S. E. Estes
H. R. Kerby
H. D. Maxey
R. M. Walker

Speech Synthesis from Stored Data

Abstract: The synthesis of speech by joining together segments derived from natural speech has not proven to be satisfactory with segments smaller than words, especially because of discontinuities in pitch and formant frequencies at the junctions. It appears that segmentation of the control signals for an analog synthesizer may avoid these difficulties. This paper describes an experimental system to investigate this method. A library of synthesizer control signals corresponding to subword segments of speech is now being developed. The equipment used to generate the library of control signals, as well as that used to synthesize connected speech from the library, is described.

The synthesizer is a transistorized terminal analog of the cascade type. The synthesizer control signals are originally derived from functions drawn on a transparent plastic belt with opaque tape and scanned by a CRT and photomultiplier. The control signal functions are varied until the speech segment being studied is satisfactory. The resulting control signals corresponding to the speech segment are then automatically digitized and recorded on punched cards for addition to the library. Connected speech may be generated by computer assembly of the synthesizer control signals corresponding to a sequence of speech segments. In the assembly of connected speech from the library segments, pitch and timing may be specified independently of the sequence of segments if desired.

1. Introduction

The synthesis of speech has potential application in two different areas: first, that in which the synthesizer is controlled by a set of parameters derived from natural speech, such as the frequency band[1] or formant tracking[2,3] vocoders; and second, that in which the basic input to the synthesizer is a set of discrete symbols such as a sequence of phonemes.[4] In the former application, the preservation of the essential qualities that identify the particular speaker are important, whereas in the latter any pleasant-sounding, intelligible voice is acceptable. It is the second application that occupies our attention.

In attempts to find suitable solutions to the problem, investigators have expended considerable effort in experiments on segmentation of natural speech,[5,6] primarily because the concept of a limited library of speech segments (from which an unlimited vocabulary can be constructed) is very attractive. The success achieved in the construction of sentences from subword segments of natural speech has been limited, and this approach has largely been abandoned. As the degree of understanding of speech has progressed, some of the basic problems encountered in joining segments of natural speech have been understood. Problems exist because of excitation discontinuities and differences in vocal tract configuration at the juncture point of the two segments. If the vocal tract configuration were continuous as a function of time across a junction and if an acoustic continuum in the vocal tract excitation could be assured when joining segments, this approach would be much more successful. This is not possible in joining segments of natural speech. However, in synthesis the effective configuration of the vocal tract and its excitation are specified by a set of slowly varying control functions. It is much easier to join these synthesizer control signals properly than to join the acoustic waveforms of natural speech.

Segments of control signals can be made which cor-

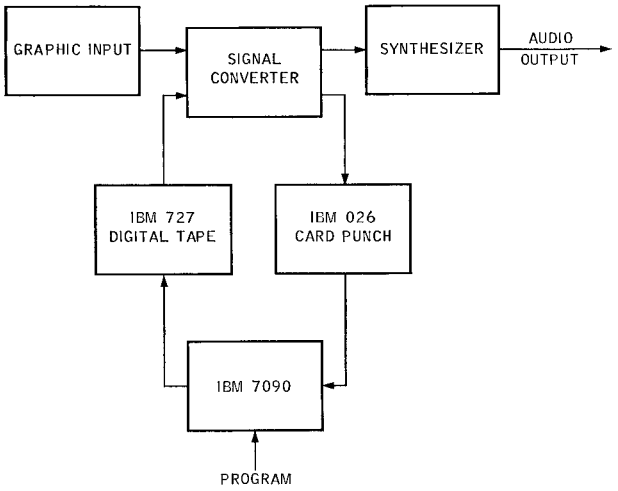


Figure 1 **Speech synthesis system.**

respond to segments of synthetic speech, and a library of speech segments can then be stored in the form of a library of synthesizer control-function segments. The synthesizer may then be controlled by composite functions assembled from the sequence of control-function segments corresponding to the desired utterance. In order to investigate this approach to speech synthesis, a terminal analog synthesizer and associated control system have been constructed. Our goal is to generate a library of reproducible control-function segments from which we can construct connected speech.

The derivation of a set of control signals for an acceptable utterance involves a combination of rules for speech synthesis, experience, and trial and error. Thus it is desirable for the operator to be able to work directly with a synthesizer and, once he has generated a satisfactory utterance, to be able to transfer the control signals to storage for recall and reuse. Shown in Fig. 1 is a functional block diagram of a system which serves this purpose. The operator may work with the graphic input in which the synthesizer control signals are drawn on a transparent belt and used to excite the synthesizer.* When the operator is satisfied with the speech segment, he may have the resulting control signals digitized and then punched on cards. The speech segment is then part of the library and the IBM 7090 may be used to assemble a digital tape from the library of segments. The tape may then be "played" into the signal converter by an IBM 727 tape unit to operate the synthesizer in real time. Although the synthesis of artificial speech by an electronic analog is not novel, we believe that this system combines the advantages of the speed and experimental flexibility of a real-time analog device with the off-line capability of a digital computer for processing the control data further.

In the following sections a description will be given of the synthesizer and its associated control system, together with results obtained to date.

2. Equipment

• *A. Synthesizer*

1. *Vocal tract analogs*

Human speech sounds are produced by the conversion of air flow from the lungs into sound sources that excite the resonant cavities of the vocal tract. For voiced sounds, the air flow is converted into a pulsating form by the vocal cords; for unvoiced sounds, the excitation arises from air flow through constrictions.

The vocal tract may be thought of as essentially a resonant pipe of nearly constant length but of variable cross-section, extending from the vocal cords to the lips. Articulatory gestures involving the pharynx, velum, tongue, teeth, and lips vary the cross-sections of this pipe so as to alter its transmission characteristics. During speech, we control the sound sources and alter the configuration of the vocal tract to produce a series of acoustic events that convey meaning according to the code of the language.

For synthesis of the acoustic waveforms of speech by electronic means, we require a pulse generator (of controllable frequency) to simulate the vocal cord source, a noise generator to simulate the unvoiced source sounds, and controllable filters that can simulate the transmission characteristics of the vocal tract.

There are three general approaches to attaining the desired filter characteristics:

(1) A multisection transmission line analog of the physical system,[7] where the equivalent cross-sectional areas of the line sections are varied electrically.

(2) A parallel bank of bandpass filters, with the individual gains variable.

(3) A terminal analog that approximates the filter characteristic by either a parallel[8] or a cascade[9] connection of resonant circuits having controllable resonant frequencies. These frequencies can be varied electrically to correspond to the primary resonances of the vocal tract. Since the cascade connection gives a better approximation to the desired filter characteristic and also requires fewer control variables, it is now the generally preferred arrangement. It is usually considered sufficient perceptually to provide for three such variable resonances, corresponding to the lowest three formants[10] in the speech

* Various forms of graphic synthesizer control have been reported in the literature, e.g., Ref. 8.

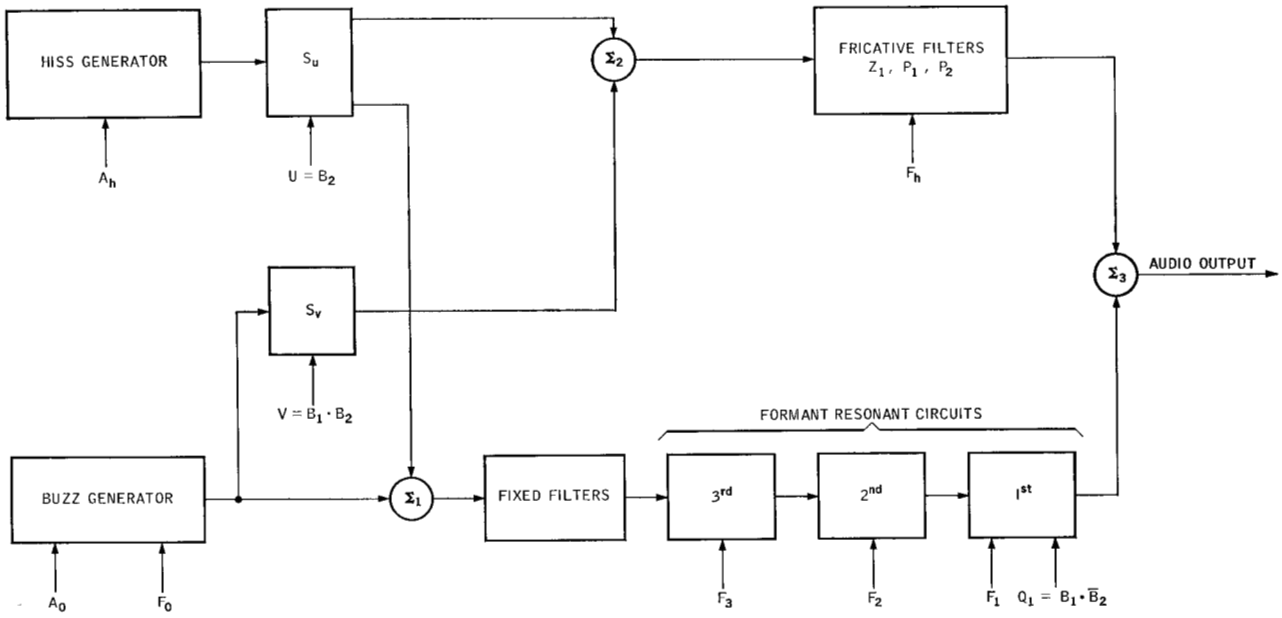


Figure 2 **Terminal analog synthesizer.**

spectrum. An additional variable filter, in a separate branch, is provided to shape the noise spectrum for unvoiced sounds.

Of these three methods, (1) and (2) require about 18 control parameters while (3) requires about 8. Furthermore, the control information needed for (3) is more easily obtained by analysis of sound spectrograms and is more easily correlated with the phonetic events. These considerations led to our choice of the cascade terminal analog type for our system.

2. *Terminal analog synthesizer*

a. *Component parts*

A functional block diagram of the synthesizer is shown in Fig. 2. It is a cascade type, transistorized terminal analog of the vocal tract. Such a system has two filter paths: one consisting of a cascade of formant resonant circuits and another consisting of the fricative filter. Either path may be excited by "buzz" energy for voiced sounds or by "hiss" energy for the various unvoiced sounds.

This synthesizer requires 2 binary and 7 analog control signals. Logical combinations of the 2 binary control signals, B_1 and B_2, generate binary functions U, V, and Q_1 as shown in Fig. 2. The binary function U directs the hiss signal to the fricative filter while $\bar{U}$ directs it to the formant resonant circuits, V directs the buzz signal to the fricative filter, and Q_1 increases the bandwidth of the first formant to simulate nasality. The analog control signals are: A_0 and F_0, the amplitude and frequency of the buzz; A_h, the amplitude of the hiss; F_1, F_2, F_3, the resonant frequencies of the three variable formant resonant circuits; and F_h, the frequency of the fricative resonance, P_1. The other fricative resonance, P_2, and the zero, Z, are made to follow P_1 with a prescribed functional relationship.

b. *Performance requirements*

For convenience in specifying the synthesizer control signals, it is highly desirable that the over-all relationship between input control signals and frequency variables be linear. Since a large amount of synthesis is to be done, it was profitable to expend the extra time to assure a sufficient degree of linearity in this over-all relationship, including both the control-signal generator and the formant filters. Since the frequency of the fricative filter is not as critical, its design is not as complex as that of the formant filters.

In addition to the linearity requirement, the stability, particularly that of the formant filters, must be considered. Flanagan[11] states that formant frequency specification need never exceed an accuracy of 3 to 5 percent to provide vowel identification. The over-all stability requirements are thus very lenient; however, this includes formant filters, control signal generator, and operator error. Possibly the most critical is operator error, so the variation in the separate components should only be a small part of 5%, preferably less than 1%. This has been provided on a direct control basis.

The perception of synthesized speech can be seriously affected by the presence of undesired signals. For example,

with several af resonant circuits in the synthesizer, the over-all noise level of the amplifiers or other components must be very low, since amplification of the noise by the resonance would generate an audible signal that would seriously impair the quality of the synthesized speech. Also, since the control signals for the synthesizer have frequency components in the audible range, it is necessary to provide a high degree of rejection for these signals.

The components of a synthesizer (formant resonant circuits, fricative filter, buzz generator, and hiss generator) that meet these requirements are described in the following sections.

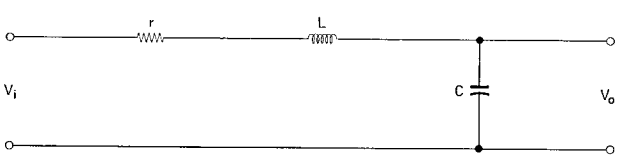
3. *Formant resonant circuits*

A formant resonance is characterized in frequency by a constant bandwidth and by unity gain at frequencies well below the resonant frequency.[12,13] These are the characteristics of a voltage-excited series resonant circuit, shown in Fig. 3, with resonant frequency varied by changing C. Such a system has the locus of poles in the S-plane as shown in Fig. 4; only the region where complex conjugate poles exist is of interest. The locus shown in Fig. 4 may be generated in several ways. One method which has been advantageous is to start with a pole pair and feedback amplifier configuration with a variable gain to generate the desired position of the locus. The basic pole pair may be a set of poles on the negative real axis (i.e., tandem RC circuits) or they may be a conjugate pair (i.e., a series LC circuit). The series LC circuit has been used, for the total gain required for a given resonant frequency is considerably less than that required if a tandem RC circuit is used.

The block diagram representation together with the over-all transfer function of this formant resonant circuit is shown in Fig. 5. From Fig. 5 it can be noticed that the frequency varies as $\sqrt{1+A}$ but since A is large (we have restricted A to lie between 10 and 400) the resonant frequency varies nearly as $\sqrt{A}$. The variable gain A is derived from a tandem arrangement of a fixed gain with two variable gains. Each variable gain changes linearly with a single control signal. Thus A varies as the square

Figure 3 **Desired form of formant resonant circuits.**

$$\frac{V_0(S)}{V_i(S)} = \frac{\frac{1}{LC}}{S^2 + \frac{r}{L}S + \frac{1}{LC}}$$



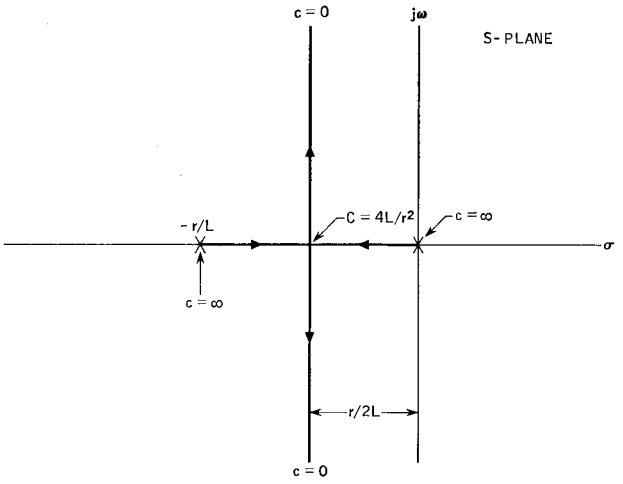
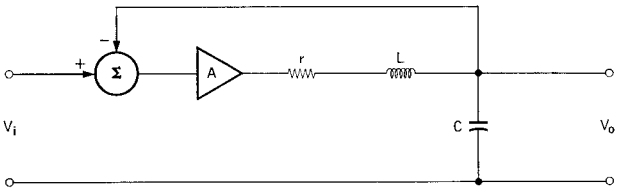


Figure 4 **Root locus for circuit of Figure 3 as C is varied.**

Figure 5 **Formant resonant circuit synthesis.**

$$\frac{V_0(S)}{V_i(S)} = \frac{A\frac{1}{LC}}{S^2 + \frac{r}{L}S + (1+A)\frac{1}{LC}}$$

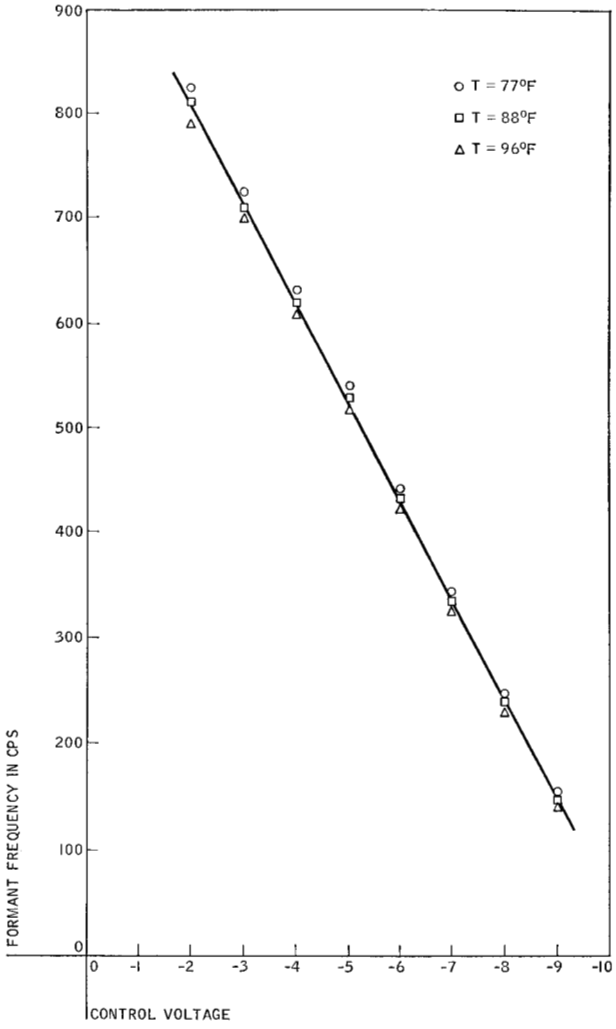


of a control signal, causing the resonant frequency to vary linearly with the control signal.

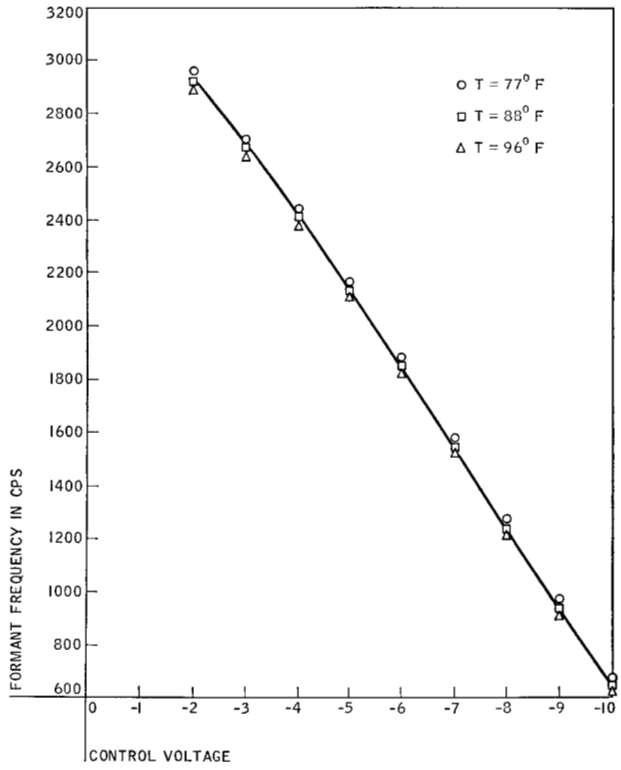
The variable gain is realized by multiplying the audio signal by a high-frequency square wave which has a variable duty ratio. The signal which results after low-pass filtering is the audio signal multiplied by the duty ratio. A pulse-ratio modulator provides a very convenient way of generating the variable duty-ratio square wave.[14] The pulse-ratio modulator, which is a combination of pulse-frequency and pulse-width modulation, has excellent stability and linearity. The performance of the formant resonant circuits constructed in this way is summarized by the plots of formant resonant frequency vs control voltage, with temperature as a parameter, shown in Figs. 6a, 6b, and 6c.

4. *Fricative filter*

The fricative filter generates a two-pole one-zero transfer function approximation to the fricative spectrum as suggested by the work of Heinz and Stevens.[15] The lowest pole is tuned by an analog control signal (F_h). The zero and the second pole are made to track in such a way as to be at the position shown in Table 1 when the lower



(*a*)



(*b*)

(*c*)

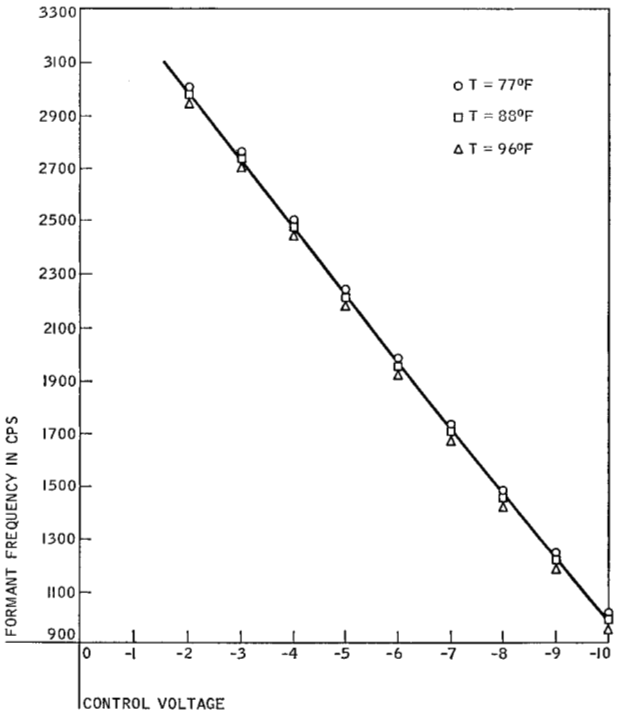


Figure 6 **Formant transfer characteristics.** *F_1, F_2, and F_3 are first, second, and third formants, respectively, in charts (a), (b), (c).*

Table 1 **Fricative filter characteristics.**

Phoneme	*Z*		*P_1*		*P_2*	
	Freq.	*BW/2*	*Freq.*	*BW/2*	*Freq.*	*BW/2*
ʃ and ʒ	1500	600	2500	250	5500	500
s and *z*	2500	500	4000	250	7000	500
f and *v*	7000	900	8000	500	—	—

pole is set for a specific fricative utterance. The resonances are generated from *LC* circuits using variable inductors.

5. Buzz generator

The buzz generator consists of a unijunction transistor relaxation oscillator which is used to generate a sequence of narrow rectangular pulses at a repetition rate determined by the pitch control signal. The repetition rate

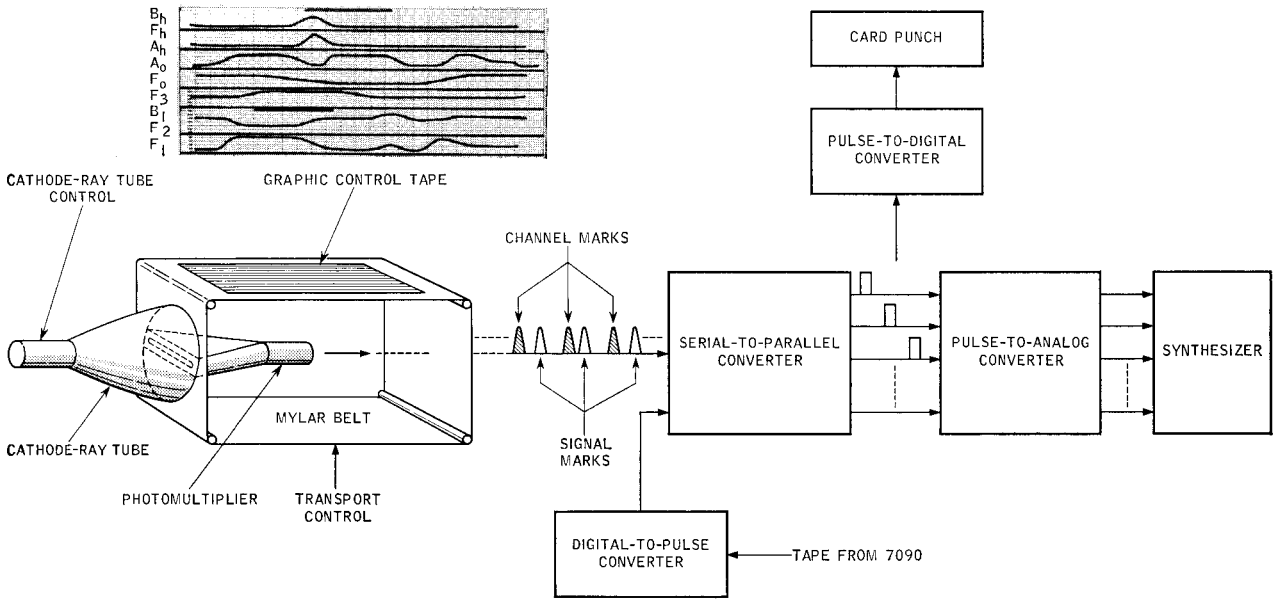


Figure 7 **Synthesizer control signal generation.**

varies from 80 to 160 pulses per second and the pulses are 100 μs wide. Thus, assuming the pulses exactly square, the first zero in the frequency spectrum of the buzz signal is 10 kc/sec and the spectrum amplitude is down 4 dB from its fundamental at about 5 kc/sec.

The actual voiced excitation from the vocal cords does not have a flat spectrum but rather one which decreases in intensity at higher frequencies. However, the cascade terminal analog of the vocal mechanism requires some high-frequency boost to compensate for the lack of higher formants. This is accomplished in part by a flat, rather than decreasing, amplitude spectrum in the buzz excitation. The amplitude of the buzz signal is controlled by clipping the top of the rectangular pulses at a level set by the control signal.

6. Hiss generator

The desired hiss signal is white noise, band limited between 100 cps and 12 kc/sec. Its amplitude control is an exponential variation in response to a linear control signal. The basic noise source is a back-biased point-contact diode. Rather than construct a linear amplitude control, it was found to be desirable to control the noise amplitude by symmetrical clipping. This is permissible if the noise input to the clipper is a square wave with random zero crossings, for then the clipped output will also be a variable-amplitude square wave with random zero crossings, and the clipper introduces no distortion. However, it is desirable to have the amplitude of the noise vary with the control voltage in such a manner as to provide sound intensity proportional to control voltage. This is accomplished by using the cubic v-i relation of a silicon carbide varistor to generate a nonlinear clipping level from the control signal. This amplitude control works over about a 60 dB range; however, only about 40 dB obeys the desired relation to a reasonable approximation.

A hiss signal generated in this way may be considered as the integral of a noise process consisting of randomly spaced impulses with the restriction that successive impulses alternate in sign. The actual power-density spectrum of this process depends on the correlation between the successive impulses, which in turn were determined by the form of noise generated by the diode together with filtering present in the amplifier. For a given temperature the noise generation process is stationary, and while it is a difficult problem to characterize the noise exactly, it is a simple problem to use filtering where necessary following the clipping, in order to flatten the spectrum over the desired bandwidth.

• *B. Synthesizer control system*

In previous sections of this paper it has been noted that the synthesizer control system serves three functions. First, a graphic input allows the operator to control the synthesizer with a set of hand-drawn control signals that can be readily changed. Second, these control signals can be automatically punched into IBM cards. Third, this punched data can be assembled, using the IBM 7090, and a digital tape output used to operate the synthesizer. Figure 7 shows a functional block diagram of the system

which is described in more detail in the following sections.

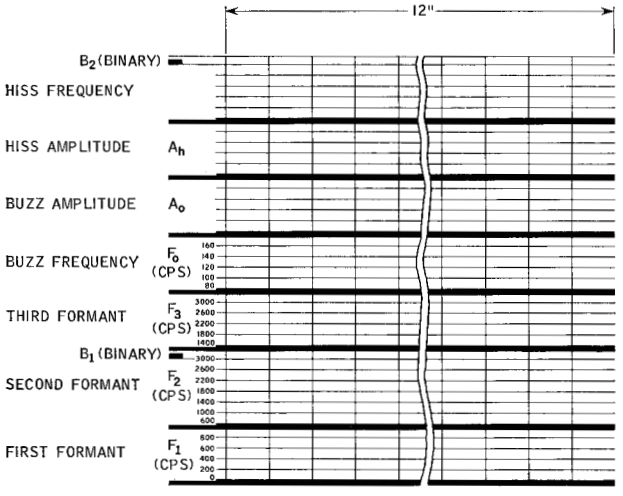
1. Graphic input control

Control signals are "drawn," using opaque tape, in their respective channels on a transparent graphic control tape on which are printed the channel and calibration marks. Figure 8 shows the format of the control tape with the identification of the separate control function channels and the calibration marks.

The tape is fastened to a transparent Mylar belt-loop which serves to transport it past the CRT-photomultiplier flying-spot scanner (Fig. 7). With the CRT sweeping perpendicular to the direction of tape travel, a pulse is generated each time the CRT spot is blanked by a line on the graphic control tape. The time separation between a channel mark and a signal mark thus represents the amplitude of the control signal for that channel. This time is converted to a pulse width. The pulses generated by each sweep are separated into their respective channels, where they are converted to analog control signals by integrate-and-hold circuits. Special logic allows the two binary signals B_1 and B_2 to be near the top of F_2 and F_h channels, respectively. The channel voltages supplied to the synthesizer have a long-term stability of about $\pm 2\%$ of their dynamic range. This drift could cause the formants of the synthesizer to deviate from the frequency values specified on the control tape by ± 15 c/s in F_1, ± 40 c/s in F_2, and ± 30 cps in F_3.

Lines "drawn" within a channel need not be continuous and can be overlapped. The first line encountered by the CRT-photomultiplier is taken to represent the channel amplitude at that sample time. Extraneous audio output from the synthesizer can be controlled by a logic gate which requires that valid data be read from all channels of the control tape before the synthesizer is connected to its audio amplifier and speaker system.

Figure 8 **Control tape format.**



The transport mechanism has several interesting features. Often only a small part of the 57-inch Mylar belt contains control tapes. Therefore the tape transport mechanism has an automatic-drive speed-change system to quickly cover those spaces where there is no control tape and then to return to the preset transport speed when a control tape is passing under the CRT READ station. An operator may select any transport speed from near 0 to about 10 inches per second, with 5 in/sec corresponding to a sound spectrogram time scale. Thus, the rate of talking may be varied over a wide range. When the talking speed is changed, there is no alteration in frequency components as when a phonograph record is played at the wrong speed; instead, only the talking speed is changed. Synthesized speech thus retains its intelligibility over a wide range of speaking rates.

The CRT sweep repetition rate, and thus the number of samples of each control signal taken per second, is variable up to 250 samples per second with a nominal rate of 100 per second generally used. This parameter, of course, will receive a great deal of attention when the actual method of encoding the library for permanent storage is considered, but at present it is set at a high value to prevent any influence on speech quality.

2. Control signal storage

An utterance which has been found acceptable by generation from the graphic input may be stored in the form of synthesizer control signals that are sampled, digitized, and punched on IBM cards. In order to make such a conversion, there must be compatibility between the possible punching rate and the sampling rate. Since the card punch is slow, this has been accomplished by putting an incremental drive on the control tape which advances the tape in 0.05-inch increments. This corresponds to a sample rate of 100 per second at a tape speed of 5 in/sec. The control tape is then repeatedly scanned by the CRT and a two-decimal digit number is punched for the control-signal amplitude in each channel; thus each sample set requires 14 columns of the punched card. The tape drive may then be automatically incremented and the procedure repeated until the control signals between two specified points along the control tape have been digitized and stored. Binary channel signals appear as special punches in conjunction with the F_2 and F_h channels. It takes about five seconds to punch a card which holds four sample sets and therefore represents 40 ms of speech. Manual punches provide card identification and special stop signals that tell the computer to ignore unused portions of the final card in a speech segment.

3. Computer control

A program has been written which can generate an output tape to control the synthesizer from the speech segment

library. A program statement must be provided which specifies the utterance in terms of the sequence of speech segments and the operations which are to be performed on the segments in assembling them. Operations which can be performed presently are:

(1) Segments having a steady state, or vowel, ending may be extended by repeating the last sample the desired number of times.
(2) The end of a word may be indicated and the interval of silence before the next word may be specified.
(3) Pitch data may be taken from library cards, may be specified to be monotone, or may be separately specified independently of the library cards.

The computer provides a tape which has a sequence of numbers that are grouped to correspond to the sample time of a 100 cps CRT sweep rate of the graphic control system and are spaced within the grouping to correspond to the separate channels. When the tape is played back on an IBM 727 tape unit in the laboratory, it then provides the proper time base for real-time control of the synthesizer.

The digital-to-pulse converter shown in Fig. 7 converts the binary number sequence from the tape into a pulse sequence with the same format as that derived from the CRT-photomultiplier scan of the graphic input. The same circuitry is thus used to generate the synthesizer control signals.

While assembly of the control signals by the computer is very fast, the present program requires that they be written on the output tape in real time in order to operate the synthesizer. At present, the computer time required to generate an utterance is equal to the length of the utterance.

• *C. Hardware*

A photograph of the equipment described is shown in Fig. 9. The IBM tape unit and card punch are on the right. The control system occupies the console with the graphic control tape on the work surface. The right control panel is for data storage and tape input, and the left control panel is for the graphic input system. The synthesizer occupies the rack to the left of the console and the rack on the extreme left contains a tape recorder and calibration instrumentation. Provision is made for synthesizer output to remote points in the recording room as inputs to the sound spectrograph and other analysis or recording equipment.

3. Results

• *A. Acoustic cues*

At the point of actually making a control tape for the function generator, the important question becomes:

Figure 9 **Synthesis hardware.**

With this particular synthesizer, how can we best approximate the human speech spectrum to ensure the preservation of the spectrum details that allow the human ear to identify the different speech sounds (phonemes) of English? These details are called *acoustic cues.*[16] Our sources of data on acoustic cues are the literature covering many years of work in the analysis of human speech, published results of work on synthetic speech, and our own experiments using spectrum analyzers and the synthesizer. The following is a summary of the synthesis techniques that we have used thus far to produce intelligible speech.

1. Vowels (i, ɪ, e, ɛ, æ, ɑ, ɔ, o, ʊ, u, ʌ, ɚ).

For our cascade-type synthesizer, the vowels are well characterized by specifying the three formant frequencies F_1, F_2, and F_3. The first two formant frequencies are different enough, however, to allow the vowels to be presented on a two-dimensional plot as in Fig. 10. For a human speaker, the vowels can be expected to fall only in the general areas indicated and will be considerably different from word to word. For our initial synthesis by assembling diphones, to be described later, the vowels will have standard, constant values for F_1, F_2, and F_3. If this simplification is found to reduce the intelligibility, variants of the vowels can be added to the segment library.

2. Diphthongs (aɪ, oɪ, aʊ).

A diphthong is characterized by smooth, slow changes in formant frequencies between two apparent "vowel positions." Although diphthongs are represented in transcription by digraphs (successions of two vowel symbols each related to a perceived "end-point") the actual formant frequencies of the end-points involved, as measured from human speech, are somewhat different from the frequencies of even the same person's vowels when independent or in isolation. In terms of the problem of synthesis, we have yet to answer the question of whether we can simulate diphthongs with transitions between

standard vowel frequencies or whether they will have to be considered as special cases. The answer will come as we build up our segment library and are able to generate words in their natural context for evaluation.

3. Consonants

The consonants have been grouped according to their similarity in terms of their synthesis. A function-generator control tape and the resulting spectrum analysis on the spectrograph are shown in Fig. 11 for a sample from each of the seven classes. Each synthesized consonant is paired with the vowel [ɑ]. The following description refers to the control tape.

a. Liquids and semivowels [rɑ], [l, j, w].

These consonants are produced by generating short, constant formants followed by slow transitions (changes in formant frequencies) to the vowel frequencies.

b. Aspirated vowels [hɑ].

The formant filters are tuned to the vowel positions and are energized first with the hiss generator, then with the buzz generator. The Q of F_1 is reduced during the aspirated portion.

c. Voiceless plosives [pɑ], [t, k].

During the transitions, the formant filters are energized with the hiss generator. The onset of the hiss amplitude is fast whereas for the [h] the onset is slow. As the formants reach the vowel frequencies, the hiss amplitude is decreased and the buzz amplitude is increased. The Q of F_1 is reduced during the aspiration as in [h].

d. Voiced plosives [bɑ], [d, g].

The transitions are the same as for the voiceless plosives. The buzz generator is turned on slightly before the transitions start and the Q of F_1 is not reduced.

e. Nasals [mɑ], [n, ŋ].

The nasals are generated with constant formants and the Q of F_1 reduced during the simulated closure of the lips, followed by rapid transitions to the vowel frequencies. During the transitions, the Q of F_1 is increased to its normal value over a period of approximately 100 ms. A critical factor is the timing and rate of change of the Q of F_1. If the Q of F_1 is increased too rapidly, the sound [mbɑ] will be generated.

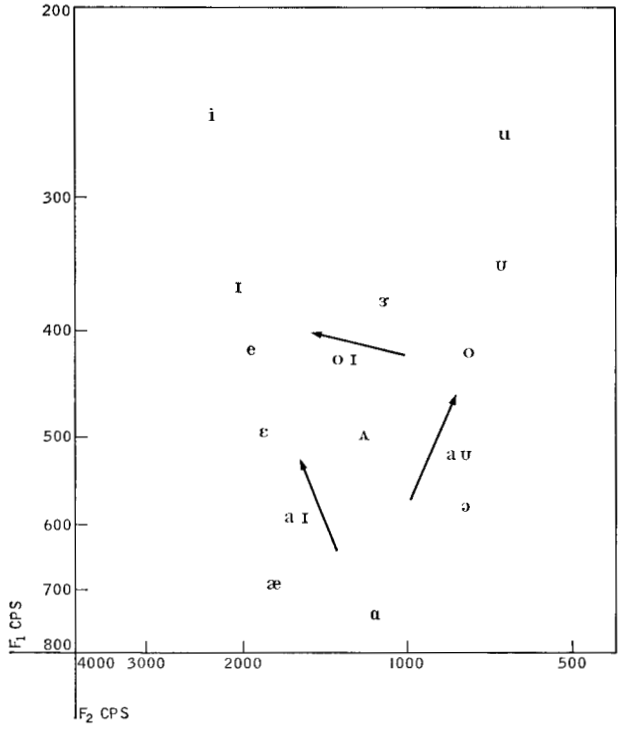
f. Voiceless fricatives [sɑ], [f, θ, ʃ, tʃ].

The binary channel B_2 is operated, connecting the hiss generator to the two-pole, one-zero fricative filter. The hiss frequency channel tunes the fricative filter to the appropriate pole-zero location for the particular fricative. The hiss amplitude channel is operated first to give the correct noise spectrum, then the buzz amplitude control together with the changing formant filters generate the voiced transitions appropriate for the fricative.

g. Voiced fricatives [zɑ], [v, ð, ʒ, dʒ].

The control tape for the voiced fricatives is very similar to the control tape for the voiceless fricatives. The difference is that the binary channel B_1 is operated, thereby adding buzz energy to the fricative filter to supply part of the voicing. The buzz amplitude channel is also operated to provide low amplitude formants during the initial part of the fricative.

Figure 10 **English vowels and diphthongs.**



B. Summary of synthesis

1. Diphones

For our first efforts in speech synthesis we have considered that, as a good approximation, we can synthesize any word in English by assembling shorter speech segments if the splicing takes place during either a steady-state condition (control signals are constant) or a silence. As an example, the words:

comet	cosset
['kɑmɪt]	['kɑsɪt]
[kɑ + ɑm + mɪ + ɪt]	[kɑ + ɑs + sɪ + ɪt]

raccoon
['ræ'kun]
[ræ + æk + ku + un]

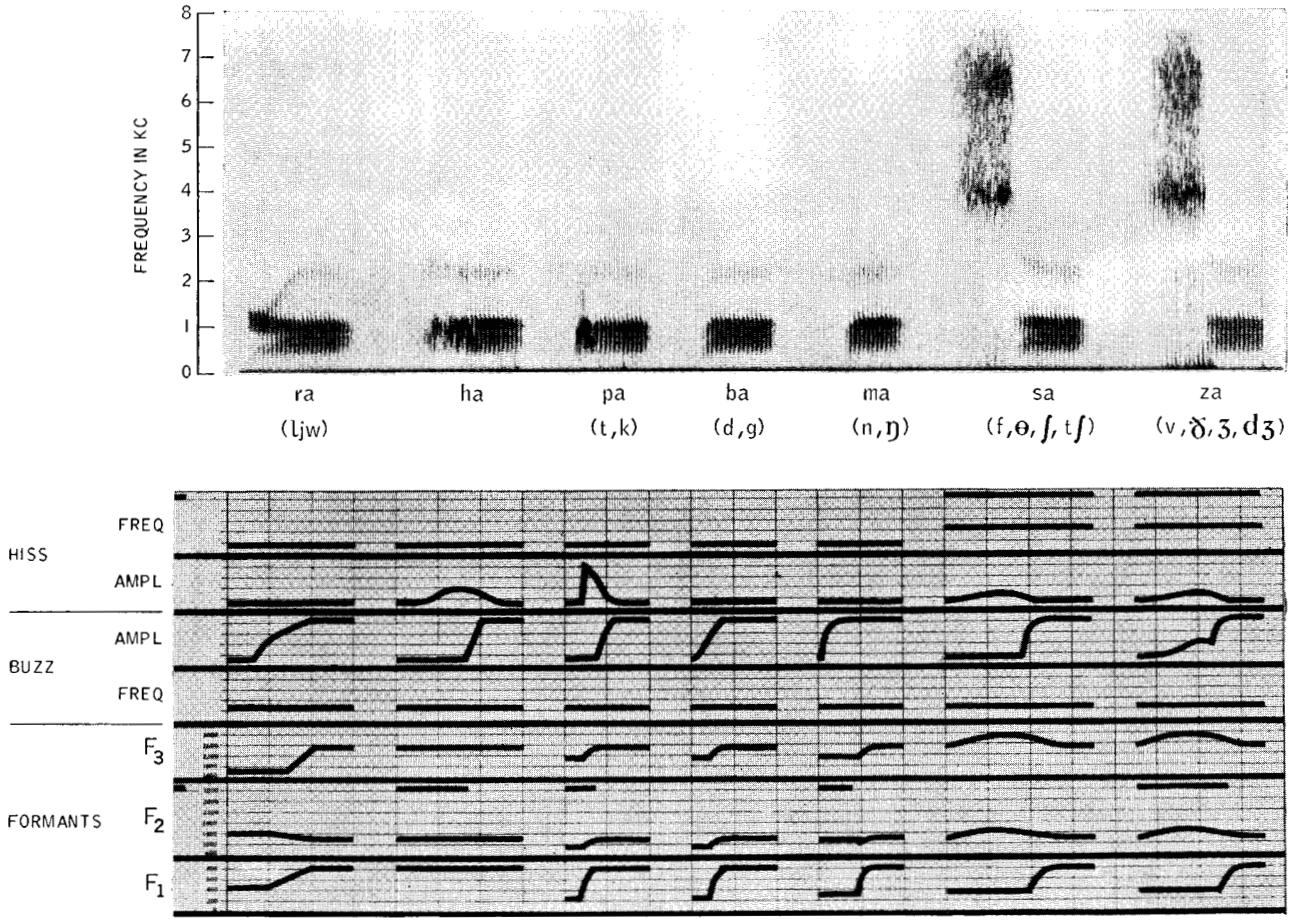


Figure 11 **Synthesis of English consonants.**

can be assembled from the segments (or diphones) shown. Since each vowel has been synthesized with a standard set of formant frequencies, an initial diphone can be assembled with any ending diphone using the same vowel. Therefore, a synthesized set of one-syllable words can be divided into diphones and the resulting diphones reassembled to form other words. Based on a published segment inventory,[17] our present estimate is that we will require a library of approximately 800 such diphones in order to synthesize any word in General American English.

2. Segment library construction

The synthesis for our initial segment library was done on the basis of synthesizing consonant-vowel (cv) or vowel-consonant (vc) diphones in isolation, as shown in Fig. 11. We found that many diphones which were judged to be very intelligible in isolation did not produce words with good over-all phonetic balance when assembled into words by the computer. Likewise, many words which were synthesized and evaluated in isolation did not always fit smoothly into sentences, usually because the consonants were too long. The initial synthesis was in monotone, but it was found that modifying the computer program to include specification of naturally occurring intonation was an aid in judging the naturalness of words. For these reasons, further synthesis is following two parallel paths. One effort is in synthesizing single words from which diphones can be extracted for entry into the segment library. The second effort is in using the available segment library to generate sentences for use in studying the effect of pitch and timing as well as to provide a natural context in which to evaluate new words.

3. Computer-assembled speech

At the present time, the synthesizer has the capability of producing most of the 800 diphones although only about 100 of them have been synthesized on the plastic control tapes so far. However, based on the diphones that are available, the computer has been used to assemble word lists containing variations of word length and endings as well as short sentences to be used in the other studies.

Some examples of the sentences that have been generated are the following:

WHICH IS HIS HAT?
THAT IS HIS HAT.
BUT, BILL SAT ON IT.
THAT IS TOO BAD.
BILL IS A NIT WIT.
TO BE OR NOT TO BE.
AN EYE FOR AN EYE.

4. Naturalness and intelligibility

The most natural-sounding synthetic speech has been obtained from manually prepared control tapes of complete sentences, where the initial values of the control parameters have been taken from sound spectrograms of human speech; these initial values were then adjusted experimentally to give the best perceptual effect.

The naturalness and intelligibility of the computer-assembled synthetic speech depends not only on having good standard segments, but also on the insertion of appropriate pitch and timing data. Efforts are being made to develop a program to generate pitch and timing data automatically, given an adequate grammatical and syntactic specification of the sentence, but at present the programmer must supply these data along with his specification of the phonetic sequence. Although the synthetic utterances produced by the computer-assembly method are judged to be less natural-sounding than those from the control tapes, informal listening tests have indicated that they are still quite intelligible, even to naive listeners.

Acknowledgments

We wish to express our appreciation of the assistance received from our consultant, Prof. P. Delattre, of the University of Colorado.

Other members of our group who have made valuable contributions to the design and experimental use of the system are R. Barber, G. Enke, and H. M. Truby. Programming assistance was supplied by I. Tang of the Advanced Systems Development Division.

We also gratefully acknowledge the contributions of L. Draper, B. Moulds, H. Kraus and H. Sano to the design and construction of the synthesizer and control system hardware.

References and footnotes

1. H. W. Dudley, "The Vocoder," *Bell Laboratories Record*, **18**, 122 (1939).
2. J. L. Flanagan and A. S. House, "Development and Testing of a Formant-Coding Speech Compression System," *Journal of the Acoustical Society of America*, **28**,1099 (1956).
3. C. R. Howard, "Speech Analysis-Synthesis Scheme using Continuous Parameters," *Journal of the Acoustical Society of America*, **28**, 1091 (1956).
4. L. J. Gerstmann and J. L. Kelly, "An Artificial Talker Driven from a Phonetic Input," *Journal of the Acoustical Society of America*, **33**, 835 (A), (1961).
5. G. E. Peterson and W. S-Y. Wang, "Segmentation Techniques in Speech Synthesis," *Journal of the Acoustical Society of America*, **30**, 739 (1958).
6. Cyril M. Harris, "A Study of the Building Blocks in Speech," *Journal of the Acoustical Society of America*, **25**, 962 (1953).
7. K. N. Stevens, S. Kasowski, C. G. M. Fant, "An Electrical Analog of the Vocal Tract," *Journal of the Acoustical Society of America*, **25**, 734 (1953).
8. W. Lawrence, "The Synthesis of Speech from Signals Which Have A Low Information Rate," *Communication Theory*, London, 1953.
9. G. Fant, "Acoustic Analysis and Synthesis of Speech with Application to Swedish," *Ericsson Technics* **15**, No. 1, 3 (1959).
10. The term *formant frequency* is an acoustic-phonetic term that refers to the frequency of the maximum of a gross concentration of energy in the spectrum of a speech sound, as indicated by a peak in the envelope of the spectrum. The formants are identified by number, with the lowest formant being identified as the first, etc.
11. J. L. Flanagan, "Perceptual Criteria in Speech Processing," *Proceedings of the Stockholm Speech Communication Seminar*, August 1962.
12. K. N. Stevens, "Synthesis of Speech by Electrical Analog Devices," *Journal of the Audio Engineering Society*, **4**, 2 (January, 1956).
13. J. L. Flanagan, "Note on the Design of Terminal Analog Speech Synthesizers," *Journal of the Acoustical Society of America*, **29**, 306 (1957).
14. R. A. Schaefer, "A New Pulse Modulator for Accurate DC Amplification with Linear or Nonlinear Devices," *IRE Transactions on Instrumentation*, **2**, No. 3, 34 (September, 1962).
15. J. M. Heinz and K. N. Stevens, "On the Properties of Voiceless Fricative Consonants," *Journal of the Acoustical Society of America*, **33**, 589 (1961).
16. Pierre Delattre, "Acoustic Cues in Speech: First Report," (includes an extensive bibliography), *Phonetica*, **2**, 108–118, 226–251 (1958).
17. Eva Sivertsen, "Segmentation Inventories for Speech Synthesis," *Language and Speech*, 27 (January/March 1961).

Received September 16, 1963