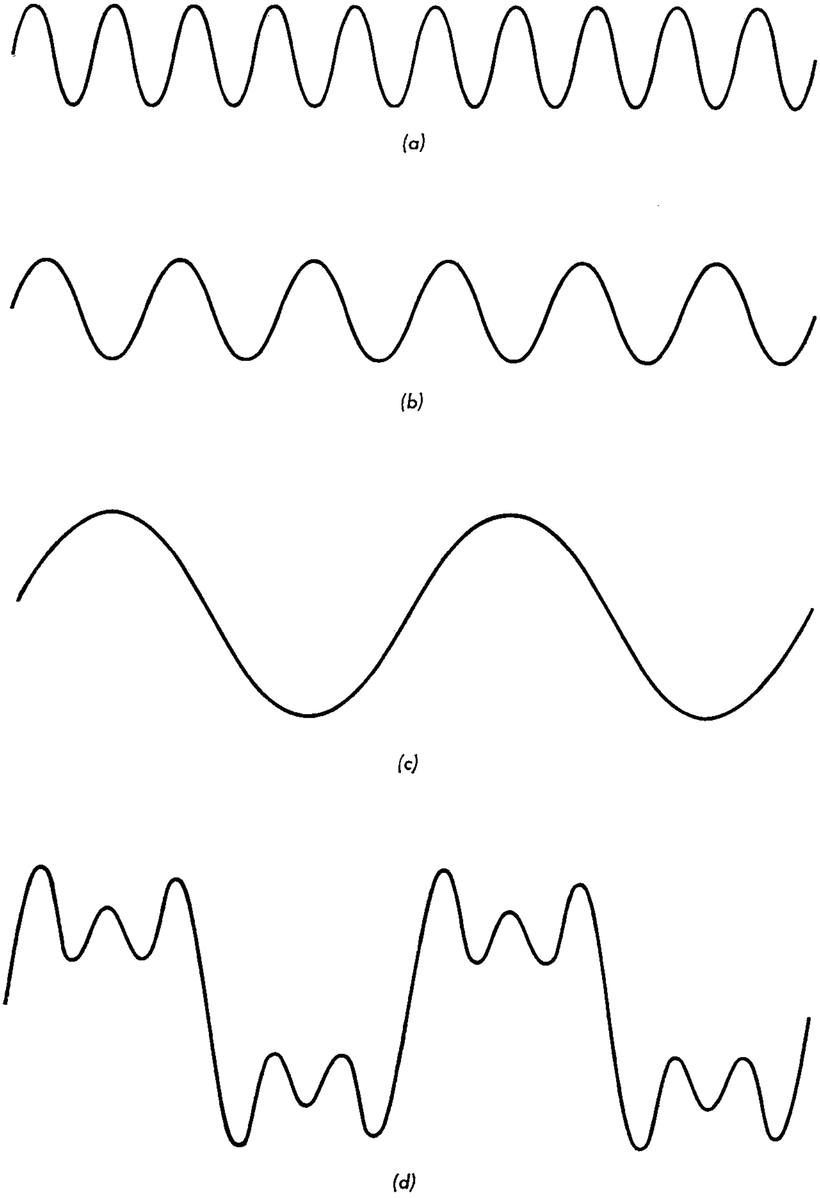


Fig. 3.11 Building up a complex wave: (a), (b) and (c) are sinusoidal components of different frequencies. Portion (a) has five times and portion (b) three times the frequency of portion (c). Portion (d) is the non-sinusoidal sum of (a), (b) and (c).

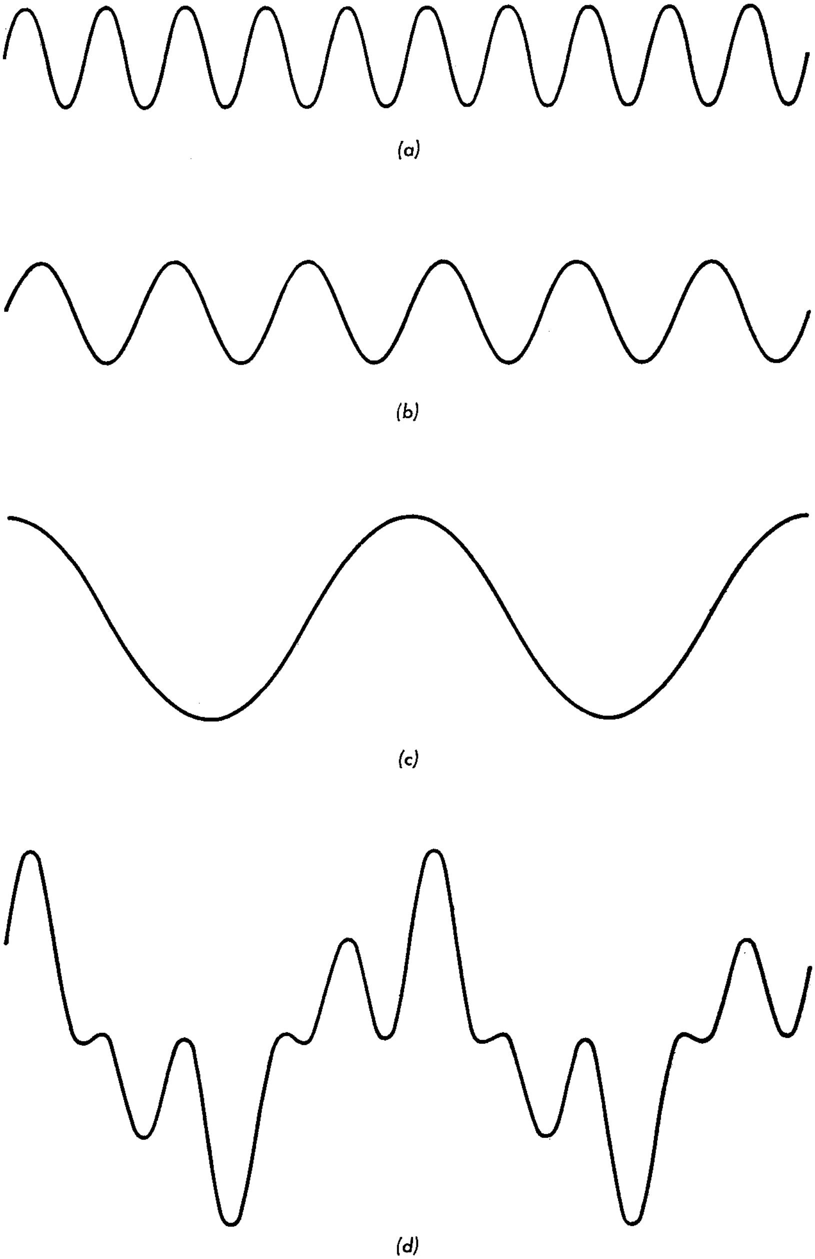


Fig. 3.12 The same component waves as shown in Fig. 3.11, but with the phase of Fig. 3.11 (c) changed; the resulting complex waveform has changed as shown in the (d) portion of this figure.

shapes by adding sinusoidal components of the same amplitudes and frequencies, but of different phases. However, our hearing mechanism cannot always detect the effect of such changes. Non-sinusoidal waves, consisting of sinusoidal waves with the same amplitudes and frequencies, often sound the same, even if their waveshapes differ because of differences in the phase relationship of their components. For this reason, we usually consider only the "amplitude" spectrum of the non-sinusoidal wave, and not its "phase" spectrum. The amplitude spectrum specifies just the frequencies and amplitudes of the sinusoidal components. In the rest of this book, we will use the term "spectrum" to refer to the amplitude spectrum alone.

Basically, we can distinguish two different types of speech wave spectra. One arises from periodic waves and the other from aperiodic waves.

For periodic waves (like the one in Fig. 3.9), the frequency of each component is a whole-number (integer) multiple of some lowest frequency, called the *fundamental frequency*. The component whose frequency is twice the fundamental frequency is called the *second harmonic*; the component three times the fundamental frequency is called the *third harmonic*, and so forth. The spectrum is usually represented by a graph, such as the one shown in Fig. 3.13. Each sinusoidal component is represented by a vertical line whose height is proportional to the amplitude of the component. It is drawn in a position along the frequency scale—marked at the bottom of the graph—cor-

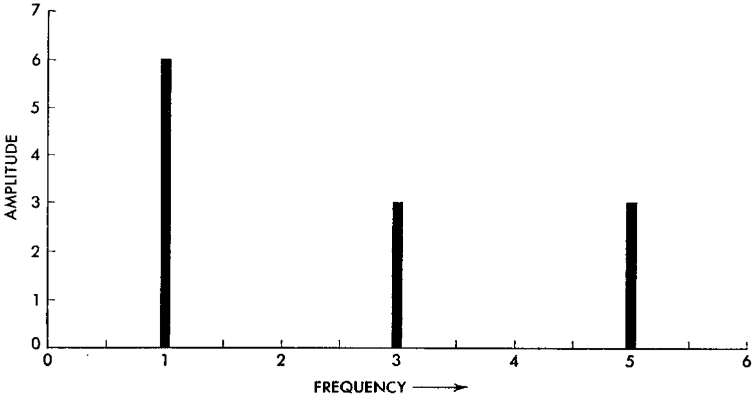


Fig. 3.13 The spectrum of the complex waves shown in Figs. 3.11 (d) and 3.12 (d).

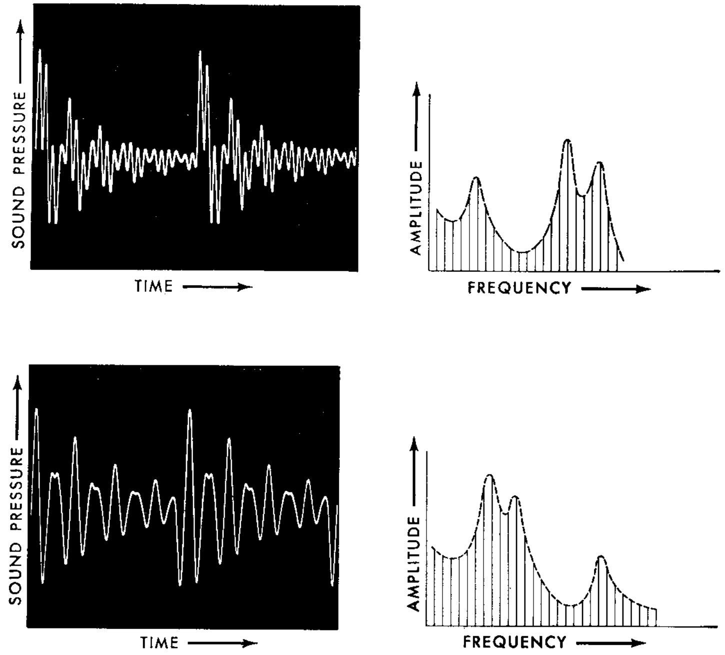


Fig. 3.14 The waveshapes and corresponding spectra of the vowels "uh" (top) and "ah" (bottom).

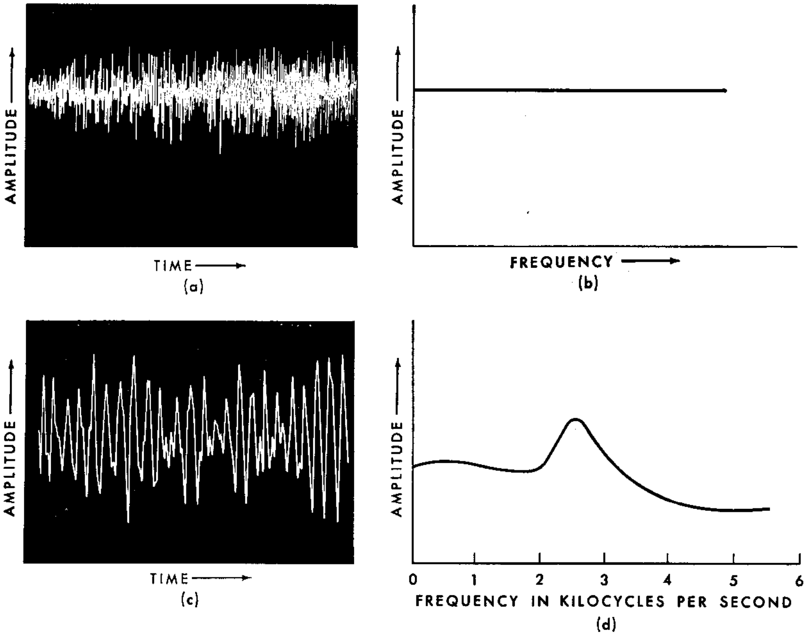


Fig. 3.15 The waveshapes and corresponding spectra of two different aperiodic waves.

responding to the frequency of the component it represents. The higher the frequency of a component, the farther to the right we draw the corresponding line. The spectrum shown in Fig. 3.13 relates to the waves of Figs. 3.11(d) or 3.12(d);consequently, the spectrum is made up of three components. Other waveshapes and their corresponding spectra are shown in Fig. 3.14.

Aperiodic waves can have components at all frequencies, rather than only at multiples of a fundamental frequency. Thus, we no longer draw a separate line for each component, but a single curve. The height of this curve—at any frequency—represents the energy in the wave near that frequency. Fig. 3.15(a) shows a typical aperiodic wave; Fig. 3.15(b) is the corresponding spectrum. It is a horizontal line, indicating that all the spectral components of this wave have the same amplitude. The waveshape of Fig. 3.15(c)—the waveshape of a typical "sh" sound (also shown in Fig. 3.10.)—is another example of an aperiodic wave. Its corresponding spectrum, Fig. 3.15(d), has a peak around 2500 cps; this indicates that, of its many components, those in this region are larger in amplitude than the other components.

We have seen that sound waves of any waveshape can be regarded as the sum of a number of waves with simple, sinusoidal shapes. This helps us deal with speech sound waves, which have a great variety of highly non-sinusoidal (complex) waveshapes. In fact, the method is so convenient that we seldom consider the waveshape of the speech wave; that is, we seldom consider the way the sound pressure or deflection of air particles varies with time. Instead, we think in terms of the corresponding spectrum.

We have a convenient and easy-to-use instrument that can measure and display the spectrum of sound waves applied to it. This is the so-called *sound spectrograph*, to be described in Chapter 8.

SOUND PRESSURE AND INTENSITY

Sound Pressure So far in our description of vibration and wave motion, we have been concerned with the movement of air particles; in other words, with their displacements from their rest positions. The air particles are moved by an external force—like the force exerted by the prongs of a vibrating tuning fork—and each particle exerts force on adjacent particles. The unit of force used most of the time in acoustics is the *dyne.* If you put a one gram mass—about 1/30 of an ounce—on the palm of your hand, the gravitational force that tends to push the mass down is equal to about 1000 dynes. *Pressure* is the amount of force acting over a unit area of surface, and the unit of pressure used here is the *dyne per square centimeter.* If, for example, the area of contact between your hand and the one gram mass is two square centimeters, we say that the mass exerts a pressure of about 500 dynes per square centimeter. Normal atmospheric pressure is equal to about one million dynes per square centimeter. In practice, we frequently use units larger than the dyne. For example, we measure tire pressure in pounds per square inch; a tire pressure of 10 pounds per square inch corresponds to a pressure of about 700,000 dynes per square centimeter. The pressures that move air particles—to produce sound waves—are very small. The smallest pressure variation sufficient to produce an audible sound wave is equal to about 0.0002 dynes per square centimeter. At the other end of the scale, sound pressures of 2000 dynes per square centimeter produce sound waves that are not only extremely loud, but strong enough to cause serious damage to the ear.

Sound Intensity

When we push against a heavy stone and move it, we do work or—looking at it another way—we expend energy. When the prongs of a vibrating tuning fork push against an air particle and move it, work is done and energy is expended. Work done is equal to the force exerted on an object, multiplied by the distance the object is moved. A frequently used unit of work and energy is the *erg.* One erg is the amount of work done when a one dyne force displaces an object by one centimeter. Frequently, we are interested in the amount of work done in a given time or, put another way, the rate of doing work. *Power* is the amount of work done in a given time; its unit is the *erg per second.* Since this is much too small for practical use, we normally reckon power in watts or in horsepower. One watt equals 10 million ergs per second, and one horsepower is equal to 746 watts.

In moving surrounding air particles, a vibrating tuning fork transfers a certain amount of energy to them. The air particles, in turn, transfer this energy to more and more air particles as the sound wave spreads out in all directions. The number of air particles affected by the vibrating tuning fork increases with distance from the source; consequently, the amount of energy available to move a particular air particle decreases with distance from the tuning fork. This is why a tuning fork sounds fainter as we move away from it. In measuring the energy levels of sound waves, we are often not interested in the total energy generated by the vibrating source, but only in the energy available over a small area at the point of measurement. The power transmitted along the wave—through an area of one square centimeter at right angles to the direction of propagation—is called the intensity of the sound wave. It is measured in watts per square centimeter. A sound intensity of 10^{-16} watts per square centimeter (one ten thousand million millionths of one watt per square centimeter) is sufficient to produce a just audible sound; a sound energy of one-hundredth of a watt per square centimeter can damage the ear.

THE DECIBEL SCALE

Most quantities are measured in terms of fixed units. For example, when we say the distance between two points is 20 meters, we mean that the distance between the points is 20 times greater than a one meter length (the reference length of a particular metal rod kept under controlled conditions in Paris). Similarly, when we measure sound intensity in terms of watts per square centimeter, we assume a reference unit of one watt per square centimeter.

Most times, however, it is more convenient to measure sound intensities along the *decibel scale*. Decibels (abbreviated, dB) are not fixed units like watts, grams and meters. When we say that the intensity of a sound wave is one decibel, we mean only that it is a certain number of times greater than some other intensity (about 1.25 times greater). The correct statement is that the sound intensity is one decibel relative to some intensity or another; for example, relative to one watt per square centimeter.

Table 3.1 — Intensity Ratios and Their Decibel Equivalents

Intensity Ratio	Decibel Equivalent
1:1	0
10:1 (the same as 10^1:1)	10
100:1 (the same as 10^2:1)	20
1000:1 (the same as 10^3:1)	30
10000:1 (the same as 10^4:1)	40
100000:1 (the same as 10^5:1)	50
1000000:1 (the same as 10^6:1)	60
10000000000:1 (the same as 10^{10}:1)	100
100000000000000:1 (the same as 10^{14}:1)	140
2:1	3
4:1 (the same as 2 times 2 to 1)	6 (the same as 3 + 3)
8:1 (the same as 4 times 2 to 1)	9 (the same as 6 + 3)
400:1 (the same as 4 times 100 to 1)	26 (the same as 6 + 20)
0.1:1 (the same as 10^{-1}:1)	−10 (minus 10 dB)
0.01:1 (the same as 10^{-2}:1)	−20
0.4:1 (the same as 0.1 times 4 to 1)	−4 (the same as −10 + 6)

Table 3.2 — Sound Pressure Ratios and Their Decibel Equivalents

Sound Pressure Ratio	Decibel Equivalent
1:1	0
10:1 (the same as 10^1:1)	20
100:1 (the same as 10^2:1)	40
1000:1 (the same as 10^3:1)	60
10000:1 (the same as 10^4:1)	80
100000:1 (the same as 10^5:1)	100
1000000:1 (the same as 10^6:1)	120
10000000:1 (the same as 10^7:1)	140
2:1	6
4:1 (the same as 2 times 2 to 1)	12 (the same as 6 + 6)
8:1 (the same as 4 times 2 to 1)	18 (the same as 12 + 6)
20:1 (the same as 2 times 10 to 1)	26 (the same as 6 + 20)
400:1 (the same as 4 times 100 to 1)	52 (the same as 12 + 40)
0.1:1 (the same as 10^{-1}:1)	−20 (minus 20 dB)
0.01:1 (the same as 10^{-2}:1)	−40
0.02:1 (the same as 2 times 0.01 to 1)	−34 (the same as + 6 − 40)

The decibel, then, refers to a certain intensity *ratio.* Specifically, the decibel equivalent of a particular *intensity* ratio is 10 times the logarithm to the base 10 of that ratio. It follows from this definition that 10 decibels corresponds to a 10-to-1 intensity ratio. However, 20 decibels *does not* correspond to a 20-fold intensity change. Rather, 20 decibels (10 *plus* 10 decibels) corresponds to a 100-fold (10 *times* 10) intensity change. Table 3.1 gives the decibel equivalents of a number of different intensity ratios. The figures in this table immediately show one of the reasons that the decibel scale is so practical. The strongest sounds we can hear without feeling pain are as much as 10 million million times greater in intensity than a just audible sound. This huge intensity ratio corresponds to 130 decibels—a much more convenient figure—on the decibel scale.

Although we can express a sound intensity in decibels relative to any intensity we like, in practice an intensity of 10^{-16} watts per square centimeter (near the level of a just audible sound) is used most frequently as a reference level. When we say the average intensity of speech (one meter from the lips) is about 60 decibels relative to 10^{-16} watts per square centimeter, we really mean that this average speech intensity is one million times greater than 10^{-16} watts per square centimeter (see Table 3.1).

It is easier to measure the pressure of a sound wave rather than its intensity. Consequently, we usually measure the sound pressure and infer the intensity from the pressure value. Sound intensity is proportional to the *square* of the pressure variations of the sound wave. Therefore, a 100-fold increase in intensity produces a 10-fold increase in sound pressure. A 10,000-fold intensity increase corresponds to a 100-fold increase in pressure, and so on. We want the same dB value to apply both to a given intensity ratio and to the corresponding pressure ratio. Consequently, 20 dB must be equivalent to a 10-to-1 pressure ratio (or 100-to-1 intensity ratio). For this reason, the dB equivalent of a particular pressure ratio is 20 times the logarithm to the base 10 of that ratio. The square-law relationship between pressure and intensity explains why the same change in decibels refers to different values of pressure and intensity ratios. Table 3.2 gives the decibel equivalents of a selection of sound pressure ratios.

ACOUSTICAL RESONANCE

We have now discussed the nature of vibration, resonance and sound waves. We will conclude this chapter by explaining acoustical resonance which, as we will see in Chapter 4, plays an extremely important part in speech production.

Enclosed volumes of air can resonate just like the spring-mass combination we described earlier. When a sound wave reaches a volume of air enclosed in a container, an increase in the sound pressure compresses the air in the container. The "springiness" of the air inside the container tends to push the compressed air out again. If the rarefaction of the sound wave reaches the container at the same time the compressed air is being pushed out, the pressure of the sound wave and the pressure of the compressed air will add together and the air particles will move with increased amplitude. If the rate of arrival of the sound wave's compressions and rarefactions (the rate being equal to the sound wave's frequency of vibration) corresponds to a natural frequency of the enclosed air, we get increased movement or *resonance*. When we fill a bottle with water, we can actually *hear* it filling up. Resonance explains this: the splashing water generates sounds of many different frequencies, but the resonance of the air column above the water level emphasizes only those frequencies in the sound that are near its own natural frequency. As the bottle fills up, the size of the air column decreases (this *increases* the column's resonant frequency), and higher frequency components of the "splashing" are emphasized. We know from experience that, when the pitch of the sound from the bottle is high enough, little air is left in the bottle and it is time to turn off the tap.

The simple spring-mass combination has only one resonant frequency; columns of air have many different resonant frequencies. We will consider only the resonances of tubes whose cross-sectional dimensions are small compared to the wavelengths of the sounds applied to them. The vocal tract is just this sort of tube for the frequencies of primary interest in speech.

A tube with a uniform cross-sectional area throughout its length has regularly spaced resonant frequencies. The values of these resonant frequencies depend on the length of the tube. Consider a tube closed at one end and open at the other. The lowest resonant frequency of this tube corresponds to the frequency of a sound wave whose length is four times the length of the tube. The value of the tube's other resonant frequencies will be odd-number multiples (three times, five times, etc.) of this lowest resonant frequency. When the cross-sectional area varies along the tube's length, the resonant frequencies are no longer uniformly spaced. They are spaced irregularly, some close together and some far apart, depending on the exact shape of the tube.

The human vocal tract is about 17 centimeters long and—at least when it produces vowel sounds—we can regard it as closed at one end and open at the other (the lips). The lowest resonant frequency of a uniform tube this long is 500 cps; its other resonant frequencies are 1500 cps, 2500 cps, 3500 cps, and so on.

When both ends of the tube are closed, the lowest resonant frequency is actually zero. The next resonant frequency, for a uniform tube, has a value corresponding to the frequency of a sound wave whose wavelength is twice as long as the tube. The values of the other resonant frequencies are even-numbered multiples of this next-to-lowest frequency.

As we will see in Chapter 4, the vocal tract is a tube of complicated shape that acts as a resonator. Its shape is varied by movements of the vocal organs. The resulting changes in its resonant frequencies play an important part in speech production.

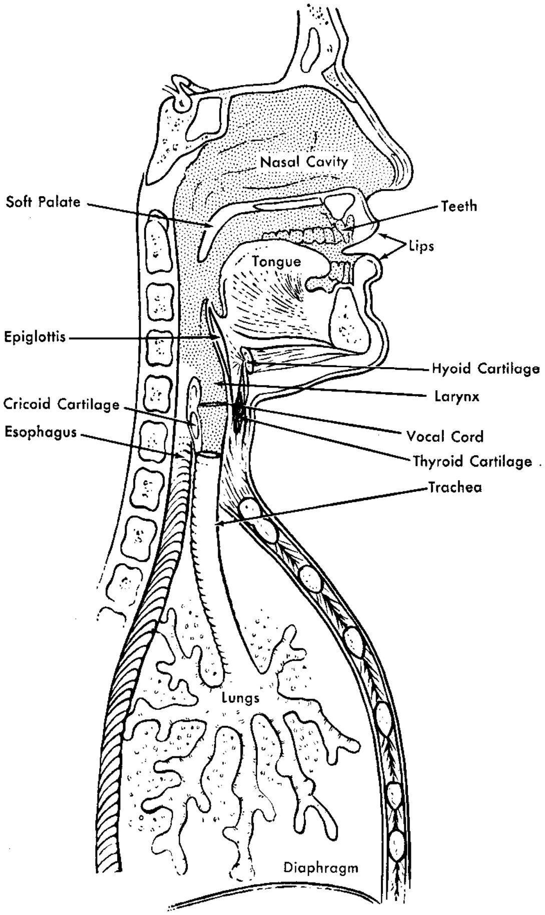


Fig. 4.1 The human vocal organs.

CHAPTER 4 Human Speech Production

The specialized movements of our vocal organs are the means whereby we generate the sound waves of speech. Expressions like "his lips are sealed," "mother tongue" and "tongue-tied," are ample evidence that man has always understood the vital contribution of these organs to speech production.

The lips and tongue, however, are not the only organs associated with speech production. In this chapter, we shall describe all the organs involved; we shall explain how they function during speech and how the sound waves are produced.

The chapter has four sections. First, we have a quick look at the speech mechanism as a whole. Next, we describe the vocal organs, one by one. In the third section, we explain how these organs move to produce each English speech sound. The last section is concerned with the acoustics (or physics) of how the vocal organs produce and shape the sound waves of speech.

A BRIEF DESCRIPTION OF SPEECH PRODUCTION

A diagram of the *vocal organs*, those parts of the body connected with speech production, is given in Fig. 4.1. The vocal organs are the *lungs*, the *trachea* or windpipe, the *larynx*, (containing the *vocal cords*), the throat or *pharynx*, the *nose* and the *mouth*. Together, these organs form an intricately shaped "tube" extending from the lungs to the lips. One part of the tube, lying above the larynx, is called the *vocal tract*, and consists of the pharynx, mouth and nose. The shape of the vocal tract can be varied extensively by moving the tongue, the lips and other parts of the tract.

The source of energy for speech production is the steady stream of air that comes from the lungs as we exhale. When we breathe normally, the air stream is inaudible. It can be made audible by setting it into rapid vibration. This can happen unintentionally; when we snore, for example. During speech, of course, we intentionally set the air stream into vibration. We can do this several ways, but the method most frequently used is by vocal cord action.

The vocal cords are part of the larynx. They constitute an adjustable barrier across the air passage coming from the lungs. When the vocal cords are open, the air stream passes into the vocal tract; when closed, they shut off the air flow from the lungs. As we talk, the vocal cords open and close rapidly, chopping up the steady air stream into a series of puffs. We can hear this rapid sequence of puffs as a buzz whose frequency gets higher and higher as we increase the vibration rate of the vocal cords. The character of the vocal cord buzz is modified by the vocal tract's acoustic properties. These acoustic properties depend on the shape of the vocal tract. During speech, we continually alter this shape by moving the tongue and the lips, etc. These movements, by altering the acoustic properties of the vocal tract, enable us to produce the different sounds of speech.

Adjusting the vocal tract's shape to produce different speech sounds is called *articulation*; the individual movements of the tongue, lips and other parts of the vocal tract are called articulatory movements.

We see, then, that air flow from the lungs provides the energy for speech wave production, that the vocal cords convert this energy into an audible buzz and that the tongue, lips, palate, etc.—by altering the shape of the vocal tract—transform the buzz into distinguishable speech sounds.

The mechanism just described is used for producing most speech waves. Two other methods are available for making the air stream from the lungs audible. In one, the vocal tract is constricted at some point along its length. The air stream passing through the constriction becomes turbulent, just like steam escaping through the narrow nozzle of a boiling tea kettle. This turbulent air stream sounds like a hiss and is, in fact, the hissy or *fricative* noise we make when pronouncing sounds like "s" or "sh."

The other method is to stop the flow of air altogether — but only momentarily — by blocking the vocal tract with the tongue or the lips, and then suddenly releasing the air pressure built up behind this block. We use the "blocking" technique to make sounds like "p" and "g," which are called *plosives.* It should be remembered that the second and third methods described are independent of vocal cord activity, although the speaker can vibrate his vocal cords simultaneously. Whichever of the three techniques is used, the resonances of the vocal tract still modify the character of the basic sounds produced by hiss, plosion or vocal cord vibration.

We may mention a few other methods that can be used for speech production, even though their use is infrequent. Producing a whisper is like making a hiss sound, except that the constriction that agitates the air stream is provided by holding the vocal cords still and close together. In some African languages, "clicks" are used. Clicks are produced by blocking the vocal tract at two points, sucking the air out from between the two blocks and then re-opening the tract. Some foreign languages use sounds produced while inhaling, but English speech is normally produced only while exhaling.

THE VOCAL ORGANS

We can now consider the action of the vocal organs in more detail. You may be interested to know, incidentally, that the primary biological function of the vocal organs is not speech production. They developed first to perform some other vital service, like breathing, chewing and swallowing, and were only later applied to the production of speech.

The lungs are masses of spongy, elastic material in the rib cage. They supply oxygen to the blood and dispose of certain waste products like carbon dioxide. The act of breathing air in and out is controlled by various muscles of the rib cage, and by muscles of the abdomen and the diaphragm, the partition that separates the chest from the abdomen. During speech, the diaphragm relaxes and the degree of abdominal muscle contraction controls the extent to which the contents of the abdomen are pressed up against the diaphragm and carried into the chest cavity, where they squeeze air out of the lungs.

Normally, the lungs contain about three quarts of air. In addition, we regularly breathe in and out about one-half quart of air. If we first inhale deeply and then breathe out as far as we can, we may exhale as much as three and a half quarts of air, leaving about one and a half quarts of residual air in the lungs.

When we exhale, the air pressure from the lungs is only slightly above atmospheric pressure. It is about one-quarter of one per cent greater than atmospheric pressure when we breathe normally, and approximately one per cent greater than atmospheric during conversation.

Normally, we breathe about once every five seconds; roughly equal parts of this period are devoted to exhaling and inhaling. During speech, we can influence our breathing rate in accordance with the needs of sentence and phrase length; since we talk only while exhaling, we can adjust this rate to devote as little as 15 per cent of the breathing cycle to inhaling.

Air from the lungs travels up the trachea (see Fig. 4.1), a tube consisting of rings of cartilage, and through the larynx toward the mouth and nose.

The larynx acts as a gate or valve between the lungs and the mouth. By opening or closing, it controls the flow of air from the lungs; when it is shut tightly, it completely isolates the lungs from the mouth. Because the larynx can close the air passages, it plays an important part in speech production, and in eating and breathing.

We take in both food and air through the mouth. When these essential commodities reach the back of the mouth — the pharynx — they face two downward openings: the larynx, leading through the trachea to the lungs, and the food pipe or *esophagus*, leading to the stomach (see Fig. 4.1). Food and air should not enter the wrong passage if our body is to function properly. We all know how unpleasant it is when food or any foreign matter finds its way into our windpipe—"goes down the wrong place," in other words. The larynx prevents this by closing automatically during swallowing to exclude food from the trachea and lungs.

Another function of the laryngeal valve is to lock air into the lungs. Animals which use their forelimbs extensively — especially tree climbing mammals — have a well developed larynx because the arms can exert greater force when they are given rigid support by the air locked in the chest cavity. You can try this yourself. See how much the power of your arms is weakened

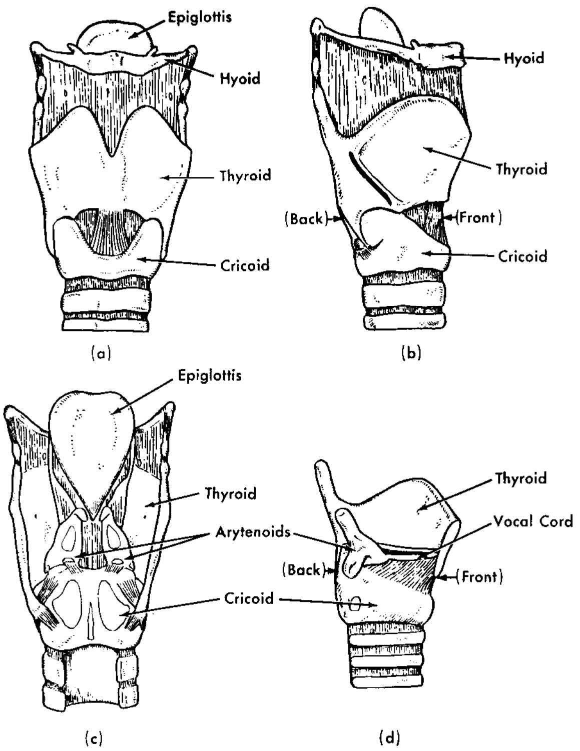


Fig. 4.2 Various views of the larynx: (a) front; (b) side; (c) back; (d) cut away side

if you fail to hold your breath. Normally, we unconsciously hold our breath when we do heavy work with our arms.

Man has learned to use his laryngeal valve to convert the steady air stream from the lungs into audible sound. He uses his larynx to break the air flow into a series of puffs, which is heard as a buzz; it is the sound wave we use in speech.

Constructionally, the larynx is a stack of cartilages. You can locate your larynx easily because one of its cartilages, the thyroid, is the projection on your neck known as the Adam's apple. Fig. 4.2 shows various views of the larynx's principal cartilages. The cartilages and their connecting muscles and ligaments form a series of rings about three inches high and less than two inches across. The larynx is not held in one rigid position; it can move up and down during swallowing and speaking.

At the top of the larynx is the pear-shaped *epiglottis.* Its narrow end is attached to the Adam's apple and its other end is free. During swallowing, the epiglottis helps to deflect food away from the windpipe, performing part of the larynx's valve function.

The valve action of the larynx depends largely on the *vocal cords.* The vocal cords are folds of ligament that extend, one on either side of the larynx, from the Adam's apple at the front to the *arytenoid* cartilages at the back. The space between the vocal cords is called the *glottis.* When the arytenoids — and, therefore, the vocal cords—are pressed together, the air passage is sealed off and the laryngeal valve is shut. The glottal opening can be controlled by moving the arytenoids apart, as shown in Fig. 4.3. The open glottis is "V-shaped" because the vocal cords, held together at the front, move apart only at the back.

The length of the vocal cords can be altered by moving and rotating the arytenoids and, sometimes, the Adam's apple. The glottis is about three quarters of an inch long and can be opened about half an inch by the arytenoids.

Just above the vocal cords is another pair of folds, the *false vocal cords.* They also extend from the Adam's apple to the arytenoids. Opinion differs on just how much effect the false cords have on speech production. They can be closed and they can vibrate but, during speech, they are probably open. Fig. 4.4 illustrates the relationship between false and true vocal cords.

We see, then, that the larynx provides a triple barrier across the windpipe through the action of the epiglottis, the false vocal cords and the true vocal folds. All three are closed during swallowing and wide open during normal breathing.

What does the larynx do during speech? When we talk, the epiglottis and false vocal cords remain open, but the vocal

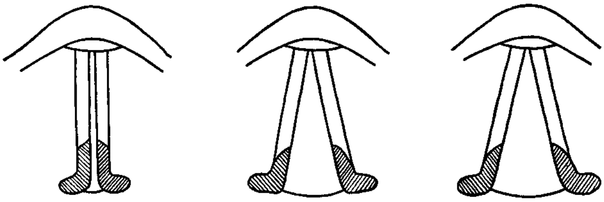


Fig. 4.3 The control of the glottal opening. The shaded areas represent the arytenoids. The curved, top portion of the figure is the Adam's apple.

cords close. Air pressure builds up behind the vocal cord barrier and eventually blows the cords apart. Once apart, the excess pressure is released, the cords return to their closed position, the pressure builds up again and the cycle repeats. The vibrating vocal cords rhythmically open and close the air passage between the lungs and mouth. They interrupt the steady air flow and produce the sequence of air puffs mentioned earlier.

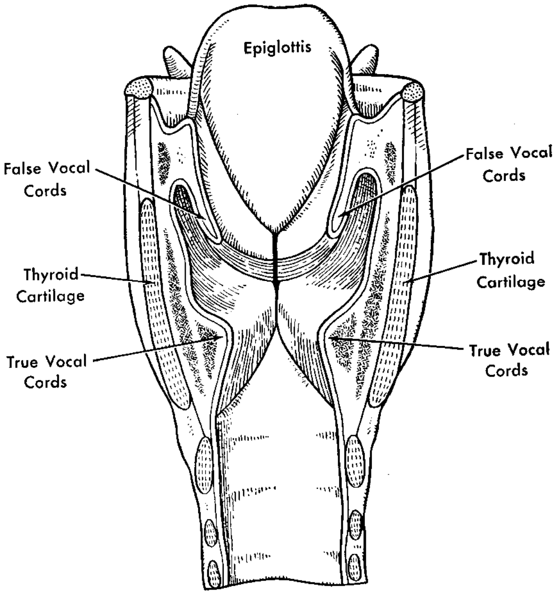


Fig. 4.4 The relationship between false and true vocal cords.

The frequency of vocal cord vibration and, consequently, the frequency of the air puffs, is determined by how fast the cords are blown apart and how fast they snap back into their closed position.

This frequency is controlled by a combination of effects. There are the massiveness of the vocal cords and their tension and length. There is also the effect of low air pressure created in the glottis by air rushing through its narrow opening into the wider space above. This draws the vocal cords back to their starting position and, consequently, increases their speed of return. Greater air pressure from the lungs enhances this effect and increases the frequency of vocal cord vibration.

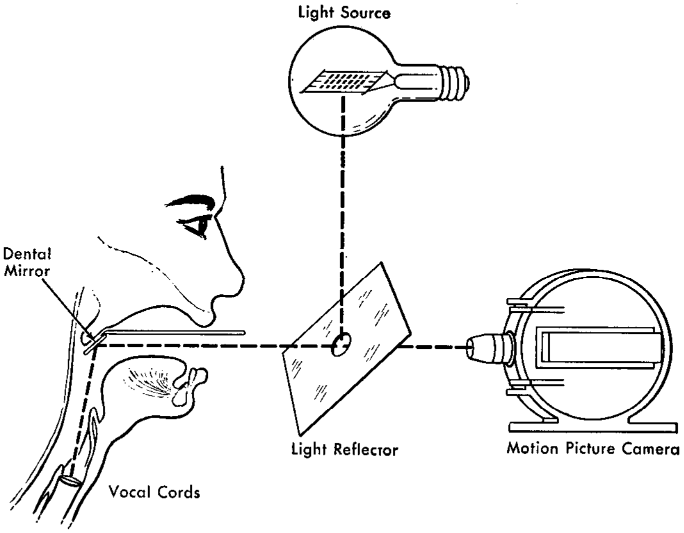


Fig. 4.5 Method of observing vocal cord movement.

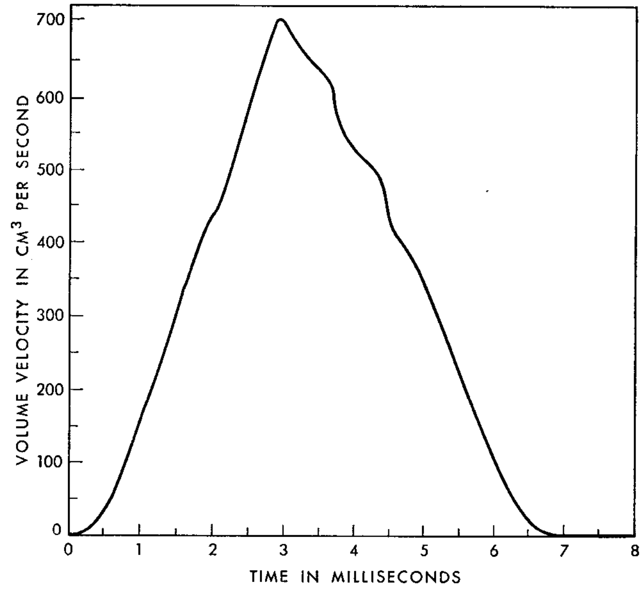


Fig. 4.6 The variation of air flow in a glottal puff. The curve repeats once every 8 ms (a frequency of 125 cps).

During speech, we continually alter the tension and length of the vocal cords — and the air pressure from the lungs — until we get the desired frequency. The range of vocal cord frequencies used in normal speech extends from about 60 to 350 cps, or more than two octaves. Higher frequencies are occasionally used. In any one person's speech, the normal range of vocal cord frequencies covers about one and a half octaves.

We can observe the vocal cords by placing a dental mirror in a speaker's mouth, as shown in Fig. 4.5. The vocal cords vibrate so rapidly, however, that their movements are not clear when observed this way. We can see much more by filming what appears in the dental mirror with a special high-speed camera, and viewing the film in slow motion. Observations of this kind show that the vibrating vocal cords move up and down as well as sideways, although the sideways movement predominates. The slow motion films also show that the vocal cords do not always close completely during their vibration cycle.

Suitable measurements have enabled us to determine how the air puffs vary throughout the glottal cycle. Fig. 4.6 shows a a typical curve. The spectrum of such pressure waves has many components, but their frequencies are always whole-number multiples of the vocal cord frequency. Their amplitudes generally decrease as their frequencies increase. In loud speech and shouting, the vocal cords open and close more rapidly and remain open for a smaller fraction of a cycle; this increases the amplitudes of the higher harmonics and gives the sounds a harsher quality.

We have seen how the energy for speech is provided by the air stream from the lungs and how vocal cord vibration generates an audible buzz. Let us go on to see how the quality of this buzz is changed by the configuration of the vocal tract. A cross-sectional view of the vocal tract is shown in Fig. 4.7 on page 54. The tract extends from the glottis to the lips—by way of the pharynx and mouth—with a side branch into the nasal passages.

The *pharynx* is the part of the vocal tract nearest the glottis. It is a tube connecting the larynx with the mouth and the nose. At its lower end, the pharynx meets the larynx and the esophagus and, at its wider upper end, joins with the back of the mouth and the nose, as shown in Fig. 4.8. We have known for some time that its shape and size are changed when swallowing, either by moving the tongue back, or the larynx up, or by contracting the pharyngeal walls. Only recently, however, has it been noticed that such changes take place during speech. These changes can be seen clearly in the vocal tract outlines shown in Fig. 4.9 on page 56. Very little is known about the pharyngeal changes we make for distinguishing one speech sound from another.

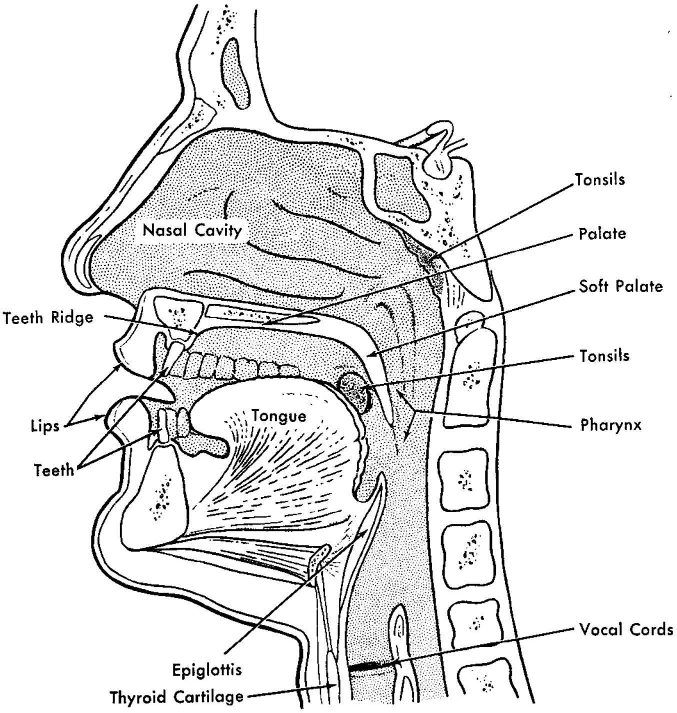


Fig. 4.7 Cross-sectional view of the vocal tract.

As a result, we shall not have much to say about the pharynx later in this chapter, when we describe the movements of the vocal organs for articulating English speech sounds.

The nasal cavity (see Fig. 4.7), extending from the pharynx to the nostrils, is about four inches long. It is divided into two sections by the *septum*, a central partition that runs along the entire length of the cavity. Ridges and folds in the cavity's

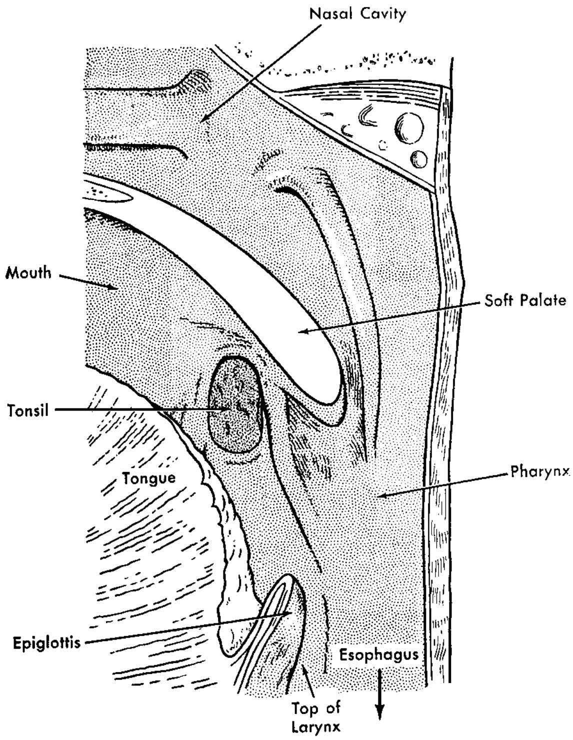


Fig. 4.8 The interior of the pharynx.

walls break up some segments of the nasal air passages into intricately shaped channels. At the back of the nose—and also lower down in the pharynx—are the *tonsils* (see Fig. 4.7). They occasionally grow large enough to influence the air flow from the lungs and, when they do, they add the characteristic "adenoidal" quality to the voice. The sensory endings of the nerve concerned with smell are also located in the nose. The nasal cavities can be isolated from the pharynx and the back of the mouth by raising the *soft palate* (to be described in a later section of this chapter).

The last and most important part of the vocal tract is the mouth. Its shape and size can be varied—more extensively than any other part of the vocal tract—by adjusting the relative positions of the palate, the tongue, the lips and the teeth.

The most flexible of these is the tongue. Its tip, its edges and its center can be moved independently; and the entire tongue can move backwards, forwards and up and down. Fig. 4.10 shows the complicated system of muscles that makes such movements possible. The tongue's covering of mucous membrane contains the nerve endings concerned with the sense of taste.

The lips, which affect both the length and the shape of the vocal tract, can be rounded or spread to various degrees, as shown in Fig. 4.11. They can also be closed to stop the air flow altogether.

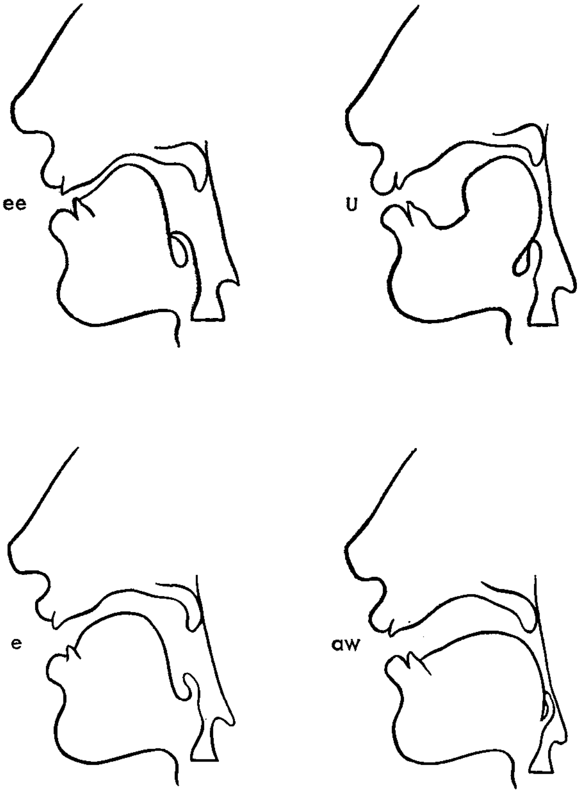


Fig. 4.9 Outlines of the vocal tract during the articulation of various vowels.

The lips and the cheeks influence speech communication in more than one way. They change the shape of the vocal tract and, consequently, the kind of speech sound produced. But, together with the teeth, they are the only parts of the vocal tract normally visible. The listener can gather information about what

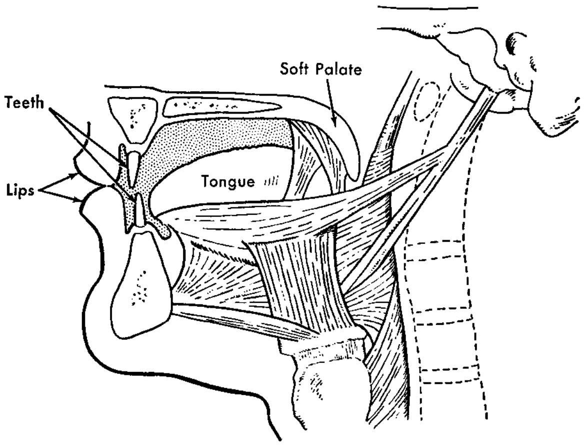


Fig. 4.10 The muscles of the tongue.

the speaker is saying by watching his face as well as listening to his voice. This is called *lip reading*, and it has a more significant effect on speech communication than most people give it credit for. If you have ever had a conversation in really noisy surroundings, you know how useful it is to see the speaker's face. Most deaf people can understand some of what you say just by watching your face.

There is still another way the lips and cheeks play a part in speech communication. Their shape contributes to setting the facial expressions that give an indication of your emotions; this can help a listener understand speech that might otherwise be insufficiently intelligible.

The teeth also affect the vocal tract's shape and the sounds it produces. They can be used to restrict or stop the air flow by placing them close to the lips or the tip of the tongue, as in the sounds "v" or "th," for example.

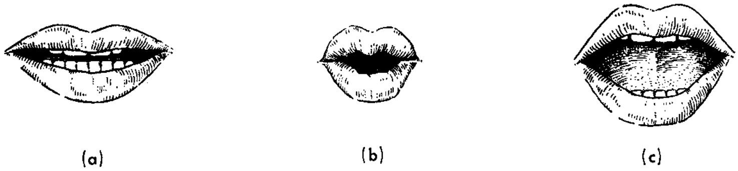


Fig. 4.11 The shapes of the lips during articulation: (a) spread; (b) rounded; (c) unrounded.

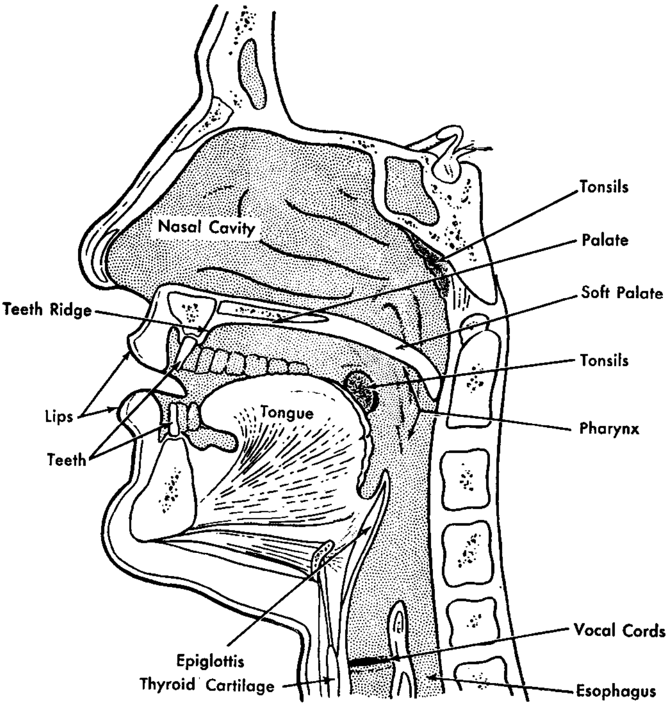


Fig. 4.12 Vocal tract configuration for articulating non-nasal sounds.

The last of the organs that shape the mouth cavity is the *palate*. We can divide it into three parts. They are the teeth ridge or *alveolus*, covered by the gums; the bony *hard palate* that forms the roof of the mouth; and the muscular *soft palate* at the back. If you stand in front of a mirror and open your mouth wide, you will see your soft palate moving up and down at the back of your mouth. The soft palate is normally lowered, taking up roughly the position shown in Fig. 4.7. It can be raised, however, and in this position it closes the opening between the pharynx and the nose (see Fig. 4.12), and the air expelled from the lungs is directed entirely along the mouth.

This completes our description of all the organs important in shaping the vocal tract. By setting the shape of the vocal tract — and its acoustic characteristics — the vocal organs enable us to distinguish one speech sound from another. Let us see, now, how these organs move in articulating the sounds of spoken English.

THE ARTICULATION OF ENGLISH SPEECH SOUNDS

For convenience, we will divide the speech sounds into two groups, vowels and consonants.

The vocal cords vibrate during the articulation of vowels; they also vibrate when making some of the consonants. Sounds produced with vocal cord vibration are called *voiced.* Sounds produced without vocal cord vibration are called *unvoiced.*

We will describe the articulation of vowels in terms of tongue and lip positions. Some speakers raise their soft palates during vowel production, shutting off the nasal cavities, while others leave it partially lowered. The added nasal quality is not used to distinguish one English vowel from another.

It is not so easy to describe the positions of the tongue. The tongue is highly mobile and its tip, edges and main body can move independently. Experience has shown that its configuration can best be described by specifying where the main body of the tongue is. The position of the highest part of the main body is called the *position of the tongue.*

The tongue positions used for making vowels are usually described by comparing them with the positions used for making a number of reference or *cardinal vowels.* There are eight cardinal vowels and they form a set of standard reference sounds whose quality is defined independently of any language. They serve as a yardstick of vowel quality against which the quality of any other vowel can be measured. The cardinal vowel system is basically a system of perceptual qualities, but X-ray experiments show substantial agreement between vowel quality and tongue position. It has become acceptable, therefore, to compare the tongue positions of vowels with those of the cardinal vowels. Strictly speaking, no written definition of cardinal vowel quality is possible because the "definition" of quality is perceived only when listening to a trained phonetician making the sounds. However, we may hazard an approximate definition. When a person moves his tongue as high up and as far forward as he can—without narrowing the width of the air passage to produce a hiss—and spreads his lips at the same time, an "ee"-like sound is produced; this is called cardinal vowel 1. If he now keeps the tongue high, moves it back as far as he can and rounds his lips, he will make

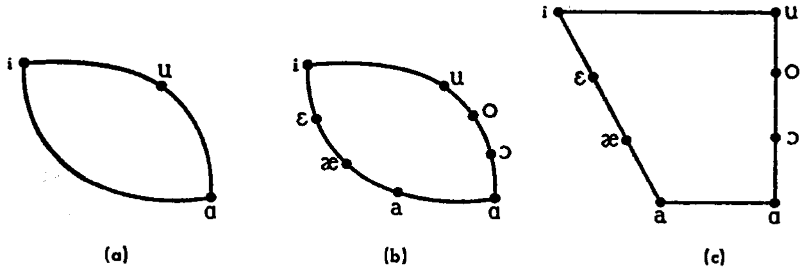


Fig. 4.13 Tongue positions for cardinal vowel articulation: (a) cardinal vowels 1, 5 and 8; (b) the eight cardinal vowels; (c) schematic representation of tongue positions for the same eight cardinal vowels as in (b).

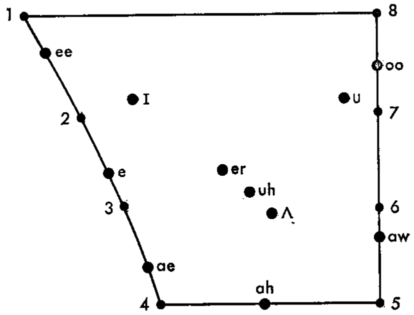
an "oo" sound; this is called cardinal vowel 8. If the tongue is moved down as far as possible, still keeping it far back, and the lips are unrounded, he produces cardinal vowel 5, a sound very much like the "aw" in the word "c*a*ll." By mapping the tongue positions for these three, we get the diagram shown in Fig. 4.13(a). The other five cardinal vowels are defined as those sounds that divide the distances between the three mapped positions into perceptually equal sections. Fig. 4.13(b) shows a map of the corresponding tongue positions and Fig. 4.13(c) the conventional form in which the tongue positions of Fig. 4.13(b) are usually shown. Fig. 4.13(c) is the so-called *vowel quadrilateral.* Cardinal vowels 2, 3, 4, 6 and 7 are shown in Figs. 4.13(b) and 4.13 (c) as ɛ, æ, a, ɔ and o, respectively. All the tongue positions of the cardinal vowels are along the outer limits of tongue movement. If the position of the tongue moves toward the center of the mouth, the quality of the sound becomes more neutral and "uh"-like.

When the tongue is near the palate, the sound produced is called a *close sound*; when the tongue is low, at the bottom of the mouth, the sound is called *open.* Sounds produced with the tongue near the center of the vowel quadrilateral are called *central* or *neutral* vowels. The sound "ee," therefore, is a *close front* vowel, and "oo" a *close back* vowel; "ah" and "aw" are *open front* and *open back* vowels, respectively.

Basically, any lip configuration could be used with any tongue position. In English, however, front vowels are usually made with spread lips and back vowels with rounded lips; as the tongue is lowered to more open positions, the lips tend to become unrounded. Native speakers of English find it difficult to go against this "rule," and some of us might even say that the muscles of our mouths are so constructed that these lip shapes and tongue positions must necessarily go together. This is really only a matter of habit, however. Other languages do have sounds with a different relationship between lips and tongue.

In Russian, for example, there is a vowel made with the tongue in a close back position (as for the English "oo"), but with the lips parted (as for the English "ee"). Again, in French, there is a vowel made with the tongue in a close front position (as for the English "ee"), but with the lips rounded (as for the English "oo").

Why not try to make this French vowel? It is that elusive sound used in "rue," the French word for "street." At first, you might find it impossible, but if you persist, perhaps even using your fingers to round your lips, while keeping your tongue in the position for an English "ee," you should succeed.



THE PURE VOWELS			
ee	heat	ah	father
I	hit	aw	call
e	head	U	put
ae	had	oo	cool
uh	the	Λ	ton
er	bird		

THE DIPHTHONGS	
ou	tone
ei	take
ai	might
au	shout
oi	toil

Fig. 4.14 Tongue positions for English vowels (the tongue positions for the eight cardinal vowels are shown by numerals).

Let us return to English sounds. Fig. 4.14 shows the tongue positions for the principal English vowels. The vowels shown in this figure are the so-called *pure* vowels, which means that their quality remains substantially unchanged throughout the syllables in which they are used. There is another group of English vowels, the *diphthongs* (pronounced, "dif-thongs"); the diphthong is a sound whose quality changes noticeably from its beginning to its end in a syllable. The principal diphthongs of English are also listed in Fig. 4.14. The tongue movements associated with these sounds are roughly movements between positions assumed for pure vowels. For the diphthong "au," for example, the tongue would move roughly from the position for the sound "ah" to the position for the sound "ʊ," and so forth.

The English consonants are best described by specifying their *place-of-articulation* and their *manner-of-articulation*; they are further distinguished by whether they are voiced or unvoiced.

The significant places-of-articulation in English are the lips, the teeth, the gums, the palate and the glottis. The categories of manner-of-articulation are *plosive*, *fricative*, *nasal*, *liquid* and *semi-vowel*.

The plosive consonants, for example, are made by blocking the air pressure somewhere along the mouth and then suddenly releasing the pressure. The air flow can be blocked by pressing the lips together or by pressing the tongue against either the gums or the soft palate. We can have plosives, then, with *labial* (lips), *alveolar* (gums) or *velar* (soft palate) places-of-articulation.

Similarly, fricatives are made by constricting the air flow somewhere in the mouth — enough to make the air turbulent —to produce sound of a hissy quality. The fricatives, like the plosives, differ according to their places-of-articulation. The nasals are made by lowering the soft palate, coupling the nasal cavities to the pharynx, and blocking the mouth somewhere along its length to produce different places-of-articulation. All other English consonants are made with the soft palate raised.

The English semi-vowels are the "w" and "y" sounds. Both are produced by keeping the vocal tract briefly in a vowel-like position, and then changing it rapidly to the position required for the following vowel in the syllable. Consequently, the semi-vowels must always be followed by a vowel in whatever syllable they are used. The consonant "y" is formed by putting the tongue in the close frontal position required for an "ee" sound, holding it there briefly and then changing to whatever vowel follows the "y." Forming the "w" is similar, except that the lips are first close rounded, as required for an "oo." The consonants "w" and "y," therefore, have a labial and an alveolar place-of-articulation, respectively. They are both voiced consonants.

The only English lateral consonant is "l." It is made by putting the tip of the tongue against the gums and allowing the air to pass on either side of the tongue. It is a voiced consonant.

A classification of all English consonants, according to place- and manner-of-articulation, is given in Table 4.1.

Table 4.1—Classification of English Consonants by Place- and Manner-of-Articulation

Place of articulation	Manner of articulation				
	Plosive	Fricative	Semi-vowel	Liquids (incl. laterals)	Nasal
Labial	p b		w		m
Labio-Dental		f v			
Dental		θ th			
Alveolar	t d	s z	y	l r	n
Palatal		sh zh			
Velar	k g				ng
Glottal		h			

The vocal tract configurations we have described are not made exactly this way every time a speech sound is produced. We have described typical (idealized) articulations and considerable deviations from these occur in actual speech. Deviations can be due to the individual habits of different speakers. They can also occur because of the influence of other sounds that immediately precede or follow the sound being uttered. For example, the sound "k" is made by pressing the back of the tongue against the soft palate. Just where the tongue and palate meet depends a lot on what the following vowel is; if it is a back vowel, like an "oo," the contact will be much further back than with an "ee." Again, in fast speech, we often start the articulation of a particular sound—that is, move the tongue or lips toward the specified position— without finishing the movement before going on to the next sound. Despite all these variations, speech is still intelligible.

Now that we have seen how the movements of the various vocal organs shape the vocal tract tube, we can consider the tube's acoustic effect on the character of sounds produced.

THE ACOUSTICS OF SPEECH PRODUCTION

You will recall that the buzz-like sound produced by the vocal cords is applied to the vocal tract. The vocal tract is, in effect, an air-filled tube and, like all air-filled tubes, acts as a resonator. This means that the vocal tract has certain natural frequencies of vibration, and that it responds more readily to a sound wave whose frequency is the same as its resonant frequency than to a sound wave of another frequency. Let us assume, for example, that the vocal cords produce a series of pulses as shown in Fig. 4.15(a). The spectrum of such a sound has a large number of components; all of them are more or less of the same amplitude and have frequencies that are whole-number multiples of the fundamental frequency. The fundamental—the spectrum's lowest frequency component—has the same frequency as the vocal cords' frequency of vibration. When such a wave is applied at one end of the vocal tract (at the glottis), and is transmitted toward the lips, the vocal tract responds better to those components of the vocal cord puffs that are at or near its natural frequency. These components will be emphasized and the spectrum of the sound emerging from the lips will "peak" at the natural frequency of the vocal tract. This is the process illustrated in Fig. 4.15. Fig. 4.15(b) shows the spectrum of the vocal cord output and Fig. 4.15(c) shows the frequency response of a simple resonator. Figs. 4.15(d) and (e) are the waveshape and spectrum of the sound wave produced when the sound of Fig. 4.15(a) is transmitted through the resonator of Fig. 4.15(c).

The resonator of Fig. 4.15(c) has only one natural frequency, but the vocal tract has many. The vocal resonator, therefore, will emphasize the harmonics of the vocal cord wave at a number of different frequencies, and the spectrum of the speech wave will have a peak for each of the vocal tract's natural frequencies. The values of the natural frequencies of the vocal tract are determined by its shape; consequently, the amplitudes of the spectral components will peak at different frequencies as we change the shape of the tract. Fig. 4.16 shows the spectra of sounds produced for three different vocal tract shapes.

Resonances of the vocal tract are called *formants*, and their frequencies the *formant frequencies.* Every configuration of the vocal tract has its own set of characteristic formant frequencies.

You may have noticed in Fig. 4.15 that the resonant frequency is not equal to the frequency of any harmonic of the spectrum. In general, the frequencies of the formants will not be the same as those of the harmonics, although they may coincide. After all, there is no reason why they should agree. The formant

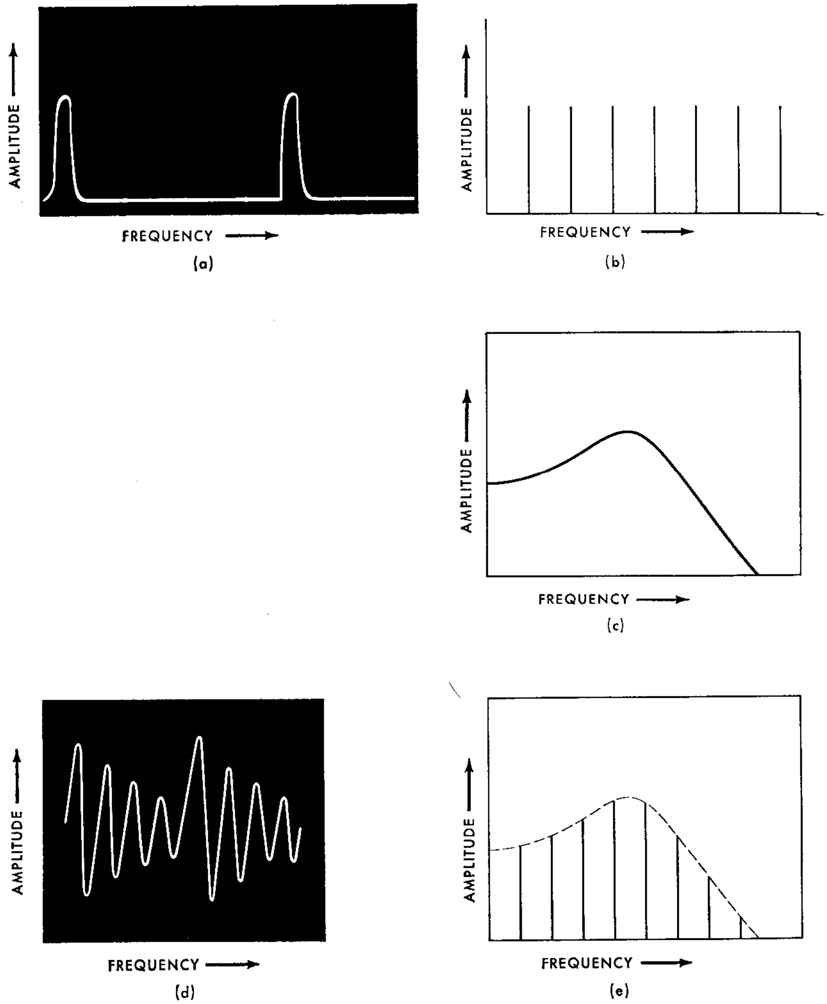


Fig. 4.15 An explanation of formants: (a) the waveshape of a pulse train; (b) a spectrum of a train of short pulses; (c) frequency response of a simple resonator; (d) and (e) are the waveshape and the spectrum, respectively, of a sound wave produced when a series of pulses, like those in (a), is applied to a resonator whose frequency response is shown in (c).

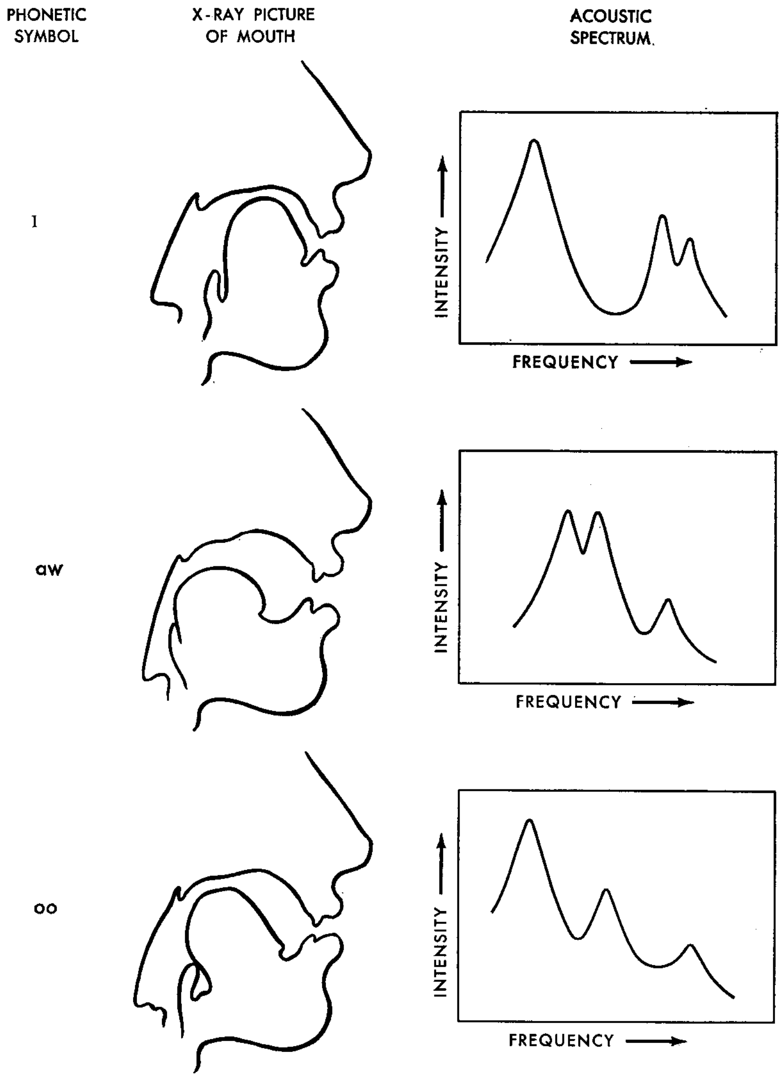


Fig. 4.16 Vocal tract configurations and corresponding spectra for three different vowels. (The peaks of the spectra represent vocal tract resonances. Vertical lines for individual harmonics are not shown.)

frequencies are determined by the vocal tract, the harmonic frequencies by the vocal cords, and the vocal tract and vocal cords can move independently of each other. The independence of vocal cord and formant frequencies is shown in Figs. 4.17

and 4.18. Fig. 4.17(a) shows the waveshape and spectrum of the sound "ah" produced with the vocal cords vibrating at 90 cps, and Fig. 4.17(b) the waveshape and spectrum of the same sound with the cords vibrating at 150 cps. Even though the frequencies of all the harmonics have changed, the frequencies of the formants (and of the spectral peaks) are unaltered because the shape of the vocal tract remained the same. In Fig. 4.18(a), we again see the waveshape and spectrum of the sound "ah" at 90 cps; in Fig. 4.18(b), the vocal cord vibration is still 90 cps, but the shape of the vocal tract has been changed to produce the sound "uh." In Figs. 4.18(a) and 4.18(b), the frequency of vocal cord vibration and, therefore, the frequencies of the harmonics, are the same; the shape of the vocal tract has changed, however, with corresponding changes in the positions of the formants (and of the spectral peaks). The figures clearly show

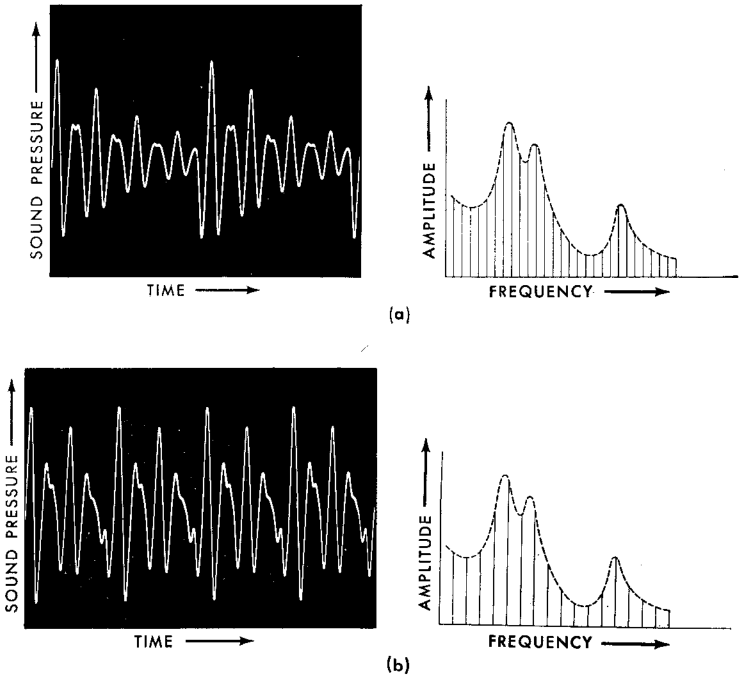


Fig. 4.17 The waveshapes and corresponding spectra of the vowel "ah" pronounced with two different vocal cord frequencies: (a) vocal cord frequency equals 90 cps; (b) vocal cord frequency equals 150 cps

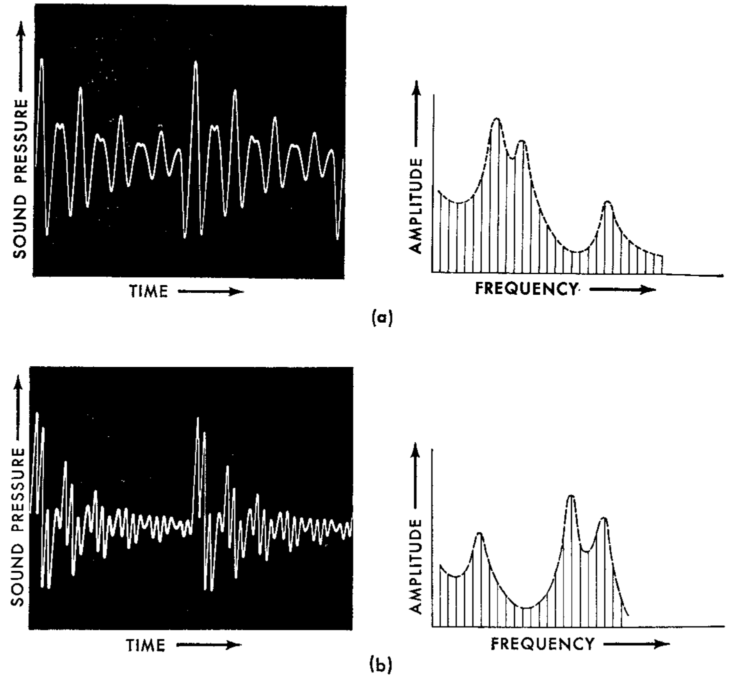


Fig. 4.18 The waveshapes and corresponding spectra of the vowels "ah" and "uh" pronounced with a vocal cord frequency of 90 cps: (a) the sound "ah"; (b) the sound "uh."

that the vocal tract does not affect the frequency of the harmonics, but simply emphasizes the amplitudes of those harmonics that happen to be similar to its own natural, resonant frequency.

Unfortunately, the sound spectra produced by the vocal cords are not always as regular as the one shown in Fig. 4.15(b). The vocal cord spectrum may have its own peaks and valleys; the vocal tract formants will just add more irregularities. The speech spectrum, then, may well have peaks that were not produced by vocal tract resonances. In transmitting speech waves, it would often be useful to know the formant frequencies, and one of the troublesome—and important—problems of present-day speech research is to find ways of determining which of the speech spectrum's numerous peaks are produced by formants and which are due to other causes.

In Chapter 3, resonance was explained in two different ways;

first, as a characteristic of oscillating systems—pendulums, springs and air-filled tubes—when exposed to vibratory forces of different frequencies. We saw that such systems respond more readily to excitation frequencies near their natural frequencies. Second, we saw that when a resonator is disturbed and left alone, it will continue to vibrate at its own natural frequency. Of course, these descriptions are just two different views of the same event. Similarly, there are two different ways to explain the effect of the vocal tract resonances on speech production. So far, we have taken the view that a resonator will respond more readily to excitation at or near its own natural frequency. We could also take the view that each time a puff of air "hits" the vocal tract resonator, the vocal tract continues to "ring" at its own natural frequency. In the simple resonator of Fig. 4.15(c), every air puff from the vocal cords will generate a sinusoidal oscillation at the resonator's natural frequency; the oscillation will decay at a rate determined by its damping. This is shown in Fig. 4.15(d). The spectrum of such a train of damped sinusoids is the spectrum already discussed and shown in Fig. 4.15(e). The vocal tract has many resonant frequencies. It will "ring" at all its natural frequencies simultaneously, and the vibration resulting from the impact of each air puff will be the sum of a number of damped sinusoids. Fig. 4.19(a) shows the waveshape of the sound "ah" and how the same oscillation repeats for every puff of vocal cord sound. Fig. 4.19(b) shows the spectrum of such a wave train. It is identical to the spectrum of "ah" in Fig. 4.17(a), and we can again see that our explanations represent two views of the same event.

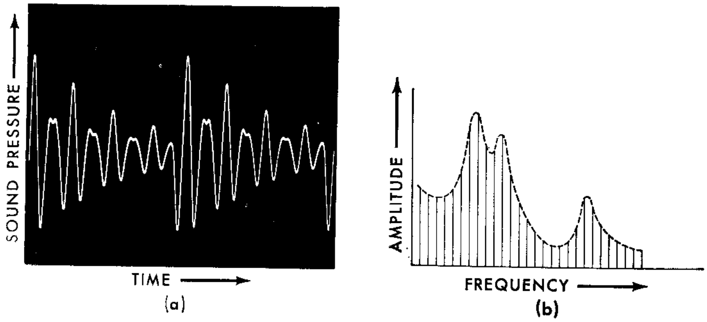


Fig. 4.19 Waveshape and corresponding spectrum of a vowel sound: (a) the waveshape; (b) the spectrum.

Formant frequency values depend on the shape of the vocal tract. When the soft palate is raised, shutting off the nasal cavities, the vocal tract is a tube about seven inches long from the glottis to the lips. For such a tube (with a uniform cross-sectional area along its whole length), the principal resonances are at 500 cps, 1500 cps, 2500 cps, 3500 cps and 4500 cps. In general, the cross-sectional area of the vocal tract varies considerably along its length. As a result, its formant frequencies will not be as regularly spaced as the resonant frequencies of a uniform tube; some of them will be higher in frequency and others lower. The lowest formant frequency is called the *first formant*; the one with the next highest frequency, the *second formant*, and so forth.

When the soft palate is lowered — coupling the nasal cavities to the mouth — a basically different vocal tract shape is formed. The vocal tract starts as a single tube in the pharynx, but separates into two branches at the soft palate, one through the nose and the other through the mouth. We now have different formants because of the additional nasal branch, and we have anti-resonances that suppress parts of the speech spectrum. The nasal cavities also absorb more sound energy; this will increase the damping and reduce the amplitudes of the formants. The speech wave produced depends greatly on whether and where the mouth cavity is obstructed, as in other speech sounds.

Much more research has to be done before we can fully explain how the characteristics of the sounds produced depend on the shape of the vocal tract. Even where the process is fairly well understood, complicated mathematics is needed for the calculation of the formants. As a result, much of what we know about the formants of speech sounds is obtained simply by examining the sounds produced when the vocal tract has a given shape, rather than by explaining how these formants came about. Examination of the acoustic characteristics of speech waves has not only produced more information about formants, but has brought to light other important features of speech waves. Some of these features will be described in Chapter 8.

CHAPTER 5 Electronic Circuits and Circuit Elements

We are now going to talk about some of the basic components you will use in putting your speech synthesizer together. Specifically, we cover capacitors, inductors, resistors and electrical resonators, and the analogies between them and their acoustic and mechanical counterparts. In addition, there is a simple explanation of transistors and their use in the vowel synthesizer you will build.

The differences between electrical and mechanical elements are not so great as you might imagine, even though there is no superficial similarity between a capacitor, for example, and a spring, or between an electrical resonator and a spring-mass combination. In fact, the only significant changes we have to make when we move from a discussion of springs to a discussion of capacitors are changes in the names of the variables involved—from forces and velocities to voltages and currents. Furthermore, you will find that our description of electrical resonance is very much like the explanation of mechanical resonance presented in Chapter 3.

ELECTRICAL VARIABLES

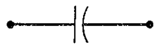
The electrical variables, *voltage*, *current* and *charge*, correspond to the mechanical variables, *force*, *velocity* and *distance*. *Charge* is a measure of sub-atomic, elementary particles called electrons. However, electrons are so small and so numerous that we seldom try to count them. Instead, we measure them by quantity, much as we might measure grains of sand by the cubic yard or sugar by the barrel. Charge is measured in units called *coulombs*. One coulomb is about the amount of charge carried by 6×10^{18} electrons.

Current is the measure of charge flow. The unit of current is the *ampere* and it is measured with an ammeter, much as we might measure the flow of water in a pipe with a flowmeter. One ampere is equivalent to a flow-rate of one coulomb of charge per second.

We can think of *voltage* as a sort of electrical pressure. Charge tends to move from higher voltage to lower voltage, as air tends to move from high pressure to low pressure areas. We could conceive of an absolute voltage corresponding to absolute or barometric pressure but, just as acoustic events depend on pressure differences, so electrical circuit behavior depends on voltage differences. The unit of voltage is the *volt* and we measure voltage differences with a voltmeter, much as we measure pressure differences with a pressure gauge. We can describe the behavior of electrical circuit elements in terms of the variables, charge, current and voltage. We will now discuss the circuit elements used in our synthesizer.

ELECTRICAL CIRCUIT ELEMENTS

Capacitors



We can think of a capacitor as an electrical spring. It consists of two conducting surfaces or "plates" separated by a thin insulator. An electrical connection is made to each plate and brought out as a terminal. In the rest condition, the amounts of charge on the two plates are equal, and the voltage difference between them is zero. Charge can be forced into one of the terminals and out of the other, but no charge flows *through* the insulator separating the two plates. Instead, a surplus of charge builds up on one plate and an equal decrease in charge develops on the other. As this "charging" of the capacitor takes place—that is, as charge is circulated through an external circuit from one terminal to the other—a voltage builds up between its two terminals whose direction opposes further circulation of charge (and tends to restore the rest condition of zero voltage).

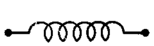
We may regard a capacitor as having an imaginary elastic membrane between the two plates. The membrane allows charge to be forced into one plate, provided an equal amount of charge is removed from the other plate; but as the operation takes place, the membrane stretches and opposes this movement. Here, then, you can see the analogy between a capacitor and a spring. As you stretch or compress a spring, a force develops to oppose the disturbance from the rest position; as you force charge "through" a capacitor, a voltage develops that opposes this disturbance.

The electrical size of a capacitor—its capacitance—is measured in units called *farads* or in *microfarads* (millionths of a farad) or in *picofarads* (millionths of a microfarad). The relationship between charge, Q, voltage, V, and capacitance, C, is expressed by the equation

$$Q = CV.$$

A charge of one coulomb produces a voltage of one volt across a one farad capacitor.

Inductors We can think of an inductor as a device that



has an inertia-like effect on any charge moving through it. An inductor is usually a coil of wire wound around a piece of magnetic metal. It develops a voltage to oppose a change in current, just as a mass develops a force to oppose a change in velocity; this is the inertia effect of the inductor. The electrical size of an inductor—its inductance—is measured in units called *henries* or in *millihenries* (thousandths of a henry) or *microhenries* (millionths of a henry). A one henry inductor develops a voltage of one volt when the current through it changes at a rate of one ampere per second.

Resistors The damping element in electricity is *resistance*

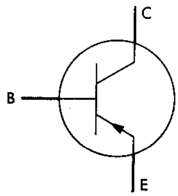
It is common to all electrical conductors, though more pronounced in some than in others. The unit of resistance is the *ohm*.

The relationship between the current, I, through a resistor, R, and the voltage, V, across the resistor is expressed by the equation

$$V = IR.$$

There is a significant difference between the action of a resistor and the actions of inductors and capacitors. Inductors and capacitors produce voltages whose directions either oppose current flow or aid it; this implies that inductors and capacitors either store energy or release it. They are energy *storage* devices. A resistor, however, always absorbs energy. The electrical energy absorbed by a resistor is converted to heat. A resistor, then, is an energy *dissipator*. This is analogous to the effect of friction in mechanical systems.

Transistors



Your synthesizer uses six transistors. The transistor is a piece of silicon or germanium to which small amounts of certain impurities have been added to produce three distinct regions. These are known as the *emitter*, the *base* and the *collector* (see Fig. 5.1 below). Small current and voltage changes applied between the base and the emitter can control large voltages and currents between the collector and the emitter. In this way, small amounts of power can be used to control much larger amounts. The transistor, then, is a power amplifier.

In our synthesizer, transistors are used to perform two different functions. In one, they act as power amplifiers of the type just described. In the second, they are used as switches that turn "on" and "off" rapidly. Both of these functions will be described in more detail in the next chapter.

ELECTRICAL RESONATORS

The following explanation of an electrical resonator paraphrases the explanation of mechanical vibrators given in Chapter 3. You can compare these two explanations to see the similarity between electrical and mechanical resonators. Look at the circuit in Fig. 5.2. A path is provided—between terminals *A* and *C* of

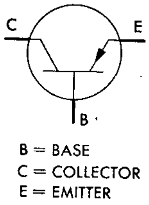


Fig. 5 1 A typical transistor.

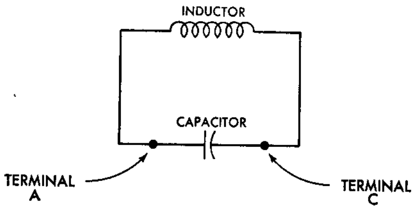


Fig. 5.2 An electrical resonator.

the capacitor—for current to flow through the inductor. In the rest condition, there is no voltage between terminals *A* and *C*, and no current flows through the inductor.

If a charge is circulated through the inductor from terminal *C* to terminal *A*, the capacitor develops a voltage— with terminal *A* having a higher voltage than terminal *C*—that tends to restore the rest condition (zero voltage) between the two terminals. Again, if the charge is displaced in the other direction—from *A* to *C*—a voltage is produced that tends to restore the rest condition.

Suppose we displace a charge from terminal *C* around through the inductor to terminal *A*, and then observe the circuit's behavior. The voltage produced by the capacitor will cause current to flow through the inductor from terminals *A* to *C*. The current will increase until the capacitor's rest condition (zero voltage) is reached and, because of the effective inertia produced by inductance, the current will continue to flow past the rest condition. Once there is a net displacement of charge opposite to the original displacement, the capacitor's restoring voltage will oppose the current flow and, eventually, bring it to a stop. Current will begin to flow in the opposite direction because of the voltage across the capacitor. This type of electrical oscillation is much like the vibration of a mechanical system.

Just as a given spring-mass combination has a particular natural (resonant) frequency determined by the size of its spring and mass, so the resonant frequency of an electrical inductance-capacitance (*L-C*) circuit is determined by the values of its inductor and capacitor. The natural frequency, *f*, of an electrical resonator is given by the equation

$$f = \frac{1}{2\pi\sqrt{LC}},$$

where *L* and *C* are inductance (in henries) and capacitance (in farads), respectively.

Damped Oscillations and Forced Response

A little resistance in a resonator is unavoidable. As we saw earlier, inductors and capacitors can store energy temporarily, but a resistor always dissipates it—converts

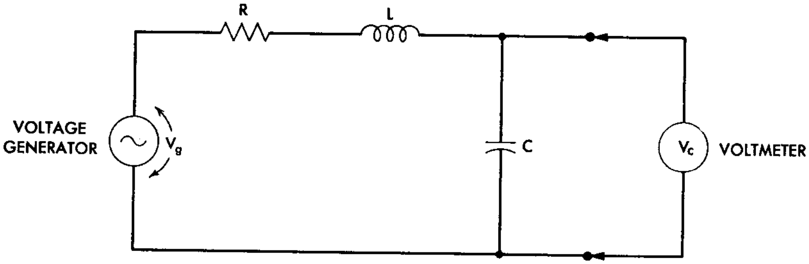


Fig. 5.3 Circuit for obtaining the forced frequency response of an electrical resonator.

it into heat. This, of course, has a *damping* effect on the circuit, and its oscillations decrease in amplitude. Like mechanical vibrators, practical inductance-capacitance circuits always have some damping. The decaying oscillations shown in Fig. 3.2 (Chapter 3) are typical of both electrical and mechanical vibrators.

We can obtain the *forced* frequency response of an electrical resonator by using a sinusoidal voltage generator, just as we obtained a forced response for the spring-mass system in Chapter 3 by moving one end of the spring sinusoidally. An electrical

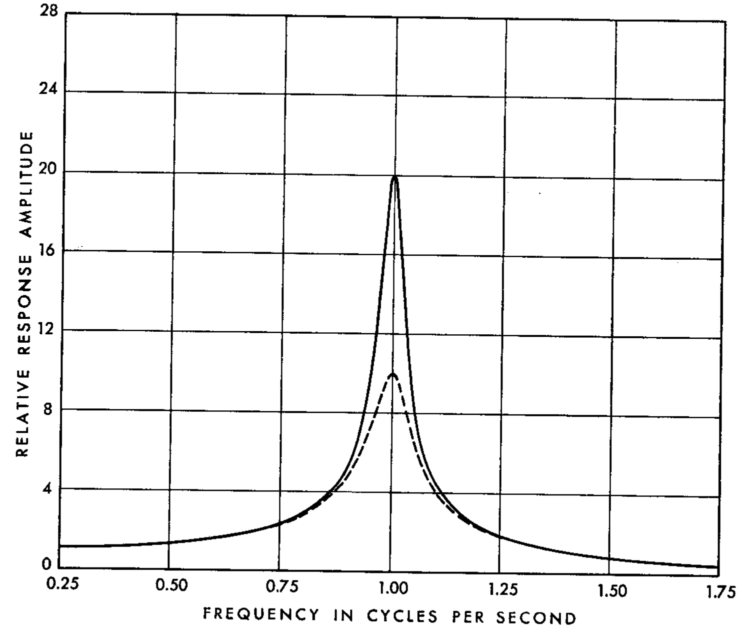


Fig. 5.4 Two frequency response curves with one cps resonant frequency. Dashed line shows oscillation with greater damping.

circuit that does this is shown in Fig. 5.3. The voltage generator is connected in series with the capacitor and inductor; that is, it is connected between the inductor and one terminal of the capacitor. A voltmeter is connected across the capacitor to indicate the amplitude of the sinusoidal voltage between the capacitor's terminals. The resistor, R, shown in the circuit, accounts for the damping of the system.

As we vary the frequency of the sinusoidal voltage generator, we can plot the ratio of the capacitor's voltage amplitude, V_c, to the generator's voltage amplitude, V_g, to obtain the frequency response. Fig. 5.4 shows two frequency response curves of this type. Note that this is identical to Fig. 3.4.

CHAPTER 6 The Electronic Vowel Synthesizer

We saw in Chapter 4 that the human speech production process has two essential parts. First, the vocal cords, which produce a sort of buzz—a sound wave whose spectrum contains many frequencies more or less evenly distributed over the audible range. Second, the vocal tract, which, acting as a resonator, brings up certain bands of these frequencies and reduces others, so that the spectrum of the speech sound wave has peaks that are associated with formants. But the individual actions of these speech organs—the vocal cords and the vocal tract—are not unique to speech production. It is possible to get this combination of a buzz source and a resonator by means other than the human vocal organs. It is reasonable to expect, therefore, that we can duplicate their combined actions and produce artificial speech.

Suppose, for example, that we build an electronic generator whose voltage spectrum contains many frequencies, like the spectrum of the vocal cord buzz. Suppose we connect this generator to an electrical circuit whose frequency response has several peaks corresponding to the formant resonances of the vocal tract. The circuit's output, then, should have a spectrum similar to the spectrum of the sound pressure produced when we make a vowel sound. Earphones and loudspeakers are devices that generate pressure variations in air proportional to the voltages and currents applied to them. We could connect the output of the circuit described above to a loudspeaker (through an amplifier, perhaps, to boost the power), and its output should sound very much like a steady speech sound.

The principles underlying such a device are common to most speech synthesizers; some sort of generator with a broad spectrum is used to excite a frequency selective element (a resonator) whose response has a number of peaks. The response of the resonator is then converted into sound pressure. The actual forms of the generator and resonator can be quite varied. As we saw in Chapter 1, the earliest synthesizers used mechanical vibrators or acoustic resonators. Helmholtz, a famous 19th century German physicist, made a speech synthesizer using tuning fork resonators. One of Alexander Graham Bell's early ideas for a telephone was, in a sense, a speech synthesizer that used vibrating reed resonators. Modern digital computer synthesizers can generate speech waves by solving the equations that describe the behavior of resonators. We will have more to say about this sort of synthesizer in Chapter 8.

Our synthesizer will be of a still different type. In it, electrical circuits consisting of capacitors and inductors are the necessary resonators; electrical voltages and currents are generated and converted to sound pressures by an earphone.

ELECTRICAL, MECHANICAL AND ACOUSTIC ANALOGS

We have touched on the similarities between electrical and mechanical resonators. We should emphasize that this similarity is not merely a superficial resemblance, but a parallel in the basic laws that govern these different phenomena—a similatity that is apparent in the mathematical equations used to describe such systems. For example, the equations for describing the behavior of a capacitor are identical to those for describing the behavior of a spring; the equations that describe the response of an electrical resonator are identical to those that describe the response of a spring-mass system. The only difference, as we mentioned in Chapter 5, is in the names of the variables. And the similarity does not stop with resonators. Indeed, the actions of a very large class of mechanical and acoustic systems can be duplicated, in close detail, by electrical circuits; conversely, electrical circuits can be duplicated by acoustic and mechanical systems. These duplicates or scale models are called *analogs*. Basically, an analog pair is a pair of systems not actually identical, but which resemble each other in behavior once the proper correspondence between their variables is recognized.

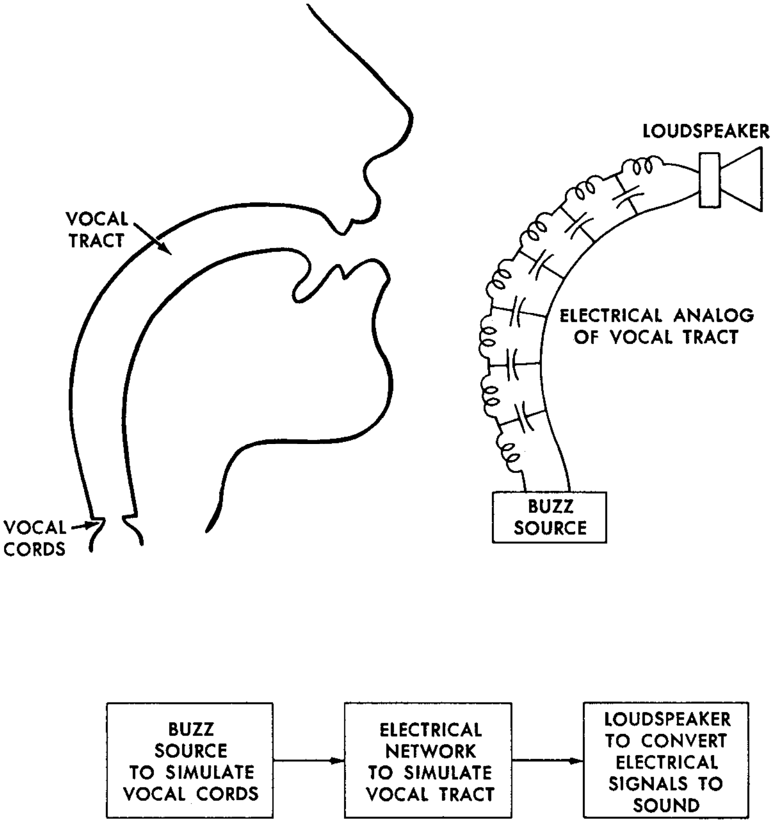

VOCAL TRACT
VOCAL CORDS
LOUDSPEAKER
ELECTRICAL ANALOG OF VOCAL TRACT
BUZZ SOURCE
BUZZ SOURCE TO SIMULATE VOCAL CORDS
ELECTRICAL NETWORK TO SIMULATE VOCAL TRACT
LOUDSPEAKER TO CONVERT ELECTRICAL SIGNALS TO SOUND

Fig. 6.1 A direct or exact analog of the vocal tract.

Vocal Tract Analogs

We can build an electrical analog of the vocal tract. Since the current-voltage behavior of an inductor is analogous to the inertia or mass behavior of air, and since the behavior of a capacitor is similar to the spring-like behavior of air, it is possible—quite literally—to build an electrical, artificial vocal tract. For example, you might use the array of capacitors and inductors shown in Fig. 6.1.

As a sort of scale model, the capacitance of a given section of the electrical circuit is made proportional to the volume into which air can be compressed in a corresponding length of the vocal tract. The inductance of a given section of the circuit is made proportional to the effective air inertia of a corresponding section of the tract. As it turns out, this means that the capacitance of a given section of the model must be proportional to the cross-sectional area of the corresponding vocal tract section, and that the inductance must be inversely proportional to the same cross-sectional area. With proper design, the voltages and currents in the model tract can be scaled-replicas of the sound pressures and particle velocities in the real vocal tract.

In a model of this type, there is a *direct* correspondence between points on the model and points on the vocal tract; in other words, a direct correspondence between voltages on the model and sound pressures in the tract. This kind of model is called a *direct* or *exact* analog.

For a simple vowel synthesizer, a direct analog vocal tract is not the easiest to build or use. Values of the electrical elements must be based on measured cross-sectional areas of the vocal tract. These are difficult to obtain without resorting to X-ray photographs or similar measurements. Furthermore, such techniques have so far not provided data of sufficient accuracy.

There is another type of model, however, a *terminal* analog, which is not based on a one-to-one *direct* correspondence between internal elements of the model and vocal tract parts, but only on a similarity in the over-all frequency response of the analog vocal tract network. In other words, with this type of model the designer is not seeking to duplicate the manner in which the vocal tract produces a given resonance pattern, but only the *end result.*

The fact that a terminal analog matches only terminal characteristics does not necessarily mean that it is simpler than a direct analog. We might, if we tried, make some pretty complicated circuits to reproduce the frequency response of the vocal tract. But it turns out that there are simple circuits that will duplicate all the vocal tract resonances below 3000 cps, the most useful band of frequencies for speech. As an added advantage, these models are described in terms of formant frequencies—data relatively easy to measure. There are comparatively large amounts of information on the formants of English vowels; measurements from several sources that agree when cross-checked. Your synthesizer will be a terminal analog of the vocal tract: one that produces sounds with spectral shapes *like* those of spoken vowels, but does not produce them the same way the vocal tract does.

YOUR VOWEL SYNTHESIZER

One way of producing the spectral peaks needed to simulate the formants of natural speech is to use simple electrical resonators. This type of terminal analog is called a *formant resonator analog.*

A single-inductor, single-capacitor resonant circuit has only one peak. The frequency response of the vocal tract is more complicated; after the first peak, there are many others. One can show, however, that the complicated frequency response of the vocal tract can be expressed as the product of many simple frequency responses, each similar to the response of an inductance-capacitance resonator. That is, one can use a simple resonance curve for each formant of the vocal tract and get the tract's complete response at any frequency by multiplying

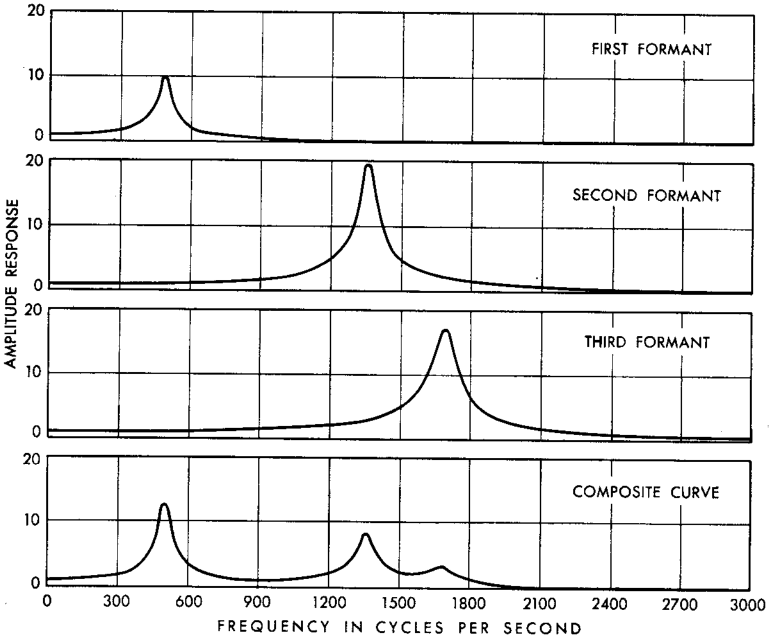


Fig. 6.2 A composite resonance curve built up from three simple resonances. The value of the composite curve for any frequency is found by multiplying the values of the three simple curves.

the values of the individual formant resonance curves at the same frequency. This principle is illustrated in Fig. 6.2.

This is a good clue to keep in mind when building the vowel synthesizer. To duplicate the frequency response of the vocal tract, we use one inductance-capacitance resonator for each formant. Since there are just three formants below 3000 cycles, and since most vowel sounds are determined by the frequencies of these formants, we use only three inductance-capacitance resonators in our vowel synthesizer. It is only necessary to connect the resonators so their over-all frequency response is the product of their individual responses. To do this, we use the output of one resonator as the input to the next in line.

One precaution must be taken in using the output of one resonator as the input to another. When we determined the frequency response of a resonator circuit, we assumed that nothing was connected to its output; that is, we assumed that no current was drawn from the output terminal. But if a substantial current *is* drawn from the output terminal, the frequency response changes. The height of the resonance peak can get lower, or the resonant frequency can change, depending on the type of *load.* Since neither of these effects is desirable—you would like the calculations to be simple—care must be taken to avoid drawing anything more than small currents from the resonator outputs. In our synthesizer, we use two techniques to reduce loading effects to a negligible amount. To minimize the current drawn from the first resonator, we design the second resonator so that the currents in its components are much smaller than those in the first resonator's components. That is, since the current through an inductor is inversely proportional to the inductance (for a given voltage and frequency), we make the inductor of the second resonator larger than that of the first. And, since the current drawn by a capacitor is proportional to the capacitance (for a given voltage and frequency), we make the capacitance of the second resonator smaller than the first's.

This principle could be applied to the second and third resonators as well, but then the size of the inductor would get quite large and the capacitor quite small. Moreover, the amplifier connected to the output of the third resonator would have to draw *very* little current because of the already small working currents in the third resonator. To avoid this, we put an

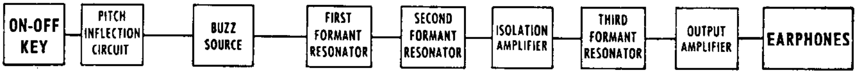


Fig. 6.3 Block diagram of the complete three-formant vowel sythesizer.

amplifier between the second and third resonators. The amplifier's function is to reproduce the voltage at the output of the second resonator, without drawing a large current from it. To keep from overloading the third resonator, we use a similar amplifier between it and the earphone. Fig. 6.3 is a block diagram of our complete *three-formant vowel synthesizer.*

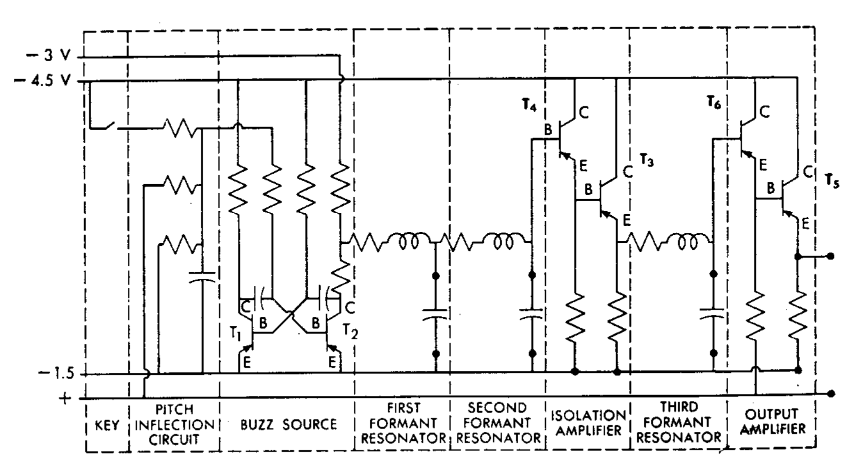
The Buzz Source The spectrum of a speech synthesizer buzz source should resemble the spectrum of the vocal cord wave. It should consist of the harmonics of a fundamental frequency between 100 and 150 cps, which is roughly the pitch of a male speaker's voice. To make sure that all formant resonances are excited, the harmonics should be smoothly distributed over a range extending to three or four thousand cycles per second —that is, over a range extending beyond the frequency of the highest formant in our synthesizer.

One simple waveform that satisfies this requirement is a train of very short pulses. A pulse train has a spectrum whose harmonics are multiples of the rate at which pulses are generated—that is, the pulse repetition rate is the fundamental frequency. The spectrum is relatively "flat" (contains harmonics of about equal amplitudes) up to a frequency whose value is one-half the reciprocal of the pulse width. This means that pulses shorter than about 1/5000 second, repeated at a rate of 100 to150 pulses per second, have the spectrum we need.

Our buzz source must meet one other requirement. The voice pitch of an adult male is usually somewhere between 100 and 150 cps. But this pitch seldom remains constant. As we talk, we continually change the pitch of our voices to add emphasis. Even when we say a single vowel, pitch changing is so much a part of our speaking habits that we automatically allow the pitch to rise or fall while we make the sound. For this reason, it is desirable to have a variable frequency buzz source. The synthesizer, therefore, has a key for turning the buzz source "on" and "off," and an additional circuit that causes the buzz source frequency to go up when the buzz is turned on and down as it goes off. This more natural inflection adds to the intelligibility of the sounds produced. The output of the buzz source drives the resonators which, through an amplifier, drive the earphone.

The Synthesizer Circuit Now that we have gone over the general plan for building the vowel synthesizer, we can examine its circuits in some detail. Look at the circuit diagram on the bottom of the box in which the experiment is packaged. You will use this box and the stiffening cardboard inside as a chassis for the synthesizer. The lighter area, marked *Buzz Source,* contains the buzz generator and its pitch control circuit. The three formant resonators are located on the right-half of the chassis. The amplifier that isolates the second and third formant resonators is above the first and second resonators, and the output amplifier is above the second and third resonators. The circuit in Fig. 6.4 is the same as the one on the box, except that conventional circuit symbols have been used and the locations of components on the drawing have been changed to make the circuit resemble the block diagram of Fig. 6.3.

The circuit used for the buzz source is known as a *multivibrator.* It has two transistors that operate as switches. Depending



TRANSISTOR TERMINALS: E-EMITTER; B-BASE; C-COLLECTOR

Fig. 6.4 Circuit diagram of the complete three-formant vowel synthesizer.

on the voltage between its base and emitter terminals, each transistor is either a short-circuit or an open-circuit between the collector and emitter. The two transistor switches are "closed" alternately. The circuit is so arranged that when one transistor switch is "closed," the other is automatically "opened." The time one transistor stays "off" depends on the sizes of the capacitor and resistor connected to its base, and on the voltage to which the other end of the resistor is connected. In our circuit, the resistor and capacitor connected to the base of transistor T_1 determine the length of the pulses generated. The resistor and capacitor connected to the base of transistor T_2 determine the pulse repetition rate. The resistor for T_2 is connected to the control key through a circuit designed to produce pitch inflection. When the control key is depressed, the voltage rises slowly to a value that makes T_2's off-time about 1/140 second. With this arrangement, the pitch starts out at a low frequency and rises to about 140 cps as the key is held down. When the key is released, the pitch decays back to a low frequency and the buzz generator stops buzzing.

The short pulses of voltage are applied to the first formant resonator. The output of the first resonator is the input to the second resonator, and its output goes to the amplifier. The amplifier consists of two transistors. In this circuit, the voltage at the emitter of each transistor closely follows the voltage at its base. In each transistor, however, the current drawn by the base is only about 1/50 the current supplied by the emitter. This is the kind of amplifier we need to isolate the resonators—one that provides a replica of the resonator output voltage, but draws little current from the resonator. The two-transistor circuit supplies about 2500 times as much current to the third resonator as it takes from the second. The third formant resonator is connected to the earphones through a similar amplifier. The output of this amplifier can be connected directly to the earphones without drawing too much current from the third resonator. Each of the three resonators consists of a resistor, inductor and capacitor similar to the resonators we discussed in Chapter 5. Terminal clips are provided on the chassis for the capacitors. This allows you to insert capacitors for tuning the resonators to formant frequencies appropriate for different vowel sounds.

At this point, you can, if you wish, start building your synthesizer, following the procedure outlined in Appendix A. However, we recommend that you begin construction only after reading the rest of the book.

CHAPTER 7 Experimenting With The Synthesizer

Before assembling your synthesizer, you may want to know some of the things that can be done with it once it is put together. This chapter includes suggestions about a few experiments you can perform, as well as a discussion of certain properties of speech that can be demonstrated with the synthesizer.

Some of the experiments involve simple modifications to the synthesizer circuit, followed by your listening to its output to hear the effect of the changes. Sometimes, you may want to have other people listen to your synthesizer and comment on the sounds it makes.

Even though we know a great deal about the characteristics of speech sounds, we know less about the relative importance of these characteristics to the listener's understanding of speech. On which features of the speech sound does the listener rely most for his understanding? Can he identify vowels from just two or three formants? Are the fluctuations of vocal cord pitch and pulse shape important to the listener? How mechanical would speech sound if the pitch were held precisely constant?

In investigating these and related questions, synthesizers can serve as useful tools for extending our knowledge about speech. Unlike a human speaker, a synthesizer can be made to produce sounds that embody some of the characteristics of real speech, but not all of them. In addition, the features it retains can be carefully controlled; consequently, by experimenting with synthetic speech sounds, we can learn something about the importance of various acoustic features to speech perception.

Our synthesizer's output is a greatly simplified representation of human speech. For example, it produces only vowel-like sounds with up to three formants. Furthermore, its pitch usually glides smoothly from one value to another, without any of the variability found in real speech. But even these features may be eliminated or varied systematically, one at a time. We can make sounds with only two formants, or with just one; we can change the frequencies of the formant resonances; and we can make the pitch of the synthesizer a monotone. Such flexibility enables us to experiment with these characteristics and note the importance of each to speech intelligibility and naturalness.

The following five experiments are examples of the type of thing you might want to do with the synthesizer. The list is by no means exhaustive. Nor are these experiments necessarily the simplest or most interesting to conduct. For example, we do not mention the possibility (for those with some knowledge of transistor circuits) of altering the waveshape produced by the Buzz Source to hear (and see on an oscilloscope) the effect this has on the sounds produced. The ambitious and patient experimenter can keep occupied with this kit for a long time.

The first experiment discusses the basic use of the synthesizer as a generator of vowel-like sounds. The others suggest further experiments you may find interesting and enjoyable.

EXPERIMENT I—SYNTHESIZING VOWELS

When you have assembled your synthesizer, as outlined in Appendix A, you should be ready to tune it to produce vowels. To do this, you have to choose capacitors that make the synthesizer resonate at frequencies equal to the resonant frequencies (formants) of the vowel sounds.

You will recall that the frequency of a resonator is given by the formula

$$f = \frac{1}{2\pi\sqrt{LC}},$$

where L is the inductance and C the capacitance of the circuit. Rearranging this equation, we see that the capacitance needed to resonate with an inductor at a frequency, f, is

$$C = \frac{1}{4\pi^2 Lf^2}.$$

The values of the inductors in your synthesizer are 1.45 henries for the first and third formants, and 6.0 henries for the second formant. Consequently, if you want to tune the first formant to 500 cps, the capacitor, C_1, would have to be

$$C_1 = \frac{1}{4\pi^2 \times 1.45 \times 250{,}000} = \frac{10^{-6}}{1.45 \times \pi^2}.$$

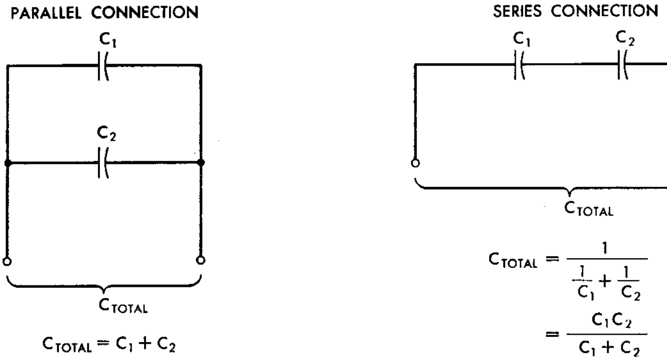


Fig. 7.1 Rules for computing the capacitance of capacitor combinations.

This approximately equals

$$\frac{10^{-6}}{14.5} = \frac{1}{14.5} \times 10^{-6} \text{ farads } = 0.07 \text{ microfarads.}$$

You can use this formula to compute capacitor values for other frequencies, or you can use the *Resonance Computer* included in the experiment kit. Instructions for its use are printed on the front of the "computer."

Eleven capacitors are provided for use in the formant generators. Their values are marked either in units of microfarads or picofarads, but you may find some capacitors whose values are marked only in numbers, with no indication of the units used. Generally, capacitors stamped with numbers less than 1 are in microfarads; numbers greater than 1 are in picofarads, unless some other units are specifically stated. A capacitor marked 700, then, would be 700 pfd., while one marked 0.01 would be 0.01 mfd. (10,000 pfd.).

In most cases, you should find that one capacitor is enough to produce the formant frequency you need. For some vowels, however, you will have to combine two capacitors to get the required frequency. The rules for combining capacitors are shown in Fig. 7.1. Notice, for example, that the total capacitance of a 700 pfd. and a 1500 pfd. capacitor in parallel is 2200 pfd., and that a 0.1 mfd. capacitor in series with a 0.05 mfd. capacitor has a combined capacitance of 0.033 mfd. Using the 10 capacitors singly or in combinations, you should be able to produce most of the vowel sounds.

We discussed the formants of English vowels in Chapter 4. Data on the formant frequencies appropriate for the various vowels is given in Fig. 7.2 and Table 7.1. In the graph, the first formant is plotted against the second formant for various vowels spoken by a number of people. The vowel sounds were pronounced in single syllable words like h*ee*d, h*a*d and h*o*d. The first two formants of any spoken vowel define a point on the graph. Contours are drawn around areas typically associated with particular vowels.

Take a good look at the graph. You can see that the first two formant frequencies are important cues for identifying vowel sounds. But for good recognition, listeners use more information than just the first two formants; for instance, the higher formants, the identity of the speaker as a male, female, child or adult, and even subtle cues about the regional accent ot the speaker that listeners can derive from pitch inflection and timing. For example, listeners can correctly identify the vowel sounds "aw" (as in c*a*ll) said by a child, and "ah" (as in f*a*ther) said by an adult, even though the two sounds have very similar first and second formants.

Since your synthesizer can reproduce none of these "extra" details (except the third formant), it is wise to pick formant frequencies that do not force listeners to rely heavily on extra cues; that is, to pick first and second formant frequencies that

Table 7.1—A Tabulation of the Average Formant Frequency values Shown in Fig. 7.2

	ee	ɪ	e	ae	ah	aw	ʊ	oo	ʌ	er
First Formant Frequency										
Male:	270	390	530	660	730	570	440	300	640	490
Female:	310	430	610	860	850	590	470	370	760	500
Second Formant Frequency										
Male:	2290	1990	1840	1720	1090	840	1020	870	1190	1350
Female:	2790	2480	2330	2050	1220	920	1160	950	1400	1640
Third Formant Frequency										
Male:	3010	2550	2480	2410	2440	2410	2240	2240	2390	1690
Female:	3310	3070	2990	2850	2810	2710	2610	2670	2780	1960

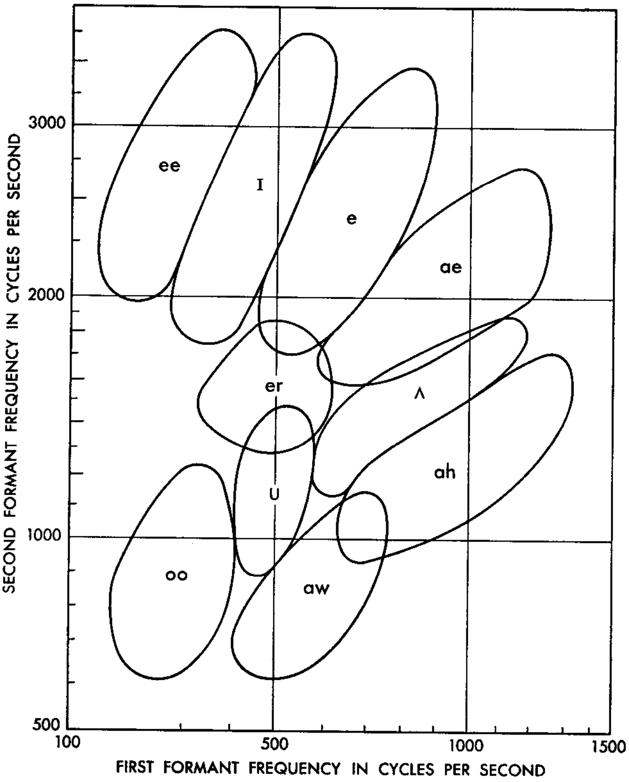


Fig. 7.2 Graph of first formant frequency versus second formant frequency for a variety of vowel sounds.

are in the centers of the areas encircled on the graph in Fig. 7.2. On your first try, you can use the third formant frequency values for males, as indicated by Table 7.1. You can experiment with frequencies 10 or 20 per cent higher and lower than the table values to get what you think are good vowel sounds.

You might use the following example in choosing your resonator capacitors. Suppose you want to synthesize the sound "ah" (as in *fa*ther). From the graph, you pick the first and second formant frequencies whose values are near the center of the "ah" contour; for the first formant, this is about 900 cps and for the second formant about 1200 cps. The value of the third formant —which you get from the Table 7.1—is 2400 cps. Using these frequencies, you calculate the capacitor values as $C_1 = 0.022$ mfd., $C_2 = 0.003$ mfd., and $C_3 = 0.003$ mfd. As a first try, you can select from the kit's stock the single capacitors whose values are closest to these values.

Now suppose you try the synthesizer with these capacitors and find that its output sounds more like "ʌ" (as in h*u*t) than "ah" (as in f*a*ther).* This can easily happen because the values of our capacitors and inductors are accurate to only ±10 per cent. Moreover, the direct connection between the first and second resonators tends to make these two formant frequencies somewhat further apart than the calculations indicate. If you add to this the fact that you have made some round number approximations in choosing your capacitors, you can appreciate why the first try might be slightly off the mark. Your synthesized speech sounds lack the additional cues that would overcome such discrepancies.

Assuming that the synthesizer *does* make the vowel sound "ʌ" when you want "ah," you can go back to the graph to determine what changes should be made to improve the "quality" of the vowel. You should notice immediately that the sound "ʌ" has a lower first formant frequency and a higher second formant frequency than the sound "ah." On the next try, the second formant frequency should be lowered; this calls for a larger capacitor in the second resonator.

Suppose this new value leads to an "aw" sound. Referring to the graph again, you can see that the chief distinction between this sound and the one you want is the difference in the first formant frequency. You should raise the frequency of the first formant; this requires a smaller capacitor in the first resonator.

You can see the value of the formant frequency graph. You use it first to locate the values of the formant frequencies for the vowel sound you want to synthesize. If your first attempts are incorrect, the graph helps you determine what changes are necessary.

After you have put in a set of capacitors, you should be very

* When experimenting with the synthesizer, you should depress the key for very brief intervals, about one half second. This uses the pitch inflection circuit to its best advantage.

critical in deciding whether or not the sound your synthesizer makes is the vowel sound you are trying to generate. Your emotional bias as builder of the synthesizer might incline you to accept a rather inexact sound for the vowel you wanted, even though an indifferent observer would identify the sound as another vowel altogether. To get the best sound from your synthesizer, you must first know when the sound is not quite correct and, when it is not, what sound you are actually synthesizing. This will help you determine the necessary changes.

EXPERIMENT II—THE EFFECT OF CONTEXT

Our understanding of a given speech sound or word depends upon the context in which it is heard. By context, of course, we mean, among other things, the rest of the sentence or paragraph, the general topic of conversation and the identity and attitude of the speaker. An aircraft pilot, for example, can understand a control tower message even when reception is so poor that most of us would understand none of it. Of course, he has more practice at listening to poor radio signals than we have but, more important, he has heard similar messages expressed in almost the same words and phrases. To him, the tower messages are like the words of a familiar song.

When we speak of context, then, we really mean everything the listener knows about what he is going to hear or what he has just heard. Indeed, people tend to hear what they expect to hear, and this depends on more than just the sound waveform itself. Synthetic speech, for example, is often most intelligible to the person who makes it. This is only partly due to his enthusiasm and emotional bias. More important is the fact that he already knows what the sounds he hears are supposed to be.

This experiment involves someone who knows nothing at all about your synthesizer—especially that it is supposed to produce speech sounds. The purpose is to determine the extent to which prior knowledge influences the listener's impressions of the sound.

First, set your synthesizer to produce what you think is a good vowel sound. Go to your uninformed listener and pose a question like this:

"*Here is an electronic device I've made to produce a certain sound. Tell me what it sounds like to you.*"

Do not give him any hint that the synthesizer's output should resemble speech; avoid using words and phrases like "*speech*," "*talk*," or "tell me what it *says*."

After demonstrating the sound two or three times, ask the listener to tell you what it sounds like to him. Since he has no idea of what it should be, it is quite possible that he will not identify it as a speech sound. Then give him a little more information by asking:

"*Suppose that this is a speech sound. What speech sound does it sound like?*"

Demonstrate it several more times and get his opinions. If he still fails to recognize the sound, give him additional information; for example, that it is a vowel sound and, possibly, that it is one of three or four vowels you pronounce for him.

EXPERIMENT III—PITCH AND NATURALNESS

You can modify the synthesizer to eliminate its built-in pitch inflection. The synthesizer's 40 mfd. capacitor produces pitch "glide" by allowing the pitch control voltage to rise to its final value at a slow rate. To give the synthesizer a monotone pitch, disconnect the 40 mfd. capacitor from the terminal where it joins the 15 K ohm (brown-green-orange) resistor. You can easily connect and disconnect the capacitor by touching it to the resistor wire.

Notice the difference that monotone pitch makes. You may wish to gather some statistics on this effect by trying it on some friends. Play several vowels for your friends with the capacitor disconnected so that the pitch is monotone. Depress the key several times for about a second each time and ask them to identify the vowels. See if your friends do better when you reconnect the capacitor to produce pitch inflection.

EXPERIMENT IV—THE IMPORTANCE OF FORMANTS

In your synthesizer, it is possible to eliminate any of the formant resonators. You can do this by placing a piece of wire across the inductor and removing the capacitor.

The procedure for this experiment will be to remove various formants from each of the vowels and note the results. You will find that eliminating either of the first two drastically changes the sound; eliminating the third formant has little effect on most of the vowel sounds, but it is an important feature of others.

EXPERIMENT V—OBSERVING WAVEFORMS

For this experiment, you will need an oscilloscope; your teacher can help you get one at school and, perhaps, tell you something about how to operate it.

Set the oscilloscope to display about two cycles of the synthesizer's waveform. Check the synthesizer's output voltage for various vowels. Carefully observe the difference between changing the pitch of the sound—as you press and release the key—and changing the frequency of the first formant.

See if you can pick out the formant frequencies in the waveform. In the sound "e" (as in b*e*t), for example, the first formant will ring about four cycles in one pitch period, while the second formant will ring about 12 to 15 cycles.

To see the different formants more clearly, eliminate two of them, using the technique suggested in Experiment IV. After studying the formants separately, you may be able to identify them more easily in the complete vowel waveform.

Notice how the amplitudes of the higher formants change as you alter the frequency of the lowest formant. You will understand why this happens if you re-examine Fig. 6.2. The output spectrum is the product of the Buzz Source spectrum and the frequency response of the three formants. Since the lower formant resonators pass less of the high frequencies as their own natural frequencies are decreased, there is less high frequency energy to excite the higher frequency resonators. This decrease in the amplitudes of the higher formants is characteristic of naturally produced speech.

You may find it interesting to observe some of the waveforms inside the synthesizer. Clip the ground side of your oscilloscope input to one of the battery terminals and the "live" side of the oscilloscope input to the top terminals of the resonator capacitors; see how the three formants are "added" to the signal as it progresses from the Buzz Source to the output.

CHAPTER 8 Complete Synthesis Systems

The synthesizer you build as part of this experiment can produce only steady vowel sounds. As we saw in Chapter 4, however, many other classes of sounds are used in continuous speech. Synthesizing continuous speech, therefore, requires a much more elaborate device than your vowel synthesizer.

In this chapter, we will discuss some techniques for synthesizing continuous speech. We will describe synthesizers used in research laboratories for generating whole sentences, and a few experiments conducted with these synthesizers. We will explain the value of synthetic speech experiments in extending our knowledge about human speech.

THE ACOUSTIC CHARACTERISTICS OF SPEECH

A complete speech synthesizer should be able to duplicate the acoustic effects produced by the human vocal organs. Let us take a closer look at these effects.

The Sound Spectrograph

A special machine has been developed to study the acoustic characteristics of speech sounds. This machine, the *sound spectrograph,* plays such an essential part in speech research that we will describe it before discussing the data it produces.

The sound spectrograph shows how the spectrum of a speech wave varies with time. The speech spectrum is plotted as a sort of continuous graph. The horizontal axis represents time and the vertical axis is the frequency scale. The darkness of the spectrogram at any point shows the relative intensity of a particular frequency at a certain time. A dark area indicates high intensity and a light area low intensity. A spectral peak, such as one made by a formant, produces a dark area at a point along the vertical axis corresponding to the formant's frequency. If that formant fre-

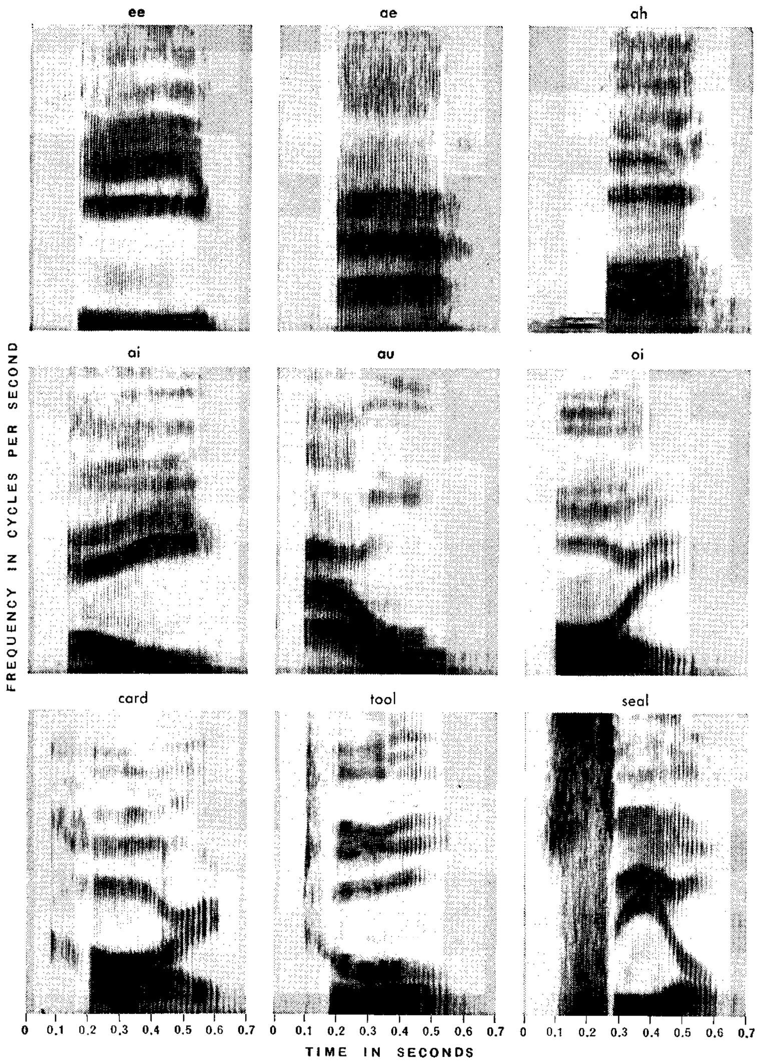


Fig. 8.1 Typical speech spectrograms. The speech sounds represented are shown above the spectrograms.

quency is left unaltered, we get a horizontal dark bar whose length depends on how long the formant frequency is kept constant. When the formant frequency is increased, the dark bar bends upward; when it is decreased, the bar bends down. The dark bar disappears when we stop making the sound.

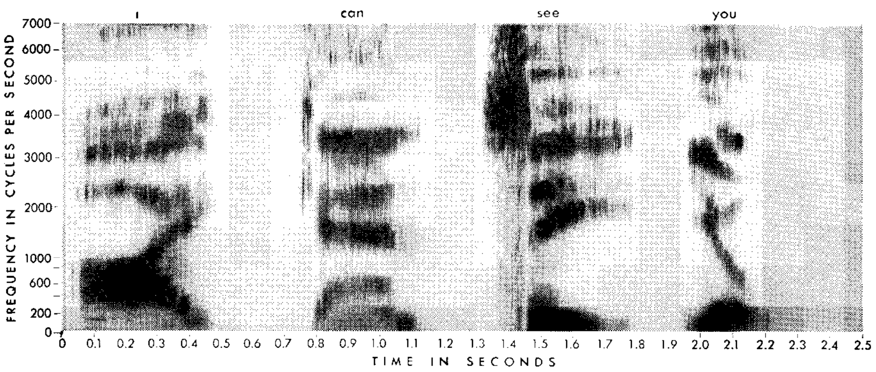


Fig. 8.1 continued

The spectrograms at the top of Fig. 8.1 show what happens when pure vowels are pronounced in isolation. The first four formants can be seen clearly; they remain constant in each of the spectrograms because the articulatory position remains unaltered. The patterns for a variety of diphthongs are in the line below. They show how the formants change as the shape of the vocal tract is altered during a diphthong. The third line shows how a consonant at the beginning or end of a pure vowel makes the formants vary. The portion of Fig. 8.1 on page 99 shows how markedly the patterns vary for a spoken sentence.

The spectrogram of a sentence said at normal speed is shown in Fig. 8.2. We can see that the frequencies and amplitudes of the many spectral peaks vary continuously. For example, look at the last word of the sentence in Fig. 8.2: "rich," or "rɪtsh" as it would be written phonetically. At the beginning of the word,

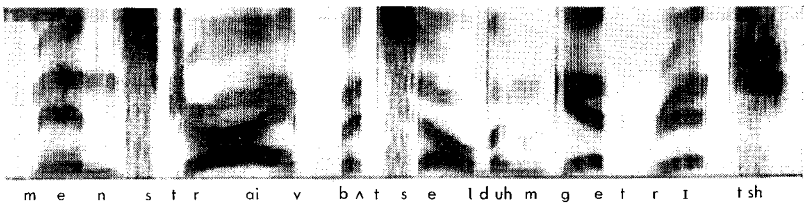


Fig. 8.2 Spectrogram of the English sentence, "Men strive but seldom get rich."

three formants are visible. The frequency of the first formant remains relatively constant, but the other two rise rapidly. Soon, the fourth, fifth and sixth formants also are visible. The frequencies of all the formants remain comparatively stable during the middle section of the word. This is followed by a sudden silence as the air flow is interrupted for the sound "t," shown by the blank segment near the end of the spectrogram. Finally, there is the fuzzy looking section when the fricative "sh" is spoken.

Looking at the whole spectrogram, the impression of many components moving in many different ways is unmistakable. Various segments stand out from the rest and catch our eye. There are the segments with the clearly defined dark formant bars and the very obvious, closely spaced vertical lines. These vertical lines are produced when the vocal cords vibrate during voiced sounds; each line corresponds to a single vocal cord cycle. There are also the blank segments indicating the absence of any sound when the air stream is stopped during plosive consonants. There are the decidedly lower intensity segments of the nasal consonants. The fricative consonants produce the fuzzy segments; they are darkest in the 4000 to 6000 cps region for "s," and in the 2000 to 3000 cps region for "sh."

The sound spectrograph shows how the acoustic characteristics of speech vary with time. By doing this, it has helped us recognize how essential this time-varying, "dynamic" feature of speech really is. At the same time, it suggests the degree of additional complexity involved in synthesizing complete sentences, rather than simple vowels. Certainly, there must be a means for varying formant frequencies, not as a series of discrete steps from one sound to another, but as smooth, continuous movements in frequency as time progresses. A "hissy" sound source must be provided for the synthesis of fricatives. There must also be provisions for switching back and forth between "voiced" and "fricative" sounds; and we must be able to stop both "friction" and "voicing" to produce the silent portions of the plosives. Finally, there must be a way of supplying simultaneous control signals to all parts of the synthesizer to keep them operating together.

COMPLETE SYNTHESIZERS

We see that many different features of the speech wave must be controlled simultaneously to produce continuous speech. Several techniques have been used to do this—none of them simple.

Early synthesizers—like those of Kempelen and Faber—were controlled manually. Mechanical buzzers, hiss sources and variable acoustic resonators were manipulated by hand to simulate the actions of the vocal organs. The 1939 World's Fair talking machine, the Voder (shown in Fig. 1.2 of Chapter 1), also was manually controlled. Ten keys were used to control the spectral shape of the sounds produced. Combinations of keys were used to produce the effect of formants. Other keys and a wrist bar controlled the outputs of the buzz and hiss generators and produced "clicks" to simulate the plosives. Pitch was controlled by a foot pedal. "Playing" the Voder was somewhat like playing an organ; the operator used a combination of finger, wrist and foot movements to keep the various parts of the synthesizer in step.

Speech synthesizers have improved since the day of the Voder, but their control is still complicated. Speech is so complex that persons who operate manually controlled synthesizers have to undergo extensive training—comparable to that of a skilled musician—before they can "talk with their hands." Modern laboratory synthesizers generally use some form of control that can be prepared in advance. Once a set of control data has been prepared, it can be run through the synthesizer many times, producing the same output every time. Synthesized speech generated with such techniques is not subject to errors of haste or confusion; in addition, there is none of the variability that arises when human beings try to perform the same operation twice in exactly the same way.

In one widely used control procedure, speech data is represented by graphs. Lines specifying the variations of the different acoustic features to be synthesized are plotted along a common time axis. The lines are drawn on a plastic sheet and scanned by a suitable device. The scanner "reads" the sheet along the time axis and develops voltages proportional to the positions of the different lines passed under it; these voltages are used to control the synthesizer.

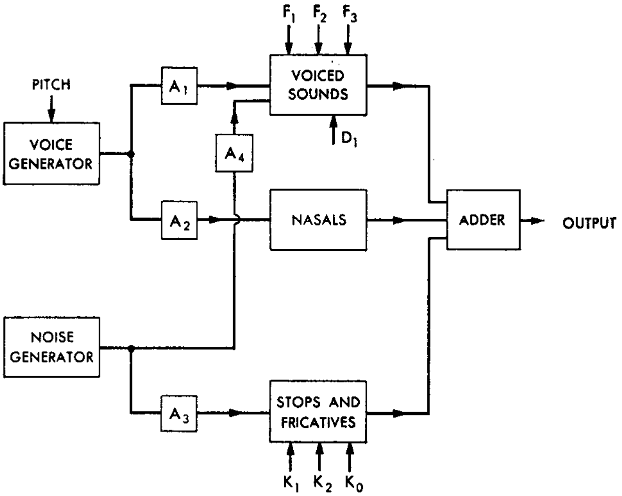


Fig. 8.3 Block diagram of the OVE II, a formant resonance speech synthesizer. A_1, A_2, A_3 *and* A_4 *are amplitude controls;* F_1, F_2 *and* F_3 *are controls for the first three formant frequencies (*D_1 *controls the damping of the first formant);* K_1, K_2 *and* K_0 *control the spectrum of fricatives.*

The OVE II Synthesizer

Some of the most realistic synthetic speech has come from the *OVE II*, a synthesizer built at the Royal Institute of Technology in Stockholm, Sweden. Fig. 8.3 is a block diagram of the *OVE* II. It has two branches excited by a buzz source to simulate voiced sounds; a third branch, excited by a hiss source, generates fricative and plosive sounds. Twelve signals are used to control the variable elements in the synthesizer. Four of these turn the buzz

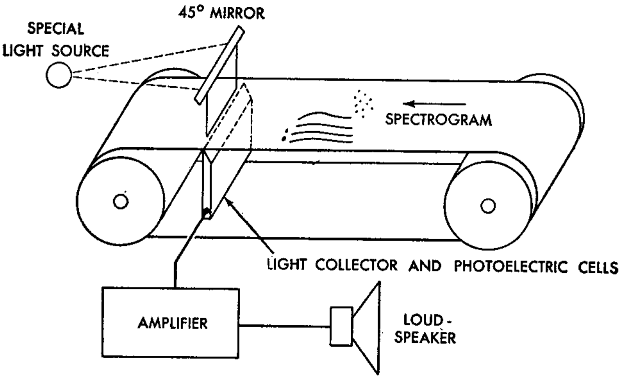


Fig. 8.4 Highly simplified schematic diagram of the Pattern-Playback.

and hiss inputs to each branch "on" and "off" and adjust the input amplitudes. Three control signals set the frequencies of the formants in the vowel-sound branch; another signal sets the damping of the lowest of these formants. Three signals control the shape of the fricative spectrum; the twelfth signal determines the pitch of voiced sounds. Resonances in the nasal branch have fixed values. The outputs of all branches are added together to produce the final synthesized speech sounds.

The input to the *OVE II* is a graph similar to the one described earlier. Curves drawn in conducting ink on a sheet of plastic represent the control signals. The sheet is clamped rigidly to a table and a scanning device moves over it on tracks parallel to the time axis of the graph; the corresponding voltages produced are used to control the synthesizer.

Producing recognizable, but poor quality, synthetic speech with the *OVE II* is relatively simple. People thoroughly familiar with its operation can start with a tape recording of a two-second sentence and produce the first synthetic copy in a few hours. But speech produced this way differs from the original in many respects. When the makers compare their first results with the original, they can hear—and see, after making spectrograms—many errors in the first simulation. With the insight gained by making this comparison, they can do a much better job on the second try. The synthesizing procedure goes on; new attempts are made, and these results are compared with the original human utterance. Eventually, the artificial and natural speech begin to sound more and more alike.

The results of such attempts at perfection can be quite striking. But the price paid for perfection, in terms of time, is very high. For example, researchers tried to duplicate a human utterance of the sentence,

"*I love the simple life.*"

The final synthetic sentence is so much like the original that even expert listeners have trouble identifying the "forgery." Many weeks were spent, however, in analyzing and correcting the synthetic copy to bring it this close to the original human utterance.

The Haskins Pattern-Playback Another fairly elaborate synthesizer is the so-called *Pattern-Playback*, built by the Haskins Laboratories of New York. Although not a complete synthesizer—it can produce only voiced sounds—

experiments with the Pattern-Playback have considerably increased our knowledge and understanding of the speech process. Fig. 8.4 is a schematic diagram of the Pattern-Playback. Spectrogram-like patterns are painted on a plastic belt; the signals needed to control the synthesizer are produced when the patterns are scanned by photoelectric cells.

The Pattern-Playback is a convenient device for determining which of the varied marks on speech spectrograms are important to speech recognition. In investigating this question through experiments with the Pattern-Playback, researchers used simplified, hand-painted versions of actual speech spectrograms. Fig. 8.5(a) is the sound spectrogram of a naturally produced sentence, and Fig. 8.5(b) is a pattern that was played on the Pattern-Playback to synthesize the same sentence. We will have more to say about the Pattern-Playback when we discuss synthetic speech experiments.

Computer Synthesizers

The synthesizers we have described so far are complicated electronic devices built for the sole purpose of generating artificial speech. Building such synthesizers and keeping them running require a great deal of time and effort. In addition, experimental circuit modifications or improvements are costly and time consuming.

Large, general-purpose digital computers offer another way of performing all the functions necessary for synthesizing speech. The time needed to get a computer-programmed synthesizer operating is usually much less than the time involved in building the equivalent, special-purpose electronic circuitry.

Electronic computers can carry out arithmetical operations—additions, subtractions, multiplications and divisions—at extremely high speeds. A list of instructions, called a *program,* is stored in the computer's memory; it specifies the sequence of operations to be performed. The actual order in which the computer carries out its instructions is not completely fixed by the program; using the results of previous calculations, the computer can make simple decisions about which portions of the program to execute at a given time. Specified in the proper sequence, these arithmetical and decision-making operations are sufficient to generate artificial speech.

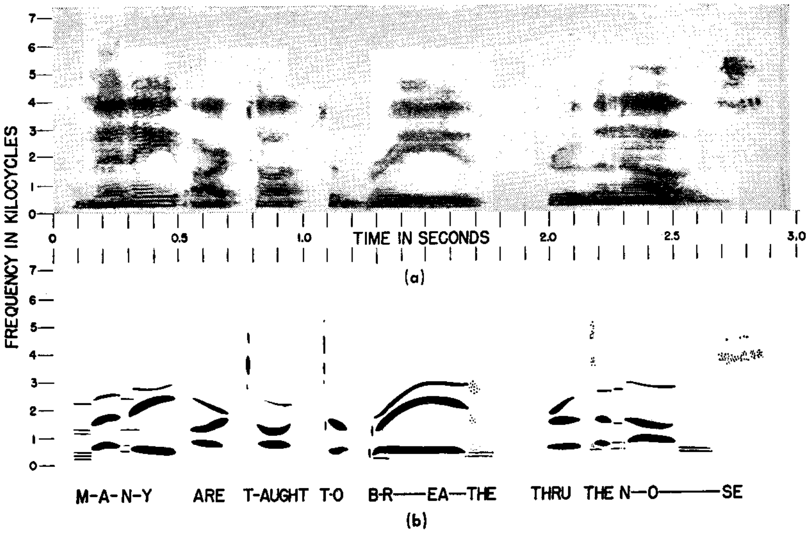


Fig. 8.5 The (a) portion of the figure is the sound spectrogram of a naturally produced sentence; (b) is the painted pattern that can be played on the Pattern-Playback to synthesize the same sentence. Only the first three formants of natural speech are represented in the painted pattern.

The result of the synthesis process in the computer is a sequence of numbers that defines the amplitude of the speech wave at successive instants of time. This sequence is converted to a voltage waveform by an electronic device and recorded on magnetic tape suitable for playing on a standard tape recorder.

EXPERIMENTS WITH SYNTHETIC SPEECH

Before ending this chapter, we will describe a few synthetic speech experiments which led to increased knowledge about the nature of the human speech process.

Much of our knowledge about consonant recognition has come from experiments on the Pattern-Playback. In painting patterns for this machine, it was found that a very short vertical mark, like any of the three shown in Fig. 8.6, was heard as a "plop"-like sound. Because of its similarity to the sound we hear when a plosive consonant ("p," "t" or "k") is pronounced, this "plop"-like sound is often called a *plosive burst.* The perceived character of the plosive burst made by the Pattern-Playback depended on the frequency at which the burst was

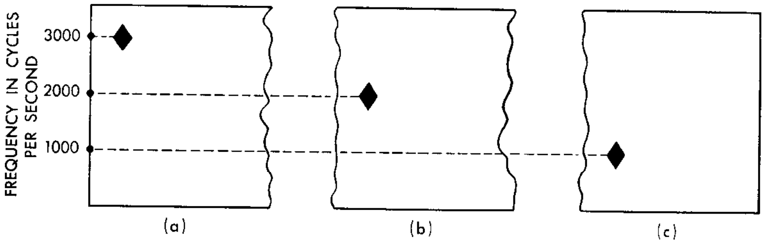


Fig. 8.6 Example of "plosive burst" patterns for the Pattern-Playback.

centered. In Fig. 8.6(a), the burst was centered at 3000 cps; in Fig. 8.6(b), at 2000 cps; and in Fig. 8.6(c), at 1000 cps.

Test syllables were generated from patterns in which a plosive burst was combined with a vowel section made up of two formants, as shown in Fig. 8.7. Many test patterns were made combining each of the 11 different vowel formant configurations with plosive bursts centered at a number of different frequencies. The test syllables generated by the Pattern-Playback were presented to a group of listeners. They were asked whether they heard the syllables as "tah," "pah" or "kah," as "too," "poo" or "koo," and so on. On the whole, no single plosive burst was consistently heard as the same plosive consonant. For example, a plosive burst centered at one frequency was heard as a "k'' when associated with one vowel, and as a "p" when associated with another vowel. In other words, the kind of plosive consonant

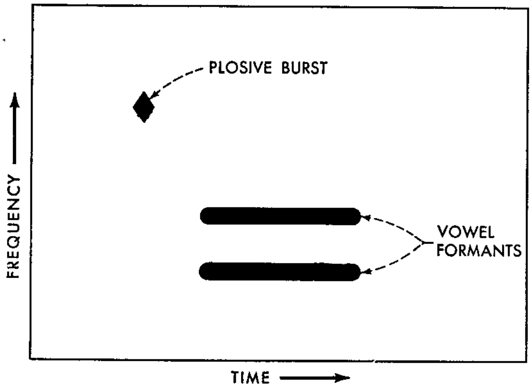


Fig. 8.7 A painted pattern which, when played on the Pattern-Playback, is heard as a plosive consonant followed by a vowel.

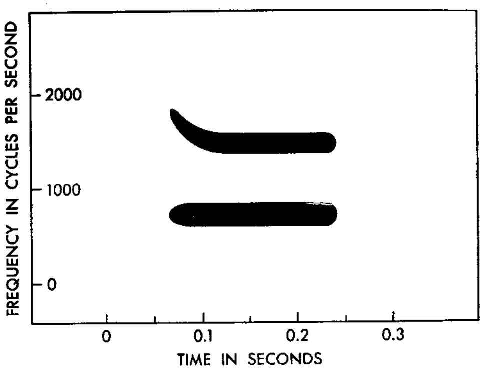


Fig. 8.8 An example of second formant transition.

we hear depends not only on the frequency of the plosive burst, but also on the nature of the following vowel. This finding was of great importance.

A steady vowel sound was heard when a pattern made up of two constant frequency formants was played on the Pattern-Playback. In experimenting with these artificial vowels, it was found that listeners heard a plosive consonant (even in the absence of a plosive burst) if the frequency of the second formant varied during the initial segment of the syllable. A typical pattern of this kind is shown in Fig. 8.8. The part of the second formant where the frequency varies is called the second formant *transition.* Patterns—like those in Fig. 8.9—were made up to test various degrees of upward and downward transitions; the test patterns were played on the Pattern-Playback and listeners were asked whether they heard the test syllables as "tah," "kah" or "pah," etc. The tests were repeated for each of the 11 English vowels. Again, as with the plosive bursts, it was found that the same kind of transition was heard as one plosive consonant or another depending on the vowel that followed.

More precisely, it was found that all the second formant transitions perceived as one particular plosive pointed toward (but never reached) the same frequency. This is shown by the composite sketch in Fig. 8.10, where we see all the second formants which, when heard on their own (combined with the first formant in the lowest part of the figure), were perceived as the consonant "t" followed by a vowel.

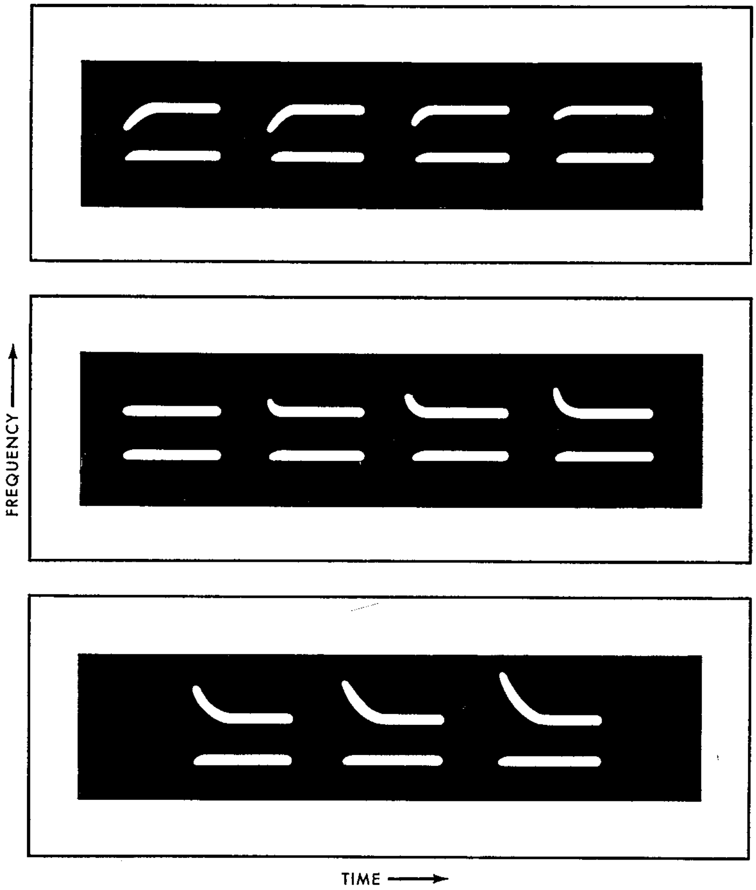


Fig. 8.9 Formants with various degrees of upward and downward transition.

For the other plosives, the second formant transitions pointed toward different frequencies; about 700 cps for "p," and, usually, about 3,000 cps for "k."

These same transitions were found to be important for the recognition of the nasals, "m," "n" and "ng." Fig. 8.11 shows the patterns which, when played over the Pattern-Playback, are heard as "pah," "tah," "kah," and "mah," "nah" and "ngah." We can see that the second formant transitions are identical for "pah" and "mah," for "tah" and "nah" and for "kah" and "ngah." The pairs of syllables differ only by what precedes the vowel segments: silence in the case of the plosives, and a low intensity buzz (heard vaguely as nasality) for the nasals.

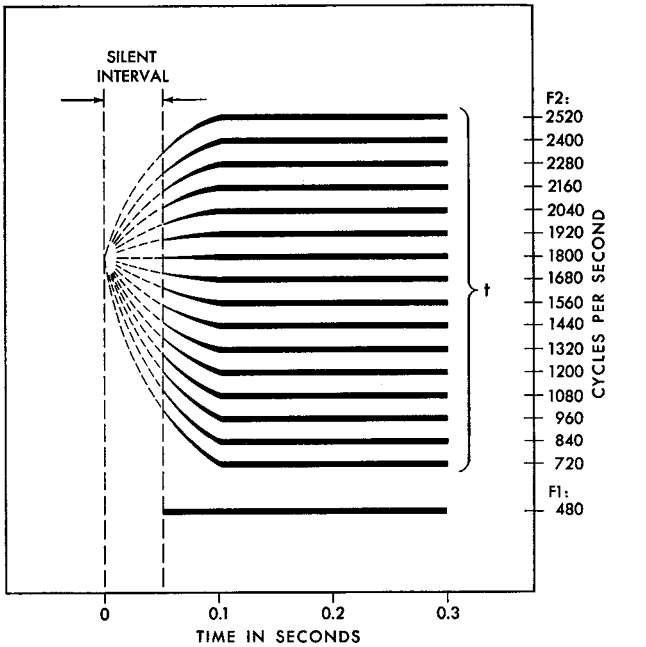


Fig. 8.10 Second formant tiansitions perceived as the same plosive consonant, "t."

The syllable pairs with *identical second formant transitions* are heard as consonants that have the *same place-of-articulation.* We conclude, then, that one way we can identify a consonant's place-of-articulation is by the nature of the second formant transition of its adjoining vowel.

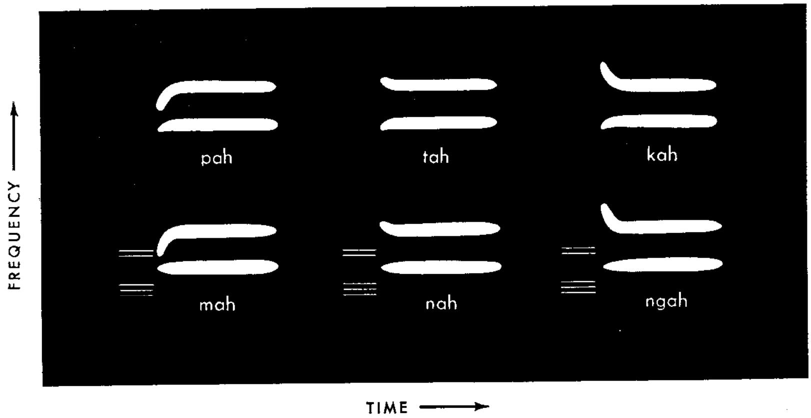


Fig. 8.11 Patterns showing the relationship between the second formant transition and the place-of-articulation of consonants.

The work of the Haskins group on the cues important for consonant recognition was eventually extended to include the development of a set of rules for synthesizing all the English phonemes. Whole sentences were synthesized according to these rules.

A computer synthesizer incorporating the Haskins rules was programmed at Bell Telephone Laboratories. The program simulated synthesizing processes similar to the ones in your vowel generator. To produce a vowel sound with your synthesizer, however, you have to specify separately the value of each formant frequency. If you wanted to generate continuous speech with your synthesizer, you would have to vary each formant frequency to get the proper formant transitions. The computer program, on the other hand, was so designed that the formant variations were automatically selected; formant frequencies appropriate for each phoneme—and the rules for making transitions between phonemes—were stored in the computer. To synthesize a given sentence, the experimenter had to specify only the phoneme sequences.

The quality of speech generated this way was found to be reasonably intelligible, provided the experimenter defined the pitch and duration of each phoneme individually. The pitch and duration appropriate for each phoneme depend to a great extent on the other phonemes in the sentence, and on stress and intonation.

So far, it has not been possible to formulate rules for controlling these variables in a way that will consistently yield natural or intelligible speech.

CHAPTER 9 A Look Toward The Future

Looking back over the previous chapters, it is clear that research performed during the past few decades has deepened our understanding of the processes involved in spoken communication. No doubt, present and future research will help illuminate areas where our understanding is at best incomplete and, at worst, incorrect. Of course, some of what we now believe to be true may someday be as outdated as the ancients' belief in air, earth, fire and water as the fundamental elements of the universe.

This is as it must be, for science is a vital, dynamic activity, not a static body of established truths. Only the rare, exceptional hypothesis remains for later generations to admire and accept as scientifically sound. Much scientific work does not stand up to the test of time. T. H. Huxley, an eminent 19th century biologist, put it this way: "It is the customary fate of new truths to begin as heresies and end as superstitions." Or, to quote him again, "The great tragedy of modern science [is] the slaying of a beautiful hypothesis by an ugly fact."

Research has a dual effect on human affairs. First, it leads to a better understanding of the world around us; it corrects our misconceptions and opens up new paths that were previously unimaginable or obscure. Second, this better understanding usually leads to the development of practical applications that directly affect our everyday lives. The X-ray machine, the television set, the transistor and the atomic reactor resulted from advances in technology that were based on advances in fundamental scientific understanding. In our present age of science and technology, the time lag between a scientific discovery and its impact on society is often very small.

We can expect research in speech synthesis to add appreciably to our knowledge about spoken communication. For example, it should help us learn more about the speech features essential for recognition and about those features that contribute to the natural, human quality of our voices, compared to the obvious "mechanical" characteristics of most synthetic speech.

The large scale digital computer will become an increasingly valuable tool in the hands of able research workers. Most speech synthesis experiments used to take months or years to complete because complex circuitry had to be built to carry them out. But this circuitry can now be "simulated" on a computer and the experiments take only days or weeks instead of months or years. As we saw in Chapter 8, speech synthesizers frequently take the form of programs or sets of instructions for computers to perform. By making relatively simple changes in the instructions, it is possible to alter the characteristics of the "talking machine" the computer is programmed to simulate. It would take much longer to accomplish the same changes if the talking machine actually existed as racks of conventional electronic circuitry, like elaborate versions of your synthesizer.

What practical applications are likely to result from a more complete understanding of the speech process at its many levels of importance? Undoubtedly, some of the most useful applications cannot be foreseen. But there are certain areas where applications are sure to be made. Some of these will involve speech synthesizers directly; in others, knowledge made available through synthetic speech experiments will be of great importance. We will consider applications in the areas of speaker identification, speech recognition and bandwidth compression systems, and in the general area of talking machines.

SPEAKER IDENTIFICATION

One of the remarkable features of the human voice is that it conveys much more information than is necessary to distinguish the words spoken. We can ordinarily tell from a person's voice whether he is happy, sad or angry, if he is asking a question or making a statement of fact. We can usually tell if the speaker is a man or a woman and, to a great extent, we can recognize a familiar voice and associate it with the person it belongs to.

What are the features of a speech wave that are peculiar to a particular speaker? We have only partial answers to this question. Voice "quality" is certainly important. This is affected by physiological factors, such as the size of our vocal cavities and the way our vocal cords vibrate. Other significant factors, such as speech timing and regional accent, are affected by where we were brought up, how we were educated and how our personalities find expression in our speech.

Although we know many of the features involved in distinguishing one voice from another, no automatic device has been built that comes close to rivalling the ability of human beings to recognize and identify voices. We can easily imagine applications for such a device. Law enforcement agencies, for example, are very interested in a technique for identifying persons from samples of their speech, just as they are now able to identify people by their fingerprints. Or, one might open his door or start his car by voice operated equipment, rather than relying on keys that get lost, forgotten or locked in the trunk.

SPEECH RECOGNITION

The problem of recognizing speech by machine is very formidable. One might say it was impossibly difficult, were it not for the fact that we have an "existence proof" of a "machine" that is able to transform acoustic waves into sequences of phonemes and words. This highly successful speech recognizer is, of course, man himself.

Primitive automatic recognizers have been built that can identify a small vocabulary of words—the 10 spoken digits, zero through nine, for example. These devices are fairly successful when the words are pronounced carefully, by a single speaker, with pauses between them. The problem of designing an automatic system to recognize "conversational speech"—that is, natural, continuous speech using a sizable vocabulary—is much more difficult. Its solution—if it ever is solved—will depend on a deeper understanding of the details and interdependencies of several levels of speech. Grammar, context, semantics and acoustics will all have to be considered in relation to the recognition problem.

If automatic speech recognizers ever become available at reasonable cost, they will have many useful applications. Telephones could be automatically "dialed" simply by speaking the desired number. Even better, it might be possible to place a call just by stating the name of the party to be called, plus some identifying information to avoid reaching a different person with the same name. Another possibility is a voice operated typewriter, capable of typing letters in final form faster than a good stenographer can take dictation.

SPEECH SYNTHESIS

We have already discussed the value of speech synthesis as a research tool. Through experiments with synthesized speech, we have learned a great deal about the speech wave features that are important for intelligibility. Much more remains to be learned through such experiments. In addition, if we ever perfect synthesizers that can be made to talk economically and conveniently, we will surely find important applications for them.

Earlier, we mentioned a "voice operated" typewriter. But how about a "typewriter operated voice"? A device of this sort would be of great value to people who cannot speak.

Voice outputs may very well be a useful feature of future digital computers. Combined with an input device that could recognize speech and accept spoken commands from a human operator, spoken outputs would permit direct verbal communication between man and machine.

Since we are having such pleasant dreams, we might just as well consider another remote possibility. Suppose we had an automatic language translating system, as well as a good speech synthesizer and a good speech recognizer. It would be possible, then, for two people who speak different languages to hold direct conversations—even over long distance telephone circuits—with each person speaking and hearing only his own language. Automatic devices would recognize speech in one language, translate it into a "second language" and prepare an input for a speech synthesizer that would "talk" in the second language.

BANDWIDTH COMPRESSION SYSTEMS

Next, let us consider the application of speech bandwidth compression systems. The *bandwidth* of a particular signal—a speech waveform, for example—is the important range of frequencies in the signal. The range of frequencies important for high quality speech is about 100 to 10,000 cps or, roughly, a bandwidth of 10,000 cps. Telephone quality speech contains frequencies between about 200 and 3400 cps, a bandwidth of only 3200 cps. Telephone speech is quite acceptable so far as intelligibility is concerned, but voice quality is noticeably altered.

A given electrical circuit can transmit only a certain limited range of frequencies; this frequency range is called the bandwidth of the circuit. If we try sending a signal that has frequency components outside the transmission band, some frequencies will be attenuated (weakened) and lost along the way. Consequently, the received signal will not be a good replica of the transmitted signal, and the resulting distortion, if severe enough, may be unacceptable.

Suppose we have a transmission link—an undersea cable, for the sake of argument—that can carry signals within a specified bandwidth. By using appropriate electronic techniques, we can transmit many "narrow band" signals over the cable, instead of only a single "broadband" signal. Specifically, suppose our cable has a bandwidth of 4,000,000 cps (four megacycles), which happens to be about the bandwidth required for transmitting one television program. We could, at a given time, use the cable to transmit a single TV program; or we could transmit 400 high quality speech signals (each having a 10,000 cps bandwidth); or we could transmit 1250 speech signals of telephone quality. Loosely speaking, our cable system's only limitation is that the total bandwidth of all signals transmitted simultaneously cannot exceed 4,000,000 cps.

If a way could be found to transmit speech using smaller bandwidths—say 250 instead of 3200 cps—it would be possible to transmit many more conversations over the same cable (about 13 times more in our example). The technique of "compressing" speech into smaller than normal bandwidths is called *speech bandwidth compression.*

Such techniques offer the possibility of increasing the communication capacity of existing transmission systems, without requiring the construction of additional expensive transmission links, such as undersea cables.

Speech bandwidth compression systems have been built on an

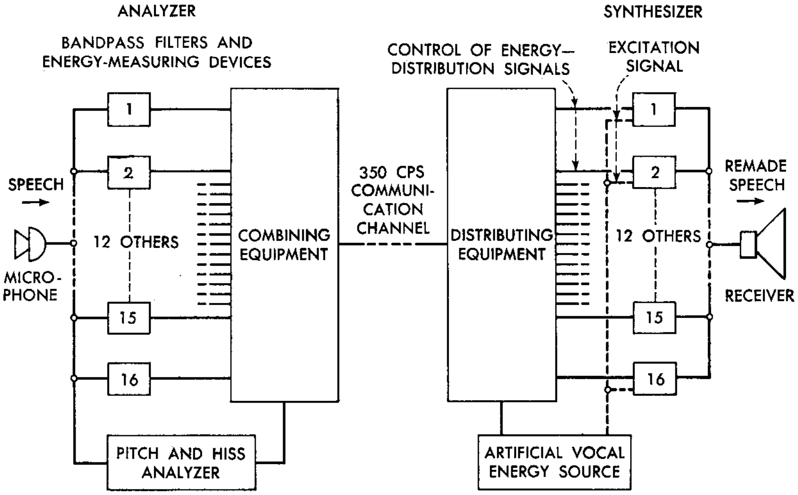


Fig. 9.1 The channel vocoder, a speech bandwidth compression system.

experimental basis. Their success rests on the fact that it is not necessary to reproduce the detailed shape of the original speech waveform in order to preserve speech intelligibility and naturalness. In fact, it is usually necessary to preserve only a rough replica of the speech spectrum. A *speech analyzer* at the transmitting end extracts signals from the original speech waveform; they are used at the receiving end to control a *speech synthesizer.* These control signals are sufficient to synthesize recognizable speech and, moreover, can be transmitted in a much narrower bandwidth then the original speech.

One device that does this is called a *channel vocoder*; it is shown schematically in Fig. 9.1. On the transmitting end is a speech *analyzer*; in the vocoder shown, this consists of 16 bandpass filters, plus a seventeenth channel to determine whether the speech is voiced or unvoiced and, if voiced, what its fundamental frequency is. With suitable electronic processing, the output of each filter represents the energy of frequencies in the speech signal that lie in the filter's "pass" band. These channel signals provide a rough measure of the original signal's spectrum.

The spectrum of the speech signal at any time is largely determined, as we already know, by the positions assumed by the articulators: the tongue, lips, etc. Consequently, the speech spectrum—or, equivalently, the outputs of the vocoder's filters—varies only as rapidly as the articulators change their positions.

These variations are quite slow compared to speech frequencies; in fact, they take place at frequencies below about 20 cps. We can, then, transmit the variations in the output of a single filter in a bandwidth of only 20 cps. By the reasoning used earlier, we can transmit all 16 filter outputs in a bandwidth of 16 x 20 cps, or 320 cps. The seventeenth channel may require another 30 cps or so, giving a total required transmission bandwidth of only 350 cps. On the receiving end of the vocoder link is a speech synthesizer; it produces an output of intelligible speech, using the 17 channel signals as inputs.

Another type of compression system is a *resonance vocoder.* Instead of transmitting spectral information in the form of filter channel signals, it transmits information directly about the formant frequencies (resonances) present in the speech signal. The resonance vocoder also offers considerable savings in bandwidth. Unfortunately, the quality of systems offering large bandwidth reductions are not yet good enough for use in commercial telephone systems.

In any compression system, the price paid for the more efficient use of transmission links is the greater complexity of the analyzing and synthesizing equipment such efficiency requires. Whether this complication is economically justified depends on the cost of increasing the communication capacity by other means—by laying more undersea cables, for example. Further research into the fundamental aspects of speech production and perception may lead to simple and effective methods offering even more bandwidth reduction than the vocoder systems.

THE PATH OF SCIENTIFIC RESEARCH

There is perhaps no better way to end this book than by quoting Georg von Békésy, 1961 Nobel Laureate in Medicine and Physiology, who has probably contributed more than any other individual to our knowledge and understanding of the hearing process. He has characterized the path of scientific research as follows:

"Of great importance in any field of research is the selection of problems to be investigated and a determination of the particular variables to be given attention. No doubt the verdict of history will be that the able scientists were those who picked out the significant problems and pursued them in the proper ways, and yet these scientists themselves would probably agree that in this phase of their work fortune played a highly important role. When a field is in its early stage of development, the selection of good problems is a more hazardous matter than later on, when some general principles have begun to be developed. Still later, when the broad framework of the science has been well established, a problem will often consist of a series of minor matters."*

He goes on to enumerate some of the forms scientific problems may take. These range from the "classical problem," which has been under attack, unsuccessfully, for a long time, to the "pseudo problem," which results from alternative definitions or methods of approach, and is not really a problem at all. Békésy warns us to beware of both the "premature problem," which is poorly formulated or not susceptible to attack, and the "unimportant problem," which is easy to formulate and easy to solve, but does not increase our fund of knowledge.

Two types of problems produce most of the worthwhile scientific results. First, the "strategic problem," which seeks data to support an intelligent choice between two or more basic principles or assumptions. Second, the "stimulating problem," which may open up new areas for exploration or lead to a re-examination of accepted principles. Of course, the strategic problems, when attacked and solved, lead to great steps forward. But one must not spend so much time and effort searching for strategic problems—they are very hard to come by—that he does nothing at all except search. It is really the "stimulating problem" that comprises most good research. A series of stimulating problems may, in the end, lead to a "strategic" result.

Let us hope the next few years will provide answers to many of the stimulating and strategic problems still unsolved. These answers will increase our understanding of the complex sequence of events involved in spoken communication.

* *Experiments in Hearing*, Georg von Békésy, McGraw-Hill Book Co., Inc., 1960, New York.

APPENDIX A Construction... by George R. Frost

This appendix is a guide to building your speech synthesizer. Since the synthesizer's circuits are rather complex, it will be of some advantage to read through the entire appendix before starting to build. You can then re-read it and follow the suggested procedures step by step.

Most of the electrical components and all of the "hardware" (nuts, bolts, etc.) for building the synthesizer are included in the experiment kit. You have to supply three flashlight batteries and a pair of headphones. Standard "D" size 1½ volt cells are most convenient, although any size cell of this voltage can be used. Magnetic headphones, of at least 2000 ohms impedance, will provide good results. In addition, if you have the following items at hand, the work of assembly will be much easier:

- Small cabinet screwdriver
- Longnose pliers
- Side-cutting pliers
- Pencil-type soldering iron
- Six feet of insulated hook-up wire (for batteries)
- Self-adhesive tape

Remove all the electrical components, hardware and partitions from the box tray to the box top, and place them in their original positions. This will help you identify the individual parts as you come to use them. Do not remove the heavy cardboard insert. The bottom of the tray, backed-up by this reinforcing insert, forms the surface of the chassis on which you will mount the parts and wire the circuits.

You will notice that the surface of the chassis is covered with tinted blocks, lines, titles and holes. These will help you mount and wire the individual parts. Compare the surface of the chassis with Fig. A.1 to see where the individual parts will be located on the finished device. Consult Fig. A.2 (page 122), the circuit diagram, to determine how the parts will be electrically interconnected.

The tinted blocks on the chassis divide the over-all Speech Synthesis circuit into four functional units; the name of each unit is printed directly under its block.

The printed designations inside the tinted blocks indicate the values of the resistors (in ohms)* and capacitors to be mounted at these points. The lines show the courses the bare tinned copper wire will take in interconnecting the parts. Where the lines are solid, the wire will run on the top surface of the chassis; where the lines are dashed, the wire will run on the under surface. Notice that some wires will thread over and under the chassis in connecting one part with another. If these paths are carefully followed during wiring, the wires will not contact one another and cause short-circuits.

Familiarize yourself with the parts themselves; Fig. A.1, and the labels on the box partitions, will help you identify them. The resistors—found in the vial—carry their ohm values in color-coded bands. Information for reading these bands in terms of resistance is on the back of the *Resonance Computer*. The capacitors, packed in the central compartment, are stamped with capacitance values in farads, microfarads or picofarads. The dark grey cup cores, also packed in a vial, are made of a brittle ferrite. These, together with the color-coded coils in the plastic bag, make up the inductors that will be used in the formant generators.

Construction is best started by mounting the hardware. The Fahnestock clips, see Fig. A.3(a), are mounted in the positions shown on Fig. A.1. The eight large clips are used at the capacitor terminals (C_1, C_2 and C_3) and at the Output terminals. The six smaller clips are used at the four battery terminals and at the Pitch resistor. Do not tighten the nuts on the clips or on any of the assemblies, since wires will later be connected to the

* Thousands of ohms are abbreviated, K; thus, a 3.9 K resistor represents 3900 ohms.

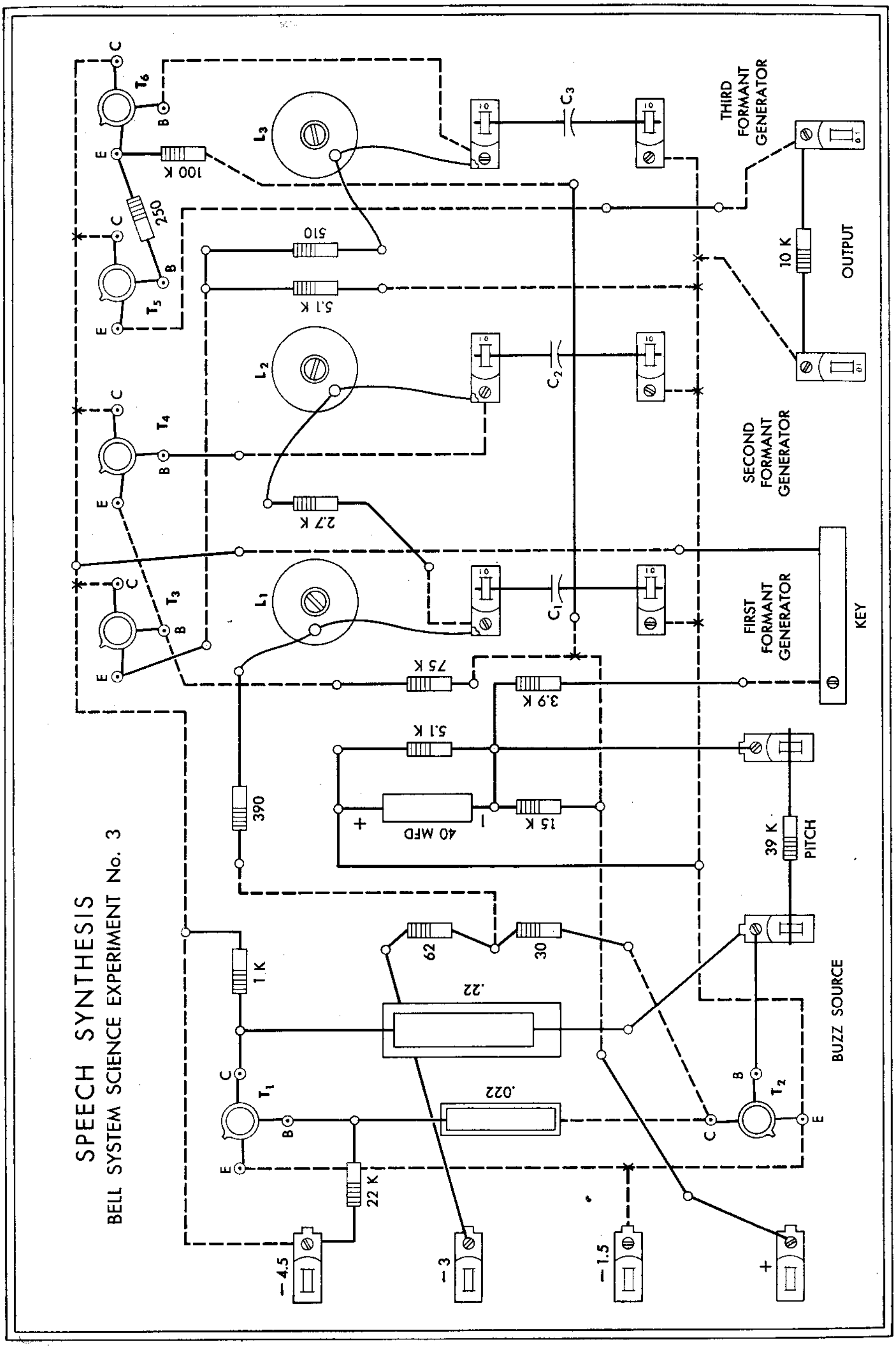


Fig. A.1 Location of individual parts on the completely assembled synthesizer.

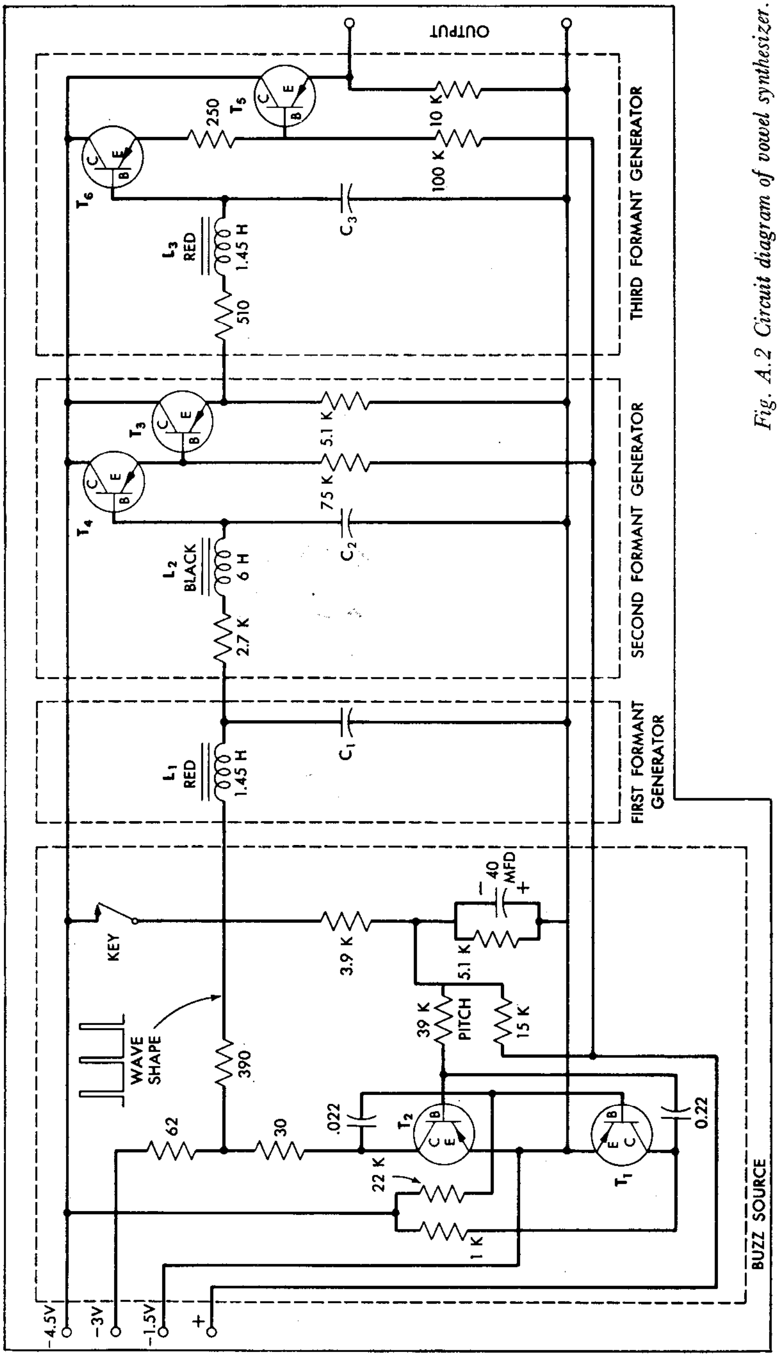


Fig. A.2 Circuit diagram of vowel synthesizer.

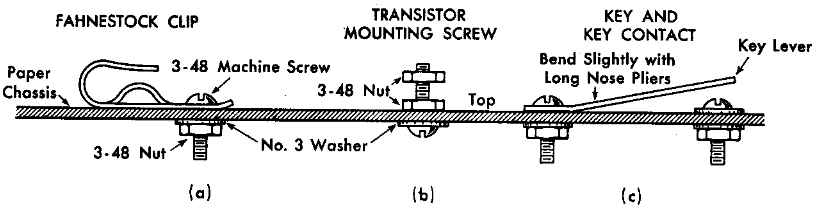


Fig. A.3 Mounting the hardware.

screws. Install the transistor mounting screws, Fig. A.3(b), and mount the key and key contact, Fig. A.3(c). This completes the hardware mounting.

You may find it helpful to follow a simple "error-avoiding" sequence in mounting and wiring component parts. The sequence begins with wiring the Buzz Source circuit. Install the 62 ohm and 30 ohm resistors as shown in Fig. A.1. Fasten these resistors in place by passing their leads through the chassis holes as shown on Fig. A.4. Now, fasten the 390 ohm resistor in place. Interconnect these resistors, the -3V Fahnestock clip and the *C* mounting screw of transistor T_2; use the bare tinned copper wire, following the indicated top (solid lines) and under face (dashed lines) paths. Solder the wire joints as shown on Fig. A.4 and bend the soldered joints (see Fig. A.4) against the chassis to lock the components firmly in place. Connect the "above" and "below" surface wires and leads to the screws, as shown in Fig. A.4. When each screw gets its full complement of leads and wires, tighten the nut firmly. For example, the nut at the -3V Fahnestock clip should now be tightened on the screw.

As each step in mounting and wiring is completed, the connections should be checked ะgainst the circuit schematic, Fig. A.2. This will provide a double check on your work.

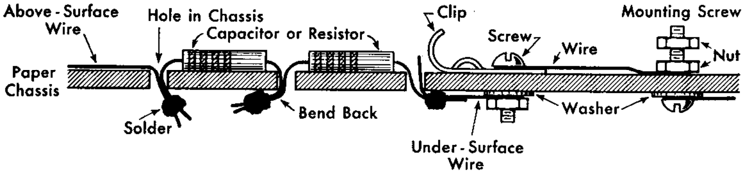
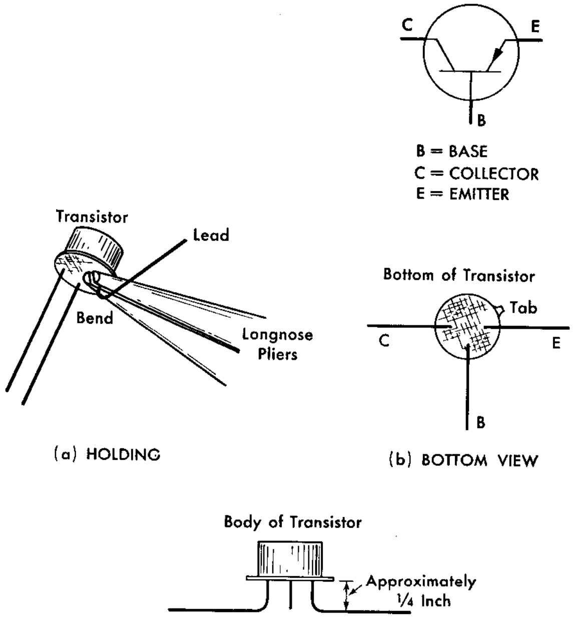


Fig. A.4 Connecting and soldering techniques.



(a) HOLDING

(b) BOTTOM VIEW

(c) SIDE VIEW

Fig. A.5 Preparing the transistors for mounting.

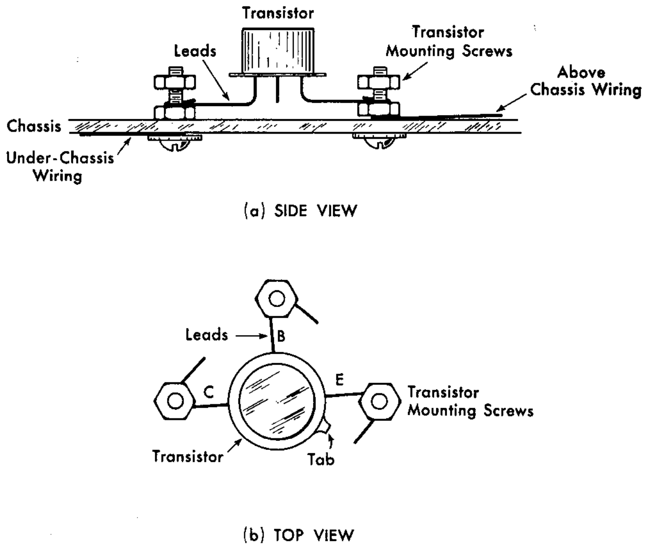
Next, mount and wire the group of components that makes up the time constant circuit of the Buzz Source (see Fig. A.1). These components are the 40 mfd. capacitor and the 15 K, 5.1 K, and 3.9 K ohm resistors. Notice especially that the 40 mfd. capacitor is polarized; that is, it has a definite positive (+) and a definite negative (−) lead. When you mount this component, be sure that its positive lead is placed in the hole adjacent the positive sign (+) on the chassis. Install all the wiring associated with these components. Proceed as before in soldering, consulting Fig. A.4 for techniques. Tighten the nuts as required, and check the completed step against the circuit schematic.

The Buzz Source components that control the output pulse-width and pulse repetition rate (pitch) can be installed in two groups. First, mount and wire the 1 K and 22 K ohm resistors and the 0.022 mfd. capacitor. Next, mount and wire the 0.22 mfd. capacitor and 39 K ohm resistor. The 39 K ohm pitch resistor is not permanently fastened into place; slip it into the Fahnestock clips as shown in Fig. A.1. This allows resistors to be changed for experiments with raised and lowered pitch.

Run the wire from the −4.5V Fahnestock clip to the 1 K ohm resistor, to the key lever, and to the *C* mounting screws of transistors T_3, T_4, T_5 and T_6. Also, run and fasten the wire interconnecting the $-1\frac{1}{2}$ V Fahnestock clip; the *E* mounting screws of transistors T_1 and T_2; the positive end of the 40 mfd. capacitor; the C_1, C_2, C_3 Fahnestock clips; and the output Fahnestock clips. This completes the wiring of the Buzz Source circuit. Check your work carefully against the markings on the chassis and the circuit diagram, Fig. A.2.

Wire the remaining mounting screws of transistors T_3, T_4, T_5 and T_6, and install the 250 ohm and 100 K ohm resistors located between T_5 and T_6. Complete this step by installing the 10 K ohm resistor (fastened permanently in place) at the output Fahnestock clips.

Install the 75 K ohm resistor between the Buzz Source and the First Formant Generator. Mount the 2.7 K ohm resistor between the First Formant Generator and the Second Formant Generator. Repeat the operation on the 5.1 K ohm and 510 ohm resistors to the right of the Second Formant Generator. This should complete all wiring and mounting, with the exception of the inductors and transistors. Check over all your work at this point.



(a) SIDE VIEW

(b) TOP VIEW

Fig. A.6 Mounting the transistors.

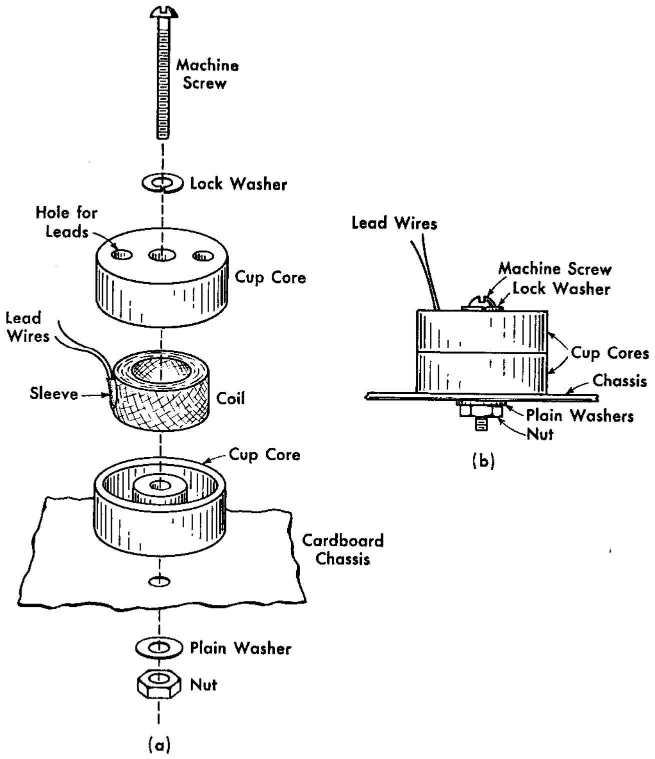


Fig. A.7 Assembling and mounting the inductors.

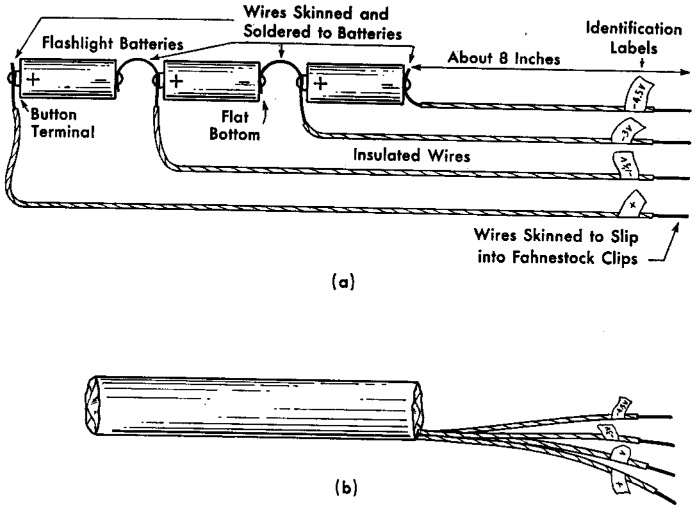


Fig. A.8 Preparing the power source.

Prepare the six transistors for mounting by bending the leads as shown in Fig. A.5. Each lead is grasped with longnose pliers—see Fig. A.5(a)—so that a bend can be made approximately one-quarter inch away from the transistor (see Fig. A.5(c).) The leads are bent out, as shown in Fig. A.5(b). Look at the lead identification shown in Fig. A.5(b); *E* is the emitter, located closest to the tab, *C* the collector and *B* the base.

Mount all six transistors by wrapping their leads around mounting screws as shown in Figs. A.6(a) and A.6(b). Cut off excess leads and fasten the nuts securely.

The formant generator inductors must be assembled before they are mounted. Begin by assembling inductor L_1 for use in the First Formant Generator. Remove a coil with *red* wire from the plastic bag and two cup cores from the glass vial. As shown in Fig. A.7(a), assemble the inductors by placing the coil in one cup core, making sure that the lead wires and sleeve pass freely through the cup core hole. Use the second cup core to enclose the coil. Press the cores together with your fingers; be certain that the lead wires are not pinched because of bad alignment.

Mount the inductor as shown in Fig. A.7(b). Tighten the nut until the lock washer is closed; this will provide proper tension. *Do not overtighten, since this may crack the core.* Solder the lead wires as shown on the chassis top and in Fig. A.1. It is unnecessary to remove enamel insulation from the lead wires—it will melt when the hot soldering iron is applied. Be sure to get the joint hot enough to melt the insulation.

Make up inductor L_2 for installation in the Second Formant Generator. Use the coil with *black* wire. Proceed as before in mounting and wiring. Do not overtighten.

Complete the construction and wiring of the Speech Synthesizer by assembling inductor L_3 (a *red* coil) and mounting it in the Third Formant Generator. Check all connections to see that they are sound and that no bare wires touch each other.

POWER

The three flashlight cells used for powering the synthesizer can be formed into an easily handled power pack.

Connect the three cells by soldering leads of skinned insulated wire to them, as shown in Fig. A.8(a). Try not to overheat the cells. Identify each lead with a label made of adhesive tape. Roll the cells in a full-sized sheet of writing paper; bind the roll with adhesive tape, as shown in Fig. A.8(b). (Do not wrap the cells, however, until testing is completed.)

TESTING

Before connecting the battery to the Speech Synthesizer, make a thorough check of your work. *Faulty wiring can lead to excessive currents, which may destroy the transistors.*

Connect the battery to the synthesizer, starting with the (+) lead and ending with the −4½ V lead. Connect the headphones to the output binding posts. Depress the key mounted on the chassis. You should hear a buzzing sound in the headphones. If you do, you are ready to conduct the experiments outlined in Chapter 7. If you hear no sound, follow the trouble shooting procedure outlined below.

Check the battery connections, making sure that the cells are properly poled and connected.

Remove the headphone leads from the output clips. Hold one headphone cord tip on the (+) Fahnestock clip and successively touch the −1½ V, −3 V and −4½ V clips with the other tip. A distinct click should be heard in each case. Now touch the tips to each adjacent pair of clips, listening for clicks. The absence of a distinct click in any of these tests indicates that a short-circuit exists in the battery wiring. If this is the case, disconnect the battery and clear up the trouble.

Connect one headphone tip to the lower C_1 clip in the First Formant Generator. While depressing the key, touch the other tip to the left-hand lead of the 390 ohm resistor. If *no* buzzing sound is heard, there is trouble in the Buzz Source circuit. If a buzzing sound *is* heard, the trouble is in the formant generators.

Should the trouble appear in the Buzz Source, check first that the electrolytic capacitor is properly connected; that is, that its (+) lead is toward the top and connected to the wiring located in the (+) hole. Carefully examine the circuit's wiring, comparing it with the courses printed on the chassis and with Figs. A.1 and A.2. Check the resistors and capacitors for correct locations and values.

If you still hear no sound after this last test, disconnect the battery and carefully remove transistors T_1 and T_2. Substitute transistors T_5 and T_6, removed from the output connections. Reconnect the battery. If you now hear the sound, transistor T_1 or T_2 (or both) has been damaged and must be replaced.

Should the trouble seem to be in the formant generators, connect one headphone tip to the lower C_1 clip in the First Formant Generator. While holding the key down, touch the other clip to the junction between the 5.1 K and 510 ohm resistors at the right of the Second Formant Generator. No buzzing sound indicates that the fault lies to the left of this point. If you hear a buzzing sound, the trouble is in the Third Formant Generator.

Again, check all circuits and components associated with the fault. Check the transistors by interchanging them in pairs. Prove the continuity of the inductor windings by momentarily placing the windings in series with the 1½ V battery and your headphones. A distinct click should be heard when you make and break this series circuit.

APPENDIX B Modifying Your Synthesizer

Listed below are three modifications you can make to your synthesizer—at very little additional cost—either to improve its performance or to make it easier to operate. The listing is not exhaustive; once you begin experimenting with the synthesizer, you may find other modifications you want to make.

(1) *Playing the synthesizer's ou'put through a loudspeaker-amplifier system*—The output of your synthesizer can be played through the amplifiers of hi-fidelity sets, public address systems, tape recorders, sound movie projectors, radios with phonograph inputs, and so forth. Make up a cable to connect the synthesizer's output terminals to the audio input socket of the amplifier you want to use. Connect the output terminal nearest the synthesizer's key to ground.

(2) *Adding a switch to make it more convenient to change vowel sounds*—In order to change the formants of your synthesizer, you now have to insert new capacitors in each of the three formant generators. This takes a good deal of time and does not allow you to switch from one vowel sound to the next fast enough to compare their qualities.

You can overcome this inconvenience by using a three-pole, multi-position switch, connected as shown in Fig. B.1. Different capacitors are connected into the circuit simply by selecting the right position on the switch.

(3) *Adding a pitch control knob*—The pitch of your synthesizer is controlled by its 39 K ohm resistor. Substituting a higher resistor will make the pitch lower, while a lower resistor will make the pitch higher. If you replace the 39 K ohm resistor with a 50 K ohm variable resistor in series with a 15 K ohm fixed resistor, you should be able to vary your synthesizer's pitch over a wide range.

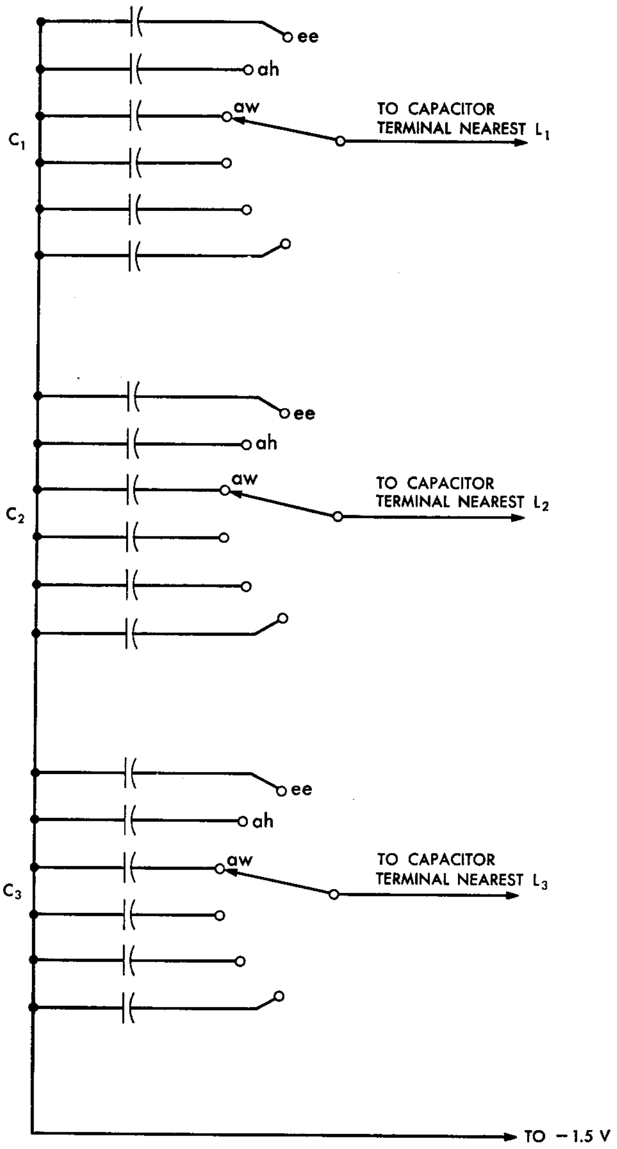

ee
ah
aw
TO CAPACITOR TERMINAL NEAREST L_1
C_1
ee
ah
aw
TO CAPACITOR TERMINAL NEAREST L_2
C_2
ee
ah
aw
TO CAPACITOR TERMINAL NEAREST L_3
C_3
TO −1.5 V

Fig. B.1 A diagram showing how to connect the multi-position switch for controlling the formants of your synthesizer. (The three rotating contacts move together to change simultaneously the capacitors connected to the formant generators.)

The Authors

Cecil H. Coker

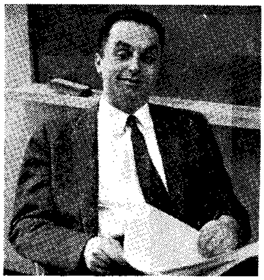
Cecil H. Coker joined the technical staff of Bell Telephone Laboratories in 1961. He is a member of the Acoustics and Speech Research Laboratory, where his chief interest is research in specialized, high efficiency methods of speech transmission.

Dr. Coker holds B.S. and M.S. degrees from Mississippi State University, and a Ph.D. degree from the University of Wisconsin.

He did research in autopilots and safety devices for light aircraft while working in Mississippi State University's Department of Aerophysics. At the University of Wisconsin, he was an assistant professor of electrical engineering.

Dr. Coker is a member of the Acoustical Society of America.

Peter B. Denes



Peter B. Denes joined the technical staff of Bell Telephone Laboratories in 1961. He is a member of the Acoustics and Speech Research Laboratory, where his chief interests are research in automatic speech recognition and speech synthesis.

Dr. Denes holds B.Sc. and M.Sc. degrees from Manchester University, England, and was awarded a Ph.D. degree from the University of London.

He was a demonstrator in the Electrical Engineering Department of Manchester University from 1941-43, and worked three years as a research engineer with Welwyn Electrical Laboratories.

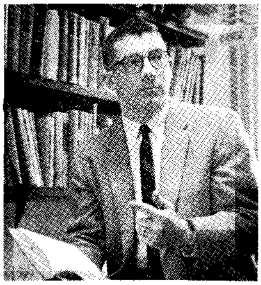
From 1946 to 1961, he was a lecturer at University College, London, where he also supervised the Phonetics Department's laboratory. He taught the physics, biology and the psychology of spoken communication, and his research work was concerned with various aspects of speech, hearing and deafness.

Simultaneously, Dr. Denes held appointments as physicist at the Royal National Throat, Nose and Ear Hospital and as honorary consultant at the Royal National Institute for the Deaf. He was a member of the Institute's Medical and Scientific Committee.

He is the author or co-author of many articles on automatic speech recognition, speech intonation, hearing tests and hearing aids.

Dr. Denes is a member of the Acoustical Society of America.

Elliot N. Pinson



Elliot N. Pinson joined the technical staff of Bell Telephone Laboratories in 1961. He is a member of the Computing and Information Research Center, where he is engaged in speech analysis and pattern recognition research.

Dr. Pinson was born in New York City and was graduated from Forest Hills High School. In 1956, he received the B.S. degree, summa cum laude, from Princeton University. He attended the Massachusetts Institute of Technology on a Whiton Fellowship, and was awarded the S.M. degree in 1957.

He continued his studies at the California Institute of Technology, where he was a Ramo-Woolridge and Space Technology Laboratories Fellow. At Cal-Tech, he did research in adaptive control systems, and received his Ph.D. degree in 1961.

He was an instructor in electrical engineering at Cal-Tech from 1960-61. He also has worked on the design and analysis of missile guidance and control systems.

Dr. Pinson is a member of Phi Beta Kappa and Sigma Xi, as well as the Acoustical Society of America and the Institute of Electrical and Electronics Engineers.