Synthetic voices for computers

The concept of giving voices to computers has been the subject of considerable study in recent years. Formant synthesis and text synthesis represent two approaches to the problem

J. L. Flanagan, C. H. Coker, L. R. Rabiner, R. W. Schafer, N. Umeda
Bell Telephone Laboratories, Inc.

The two methods described for giving voices to computers recognize the importance of economical storage of speech information and extensive vocabularies, and consequently are based on principles of speech synthesis. The first, formant synthesis, generates connected speech from low-bit-rate representations of spoken words. The second, text synthesis, produces connected speech solely from printed English text. For both methods the machine must contain stored knowledge of fundamental rules of language and acoustic constraints of human speech. Formant synthesis from an input information rate of about 1000 bits per second is demonstrated, as is text synthesis from a rate of about 75 bits per second. To give the reader an opportunity to evaluate some of the results described, a sample recording is available; see Appendix A for details.

Voice output from machines

If computers could speak their answers—as well as print them and display them graphically—digital machines could be applied effectively to an expanded range of problems. Typical uses would include automatic information services, computer-based instruction, reading machines for the blind, and spoken status reports from aircraft and space-vehicle systems. Vast amounts of information would be only as far away as the closest push-button telephone. For example, a physician sitting in his office might need information on some obscure disease. It would be convenient if he could dial a computer, key in a reference number, and hear a page or two "read" to him out of a medical encyclopedia. A prospective air traveler might dial a computer, enter destination and desired departure time, and have the computer make combinational searches through timetables and report verbally the convenient connecting flights.

For such applications, the computer must have a large and flexible vocabulary. It therefore must store sizable quantities of speech information, and it must have the information in a form amenable to producing a great variety of messages.* Speech generated by the machine must be as intelligible as natural speech. It need not, however, sound like any particular human and might even be permitted to have a "machine accent." Toward these objectives we have explored two methods of obtaining voice response from a computer, both of which at present appear feasible and attractive.

The first method is called *formant synthesis*. It depends upon an initial, automatic analysis of human speech to produce a synthetic vocabulary. Word libraries are analyzed and stored in terms of formant frequencies. Formants are the natural resonances of the vocal tract, and they take on different frequency values as the vocal tract changes its shape during talking. Typically for nonnasal, voiced sounds three such resonances occur in the frequency range 0 to 3 kHz. The word-length formant data are accessed upon program demand, and are concatenated to form complete formant functions for an utterance. The formant functions have to be interpolated naturally across word boundaries, and voice pitch and word duration have to be calculated according to linguistic rules. Economy in storage derives from the fact that the formant and excitation parameters change relatively slowly and can be specified by fewer binary numbers (bits) per second than can, for example, the speech waveform.

The second method is *synthesis from printed text*—that is, speech synthesis literally from the printed page. In this method, no element of human speech is involved.

* Voice response is already being used in a number of limited-vocabulary applications. Present methods mainly employ prerecorded messages, which are stored and accessed on demand. The limitations of storage and vocabulary size are factors that synthetic speech aims to overcome.

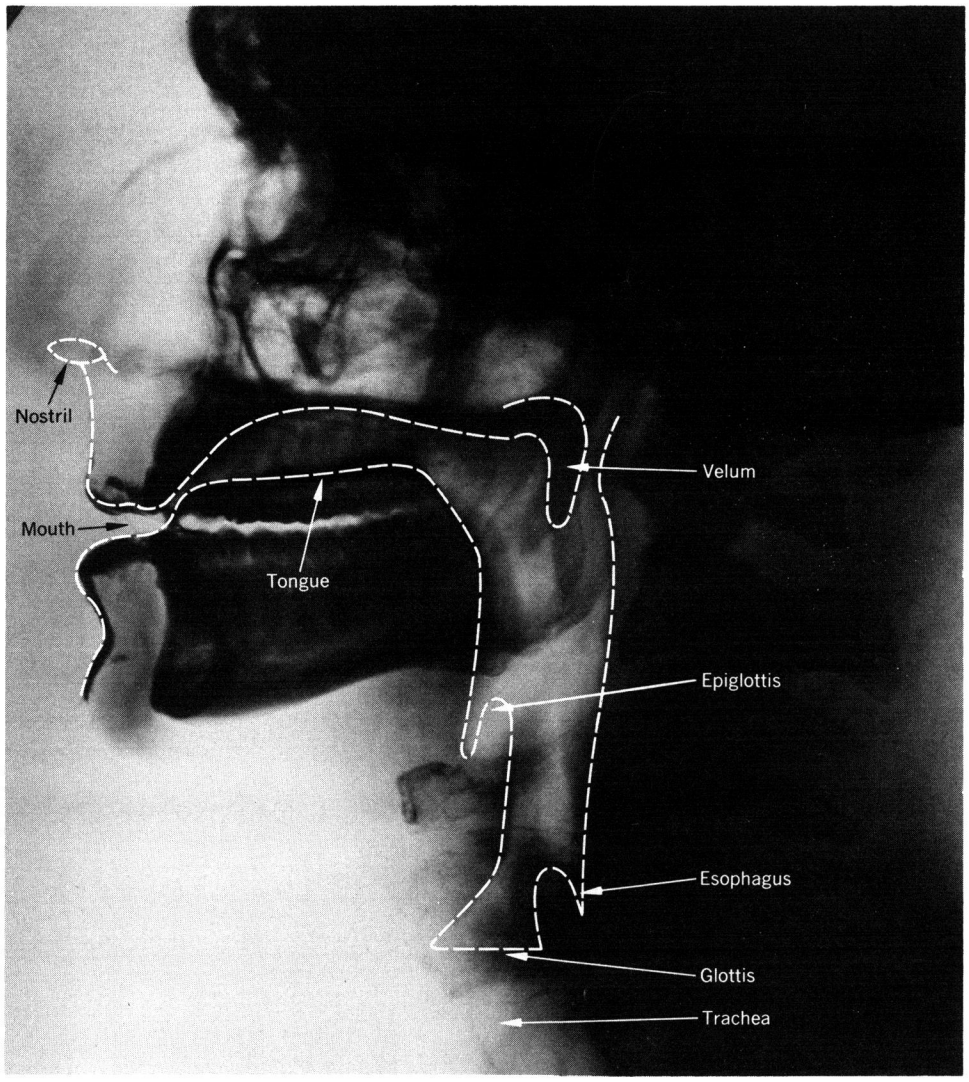


FIGURE 1. Sagittal-plane X ray of a man's vocal tract.

The method depends solely upon the machine knowing, as best it can, the rules and constraints of human speech production and language. An automatic syntax analysis is first made of the text to be spoken. Sound pitch and duration are computed from stored rules about English language. A sequence of vocal tract shapes is then calculated to correspond to the message. Economy of storage results from being able to represent the alphabetic text characters by very few bits per second.

The complexities and ambitions of the two methods can be put into focus at the outset. This is conveniently done by comparing their typical data rates to the data rate for a digitized speech waveform.

I. Information rates for different forms of digital speech data

Stored Data	Bit Rate, b/s	Duration of Speech per 10^6 Bits of Storage
Digitized waveform (PCM)	50 000	20 seconds
Formant data	1 000	17 minutes
Printed text	75	4 hours

Parametric (reduced-data) storage of speech information. A comparison of information rates corresponding to digitized waveform, formant data, and printed text encodings of speech information is given in Table I. The duration of speech that can be stored in 10^6 bits is also shown for the three cases.

It can be seen that the waveform, without further coding, requires around 50 000 b/s (bits per second); that is, the signal is typically sampled at the Nyquist rate and quantized to about 7 bits on a logarithmic scale. Storage capacity of 10^6 bits can therefore accommodate only about 20 seconds of speech in this form. Further, this signal cannot be satisfactorily chopped up and used to fabricate messages different from that originally spoken.

Formant data, on the other hand, require an information rate around 1000 b/s—a reduction of 50:1 over the waveform. (The subsequent discussion will reveal the nature of the compression.) In this case, a store of 10^6 bits can accommodate about 17 minutes of continuous speech. Equally important as the saving in storage is the

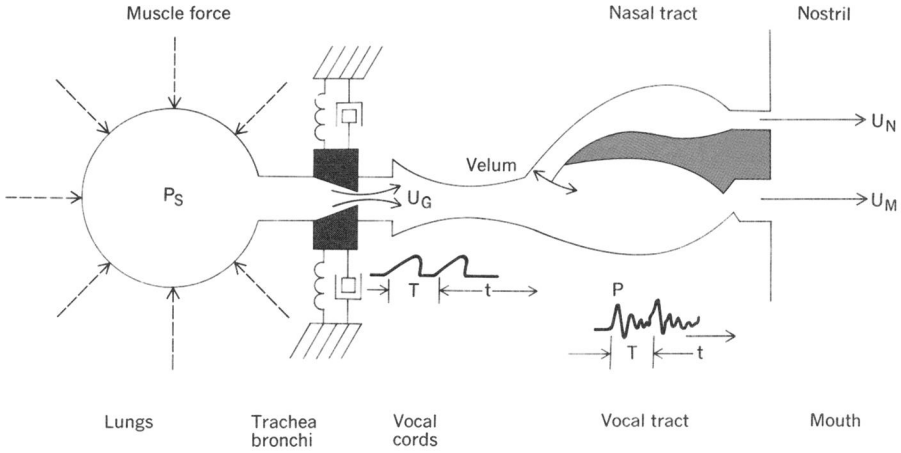
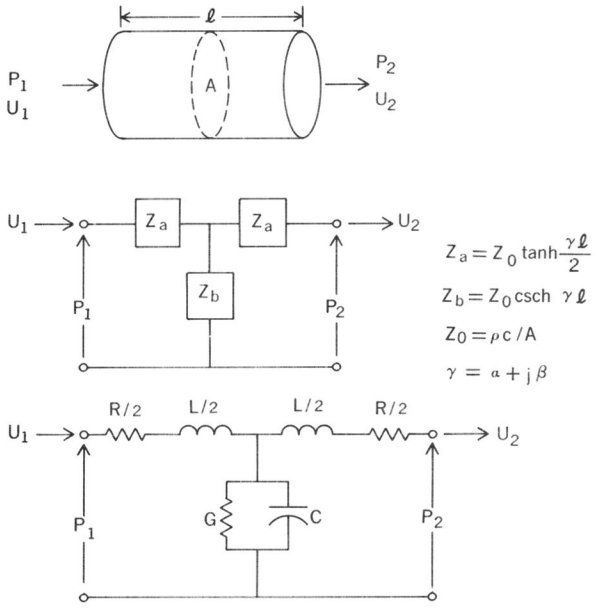


FIGURE 2. Schematic diagram of the human vocal system.

FIGURE 3. Equivalent network for plane-wave propagation in a right circular pipe.



fact that formant-coded words and phrases can be connected together in a relatively natural way to form a variety of messages. This possibility makes formant synthesis attractive for a range of voice services, such as inventory reporting, automatic information operators, and machine instruction.

The ultimate in data reduction and flexibility is printed text, the third entry of Table I. Printed text converted to conversational speech corresponds to an information rate around 75 b/s. About 5 bits are required to specify an alphabetic character and punctuation; words on the average have about 5 characters; and conversational-rate speech corresponds to about 3 words per second. A storage of 10^6 bits therefore provides about four hours of speech. If the available storage were, say, 10 million 36-bit words (not uncommon with larger machines), then about two months of continuous voice output could be provided. In this form, then, the storage of encyclopedic material is quite feasible. The synthesis problem, however, is relatively complex, and the computations that are used to obtain the synthetic output are more consuming of machine time.

In the following sections we explain these two methods and indicate their stages of development. The acoustic and linguistic fundamentals that underlie both synthesis methods are the same. A convenient point of departure is a summary of these relations.

The speech signal

Acoustics of speech production. The major parts of a man's vocal apparatus are shown in the sagittal-plane X ray of Fig. 1. The vocal tract proper is a nonuniform acoustic tube about 17 cm in length. It is terminated at one end by the vocal cords (or by the opening between them, the glottis) and at the other by the lips. The cross-sectional area of the tract is determined by placement of the lips, jaw, tongue, and velum, and can vary from zero (complete closure) to about 20 cm^2.

An ancillary cavity, the nasal tract, can be coupled to the vocal tract by the trapdoor action of the velum. The nasal tract begins at the velum and terminates at the nostrils. In man the cavity is about 12 cm long and has a volume of about 60 cm^3. In nonnasal sound the velum seals off the nasal cavity and no sound is radiated from the nostrils.

Sound can be generated in the vocal system in three ways. Voiced sounds are produced by elevating the air pressure in the lungs, forcing a flow through the vocal-cord orifice (the glottis) and causing the cords to vibrate. The interrupted flow produces quasiperiodic, broad-spectrum pulses, which excite the vocal tract. Fricative sounds are generated by forming a constriction at some point in the tract, usually toward the mouth end, and forcing air through the constriction at a sufficiently high Reynolds number to produce turbulence. A noise source of sound pressure is thereby created. Plosive sounds result from making a complete closure, again usually toward the front, building up pressure behind the closure and abruptly releasing it. All these sources are relatively broad in spectrum. The vocal system acts as a time-varying filter to impose its spectral characteristics on the sources.

The vocal system can be schematized as shown in Fig. 2. The lungs are represented by the air reservoir

at the left. The force of the rib-cage muscles raises the air in the lungs to subglottal pressure P_S. This pressure expels a flow of air with volume velocity U_G through the glottal orifice and produces a local Bernoulli pressure. The vocal cords are represented as a mechanical oscillator composed of a mass, spring, and viscous damping. The cord oscillator is actuated by a function of the subglottal and Bernoulli pressures. The sketched waveform shows the form of the U_G flow during voiced sounds. The vocal tract and nasal tract are shown as tubes whose cross-sectional areas change with distance. The acoustic volume velocities at the mouth and nostrils are U_M and U_N, respectively. The sound pressure P, in front of the mouth, is approximately a linear superposition of the time derivatives $\dot{U}_M$ and $\dot{U}_N$.

A factor of primary interest is the transmission characteristic of the vocal system. The tract length is comparable to a wavelength at all speech frequencies of interest. Its cross dimensions, however, are relatively small. If the tract is considered hardwalled, lossless, and with no side branches, a numerical solution of a one-dimensional steady-state wave equation with nonconstant coefficients (Webster's horn equation) yields the undamped eigenfrequencies (or resonances) of the system.[1,2] On the other hand, a bilateral transmission-line equivalent, useful for digital simulation of actual pressure and velocity relations, including those of the vocal cords, nasal tract, and the subglottal system, can also be obtained.[1]

Consider the variable-area pipe to be composed of elemental right-circular pieces, one of which is shown in Fig. 3. Sound pressure and volume velocity at the two ends are represented by P_1, U_1 and P_2, U_2, respectively. For plane-wave propagation, the pipe element of length l has an equivalent T-section in which the impedance elements Z_a and Z_b are hyperbolic functions of the complex acoustic propagation constant γ. (The acoustic network has exactly the same form as that for a uniform electrical line.) The characteristic impedance Z_0 is the product of air density ρ and sound velocity c divided by the cross-sectional area A. For a given quantal length, then, the impedance elements are determined solely by the cross-sectional area. The first terms in the series expansions of the hyperbolic functions for Z_a and Z_b give the acoustic elements R, L, G, C of Fig. 3. The loss R arises from viscous loss at the walls of the pipe; the inertance L is due to the mass of air in the elemental cylinder; the loss G results from the heat conduction at the walls; and the capacity C arises from the compressibility of air in the volume Al.*

The schematic system of Fig. 2 can therefore be decomposed into elemental right-circular pieces and represented for computation and simulation by the bilateral network of Fig. 4. Network elements correspond to those parts shown in Fig. 2. Consider voltage analogous to pressure and current analogous to volume velocity. The lung volume is represented by a capacity and loss whose sizes depend upon the state of lung inflation. The lungs are connected to the vocal cords by the trachea and bronchi tubes, represented in Fig. 4 as a single T-section. The impedance of the vocal cords Z_G is both time-varying and dependent upon the glottal volume velocity U_G.[3] The vocal tract is approximated as a cascade of T-sections in which the element impedances are determined by the cross-sectional areas $A_1 \cdots A_N$. The line is terminated in a radiation load Z_M at the mouth, which is taken as the radiation impedance of a circular piston in a plane baffle. U_M is the mouth current and, for simulation of dc quantities, the battery P_A represents atmospheric pressure.

The nasal tract is coupled by the variable velar impedance Z_V. The nasal tract is fixed in shape, and the nostril current U_N flows through the radiation impedance Z_N.

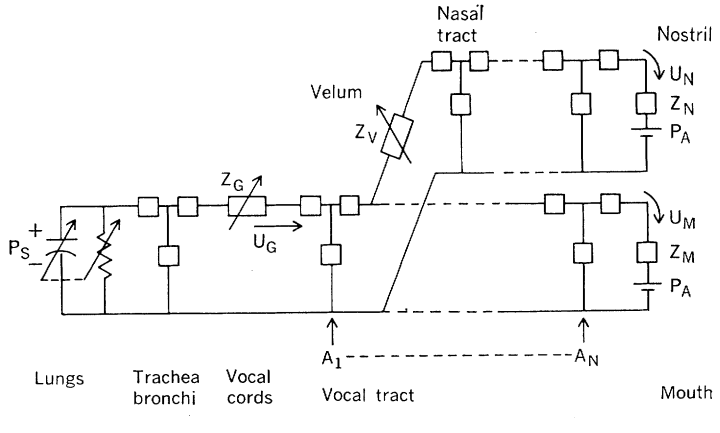
This formulation of the vocal system actually can simulate respiration and phonation. The glottis is opened (Z_G is reduced), the rib-cage muscles enlarge the lung capacitor (volume), and the atmospheric pressure forces a charge of air through the tract and onto the capacitor. The glottis is then clenched and increased in impedance; the rib-cage muscles contract, raising the voltage (pressure) across the lung capacity, and force out a flow of air.

Under proper conditions, the vocal-cord oscillator is set into stable vibration, and the network is excited by periodic pulses of volume velocity. The lung pressure, cord parameters, velar coupling, and vocal-tract area all vary with time during an utterance. A difference-equation specification of the network, with these variable coefficients, permits calculation of the Nyquist samples of the output sound pressure.*

For purposes of computer synthesis, sound may be computed from a bilateral network such as Fig. 4, or the eigenfrequencies (formants) of the system may be obtained and used in a "terminal analog" synthesis.[1] The latter involves control of the transfer function of a variable network so that its unilateral transmission simulates that of the vocal tract. The nature of the unilateral transmission is simply illustrated for a vocal tract in the shape of a straight pipe in Fig. 5. The hard-walled pipe is assumed to be excited by a high-impedance volume velocity source U_G, and the mouth radiation impedance is assumed negligible for simplicity. The Fourier transform of the ratio of mouth and glottal currents, U_M/U_G, is the transmission function of interest.

* One of our computer programs is a vocal-tract synthesizer represented just this way.

FIGURE 4. Network simulation of the vocal system.



* Mechanical yielding of the vocal-tract walls modifies the shunt parameters of the equivalent circuit (see Ref. 1).

As shown, $|U_M/U_G|$ has peaks, or formants, at frequencies where the pipe is an odd quarter-wavelength—that is, at frequencies $f_n = (2n - 1)c/4l$, for $n = 1, 2, \cdots$. For a tract length of 17 cm and a sound velocity of 340 m/s, these frequencies are 500, 1500, 2500, $\cdots$ Hz. The resonances are simple and appear as single complex-conjugate poles in the transmission function. The half-power bandwidths of the resonances are conditioned by the losses in the system and are roughly constant for each formant. The phase shift of the transmission passes through π radians as each formant is traversed in frequency. The phase response is a perturbation about a line of constant slope, namely the transit delay through the tube, l/c. This transmission characteristic is the "filter" that operates on the vocal sound source. Hence the radiated sound bears these resonances.

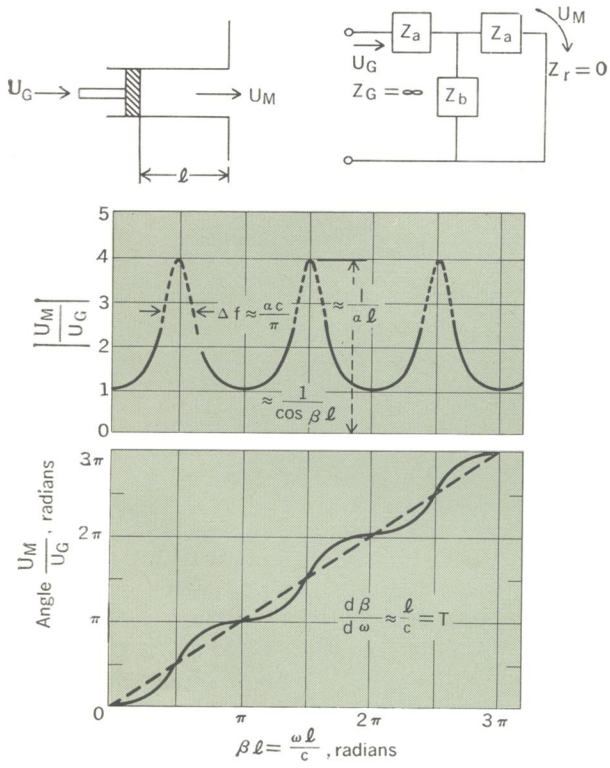


FIGURE 5. Acoustic transmission properties of a straight hardwalled pipe.

For nonnasal voiced sounds, three formants typically fall in the frequency range 0–3 kHz. Because the resonances are simple and have relatively constant bandwidths, the formant frequencies effectively specify the spectrum everywhere. For voiceless sounds, because the sound source is located forward in the tract, one resonance (pole) and one antiresonance (zero) typically describe the transmission in the frequency range 0–3 kHz.

Every shape of the vocal tract has a unique set of formant frequencies and the distinctive sounds of a language have perceptually distinctive formant positions. Idealized vocal-tract transfer functions for several vowels are shown in Fig. 6. Note, for example, the vowel /i/ (as in eat) has typically a low first formant frequency and a high second formant.* By contrast, the vowel /a/ (as in father) has a high first formant proximate to its low second formant. The overall spectral shapes of the two sounds are notably different.

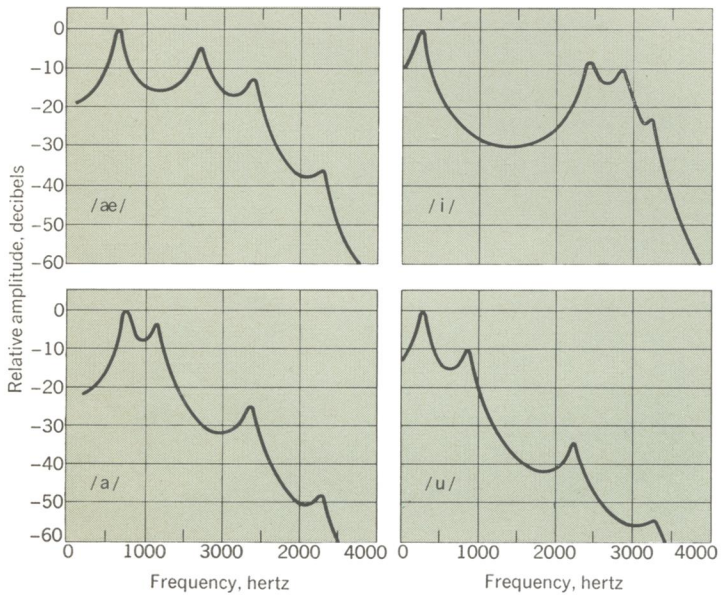
In continuous speech the formant resonances move around as the vocal tract changes shape. Figure 7 shows a sound spectrogram (time–frequency–intensity plot) of a sentence in which the first three formant frequencies are traced. (Dashed lines in Fig. 7 are idealized and may not accurately plot formant transitions—especially in the context of stop and nasal consonants.) These parameters vary slowly because of the physical limitations on how quickly the vocal-tract shape can be changed. Hence they occupy only a small bandwidth. If these resonances can be determined accurately, and preferably automatically, they can be used with data about fundamental voice frequency and intensity to synthesize signals similar to natural speech.

Discrete and dynamic aspects of speech. Speech has long been viewed as a discrete process. The pronunciation key of any dictionary expresses the sounds of the language as a finite set of discrete symbols, each having a relatively invariant sound and vocal-tract shape. One might think, therefore, that a simple means could be devised to represent speech sounds by a truly small inventory of these basic speech sounds, the *phonemes*, and that messages might be composed by concatenation of these small units.

A look at a speech spectrogram, such as Fig. 7, however reveals that speech, at the acoustic level, is not particularly discrete. Spectrograms and X-ray motion pictures show that the articulations of adjacent phonemes interact and that transient movements of the vocal tract for the production of any phoneme last much longer than the average duration of the phoneme (total time divided by number of phonemes); that is, the articulatory gestures overlap and are superposed.

The transient motions of the vocal tract are perceptually important. Experiments show that much information about the identity of a consonant is carried not by the

FIGURE 6. Frequency spectrums of vocal-tract transmission for four vowel sounds.



* Symbols enclosed in slashes are those for the International Phonetic Alphabet.

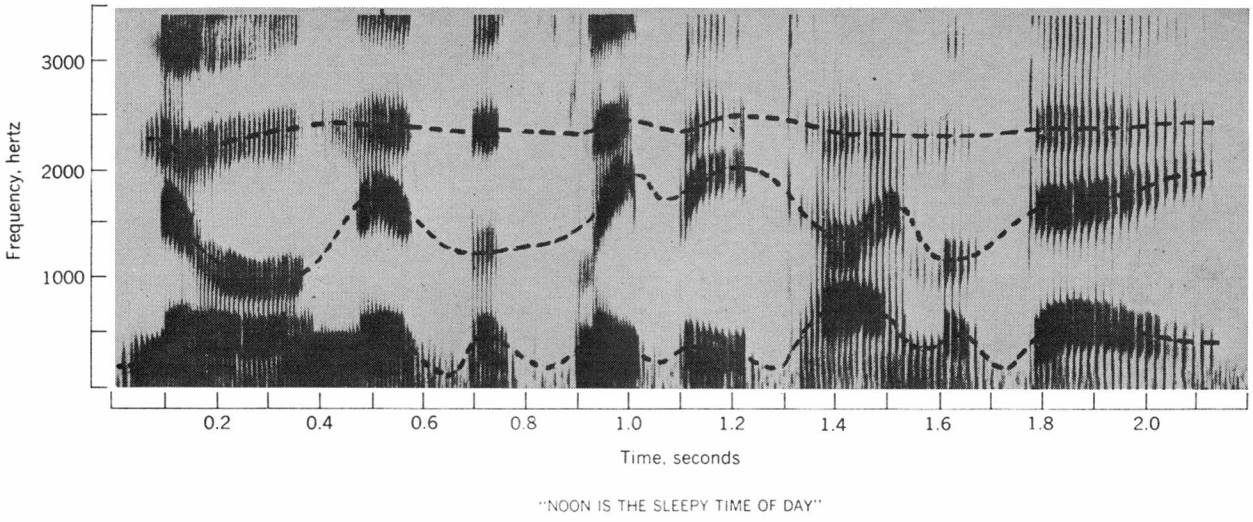


FIGURE 7. Dynamic variation of formant frequencies in connected speech.

spectral shape at the "steady-state" time of the consonants, but by its dynamic interactions with adjacent phonemes.

Speech synthesis is therefore strongly concerned with dynamics. A synthesizer must reproduce not only the characteristics of sounds when they most nearly represent the ideal of each phoneme, but also the dynamics of vocal-tract motion as its progresses from one phoneme to another.

This fact highlights a difference between speech synthesis from word or phrase storage and synthesis from more elemental speech units. If the library of speech elements is to be a small number of short units, such as phonemes, then the concatenation procedures must approach the complexity of the vocal tract itself.

Conversely, if the library of speech elements is a much larger number of longer segments of speech, such as words or phrases, then concatenations can be made at points in the message where information in transients is minimal.

The form in which speech is represented for either long-or short-element storage is somewhat flexible. For short elements, a representation as coordinates of the articulatory system seems advantageous. For word- or phrase-length storage, a formant characterization is especially appropriate.*

Speech synthesis from formant data

Automatic formant analysis of speech.[12] Because speech parameters vary slowly with time, the concept of the short-time spectrum is a basic tool in speech analysis. The Fourier transform of a short segment of the speech waveform reflects features of the excitation and formant frequencies for that segment. Figure 8 illustrates the way in which short-time spectral analysis can be employed in the estimation of speech parameters.

Figure 8(A) depicts the analysis of voiced speech. The waveform at the left is a segment of voiced speech of approximately 40-ms duration, which has been multiplied by a data window (which reduces the undesirable effects of analyzing a finite amount of data). Over such a short time interval, the speech waveform looks like a segment of a periodic waveform. The detailed time variation of the waveform during a single period is determined primarily by the vocal-tract response, whereas the fundamental period (pitch period) reflects the rate of vibration of the vocal cords.

The logarithm of the magnitude of the Fourier transform of this segment of speech is shown as the rapidly varying spectrum at the right. This function can be thought of as consisting of an additive combination of two components: a rapidly varying periodic component associated primarily with the vocal-cord excitation, and a slowly varying component primarily attributable to the vocal-tract transmission function. Therefore, the excitation and vocal-tract components are mixed and must be separated to facilitate estimation of the parameter values. The standard approach to the problem of separating a slowly varying signal and a rapidly varying signal is to employ linear filtering. One technique for achieving this filtering is through the intermediate computation of the *cepstrum*,[13,14] the inverse Fourier transform of the log-magnitude spectrum.

The cepstrum is plotted in the middle diagram of Fig. 8(A). The rapidly varying component of the log magnitude corresponds to the cepstral peak at about 8 ms (the value of the pitch period). The slowly varying component corresponds to the low-time portion of the cepstrum. Therefore, the slowly varying component can be extracted by first smoothly truncating the cepstrum values to zero above about 4 ms, and then computing the inverse transform. This yields the slowly varying curve that is superimposed on the short-time spectrum, shown at the right in Fig. 8(A). The figure of 4 ms was chosen to be representative of the lower limit of the pitch period for male speakers.

The formant frequencies correspond closely with the resonance peaks in the smoothed spectrum. Therefore, a good estimate of the formant frequencies is obtained by simply determining which peaks in the smoothed spec-

* Many people have investigated approaches to storage of speech for computer voice response. These methods have varied widely in their efficiency and flexibility; see Refs. 4–11.

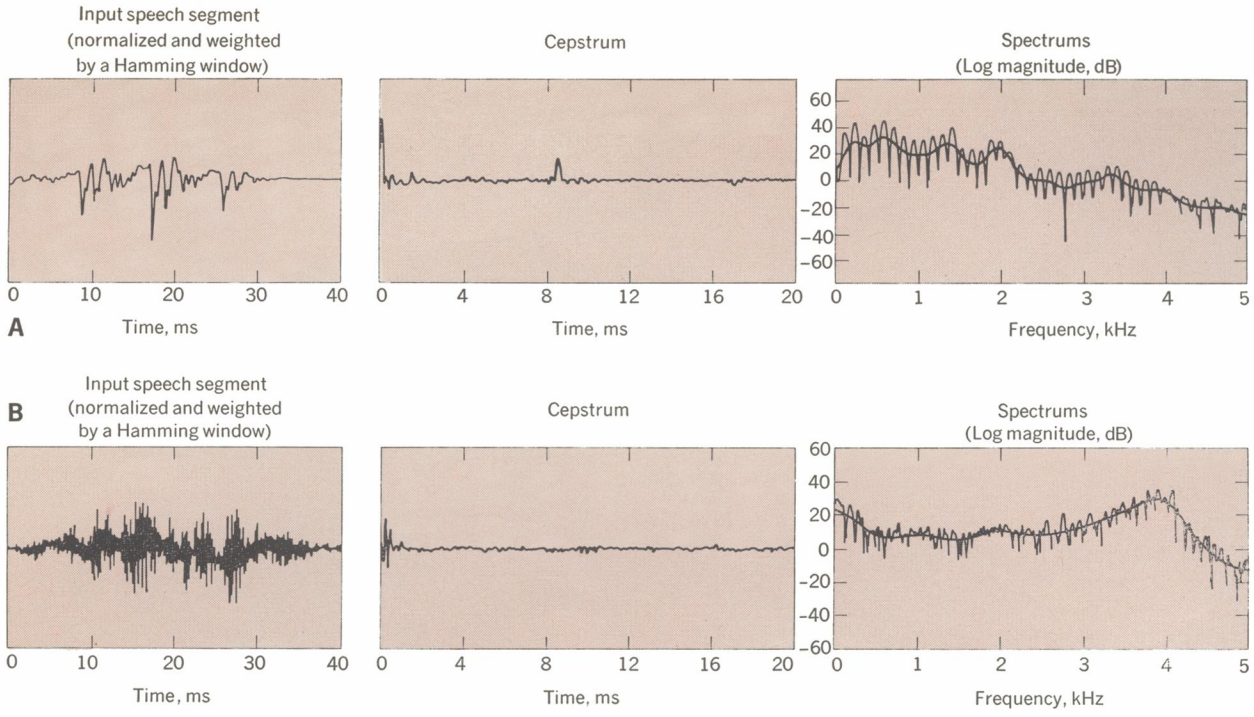


FIGURE 8. Short-time spectrum and cepstrum analysis of (A) voiced and (B) unvoiced speech.

trum are vocal-tract resonances. Acoustic constraints on formant frequencies and amplitudes are incorporated into an algorithm that locates the formant peaks in the smoothed spectrum.[12]

The analysis of unvoiced speech segments is depicted in Fig. 8(B). In this case, the input speech resembles a segment of a random-noise signal. As before, the log-magnitude spectrum of the speech segment can be thought of as consisting of a rapidly varying component associated with excitation, plus a slowly varying component due to the spectral shaping of the vocal tract. In this case, however, the rapidly varying component is not periodic but random. Again the low-time part of the cepstrum corresponds to the slowly varying component of the transform, but the high-time peak present in the cepstrum of voiced speech is absent for unvoiced speech. Therefore, the cepstrum can also be used in deciding whether an input speech segment is voiced or unvoiced. If voiced, the pitch period can be estimated from the location of the cepstral peak.[15] Truncation of the cepstrum and subsequent Fourier transformation produce the smoothed spectrum curve that is superimposed on the short-time transform at the right of Fig. 8(B). An adequate specification of the spectrum of an unvoiced sound can be achieved by estimating the frequency locations of a single wide-bandwidth resonance and a single antiresonance—that is, a single pole and zero.

Continuous speech is analyzed by performing these operations on short segments of speech, which are selected at equally spaced time intervals. Figure 9 illustrates this process for a section of voiced speech. The short-time spectrum and smoothed spectrum corresponding to each cepstrum are plotted adjacent to the cepstrum. Time increases from top to bottom, and each set of curves corresponds to a segment of speech offset 20 ms from the preceding segment. The formant peaks are connected by straight lines. One notices that formants occasionally come close together in frequency and pose a special problem in automatic estimation.

In the third and fourth spectrums* from the top of Fig. 9, the second and third formants are so close together that there are no longer two distinct peaks in the Fourier spectrum. A similar situation occurs in the last four spectrums, where the first and second formants are not resolved. A procedure for detecting such situations and for enhancing the resolution of the formants is shown in Fig. 10.[12]

The upper curve is the smooth spectrum as evaluated along the $j\omega$-axis of the s-plane. (The lowest three eigenfrequencies are depicted in their approximate locations.) Because formants two and three (F_2 and F_3) are quite close together, only one broad peak is observed in the spectrum. However, when the spectrum is evaluated on a contour that passes closer to the poles, two distinct peaks are in evidence, as shown in the lower curve. A computation algorithm known as the Chirp z-transform algorithm facilitates this additional spectral analysis.[16]

Formant synthesis. Once the excitation and transmission parameters are obtained, they are used to synthesize a waveform that approximates the original speech signal. Numerous systems, both analog and digital, have been devised for formant synthesis.[17-22] A digital system is illustrated in Fig. 11. The upper branch produces voiced speech. Its excitation source produces a train of impulses with spacing equal to τ (the fundamental pitch period). The signal A_V, also estimated from the natural speech, controls the intensity of the pulse excitation

* The authors take no credit for the plural terminology—such as spectrums, cepstrums, and phenomenons—used in this article.

applied to a cascade of variable digital resonators. The resonator system is specified (under steady conditions) by the system function

$$H_V(z) = \prod_{k=1}^{4} \frac{1 - 2e^{-\alpha_k T}\cos(2\pi F_k T) + e^{-2\alpha_k T}}{1 - 2e^{-\alpha_k T}\cos(2\pi F_k T)\, z^{-1} + e^{-2\alpha_k T} z^{-2}}$$

where T is the sampling period, the F_k's are the formant frequencies (only three of which are time-varying), and the α_k's are the formant bandwidths (all of which are fixed). The output of this system excites a fixed system whose transfer function is

$$S(z) = \frac{(1 - e^{-aT})(1 + e^{-bT})}{(1 - e^{-aT}z^{-1})(1 + e^{-bT}z^{-1})}$$

This network is a cascade of two simple poles, and is designed to approximate the spectral shaping due to radiation and source properties.

The lower branch of Fig. 11 produces unvoiced speech. A random-noise generator, whose intensity is controlled by the signal A_N, excites a digital filter whose steady-state transfer function is given by the relation

$$H_U(z) = \frac{AB}{CD}$$

where

$$A = 1 - 2e^{-\beta T}\cos(2\pi F_P T) + e^{-2\beta T}$$
$$B = 1 - 2e^{-\beta T}\cos(2\pi F_Z T)\, z^{-1} + e^{-2\beta T} z^{-2}$$
$$C = 1 - 2e^{-\beta T}\cos(2\pi F_Z T) + e^{-2\beta T}$$
$$D = 1 - 2e^{-\beta T}\cos(2\pi F_P T)\, z^{-1} + e^{-2\beta T} z^{-2}$$

In these expressions, F_P and F_Z are, respectively, the time-varying pole and zero center frequencies for the unvoiced sound and β is the fixed bandwidth of both the pole and zero. The output of this system is passed to the fixed spectral compensation filter to provide the unvoiced speech output. Control parameters are supplied to the synthesizer at least at their Nyquist rates, and output samples are generated at a 10-kHz rate.

Figure 12 illustrates an example of automatic analysis and synthesis. Figure 12(A) shows the pitch period and formant signals (each band-limited to 16 Hz) as automatically estimated from a natural utterance. Figure 12(C) shows the spectrogram of speech synthesized from the estimated control parameters. For comparison, Fig. 12(B) shows the spectrogram of the original signal. Figure 13 shows spectrograms of original and synthetic versions of another utterance. Comparison of the spectrograms for the original and synthetic signals indicates that spectral properties are reasonably well preserved. [The best way to evaluate the results is actually to listen to them. For this purpose, a recording is available in conjunction with this article. Appendix A gives the contents of the record and information on how it may be obtained. Section 1 of the recording illustrates automatic analysis and synthesis.]

Efficiency of formant synthesis. The storage efficiency of the formant representation of speech depends on the precision with which the basic parameters must be specified. Synthetic speech of high quality can be obtained if the pitch period is specified to the nearest 0.1 ms, the gain specified to one place in 100, and the formant frequencies to the nearest 1 Hz. Since the parameters are estimated and supplied to the synthesizer 100 times per

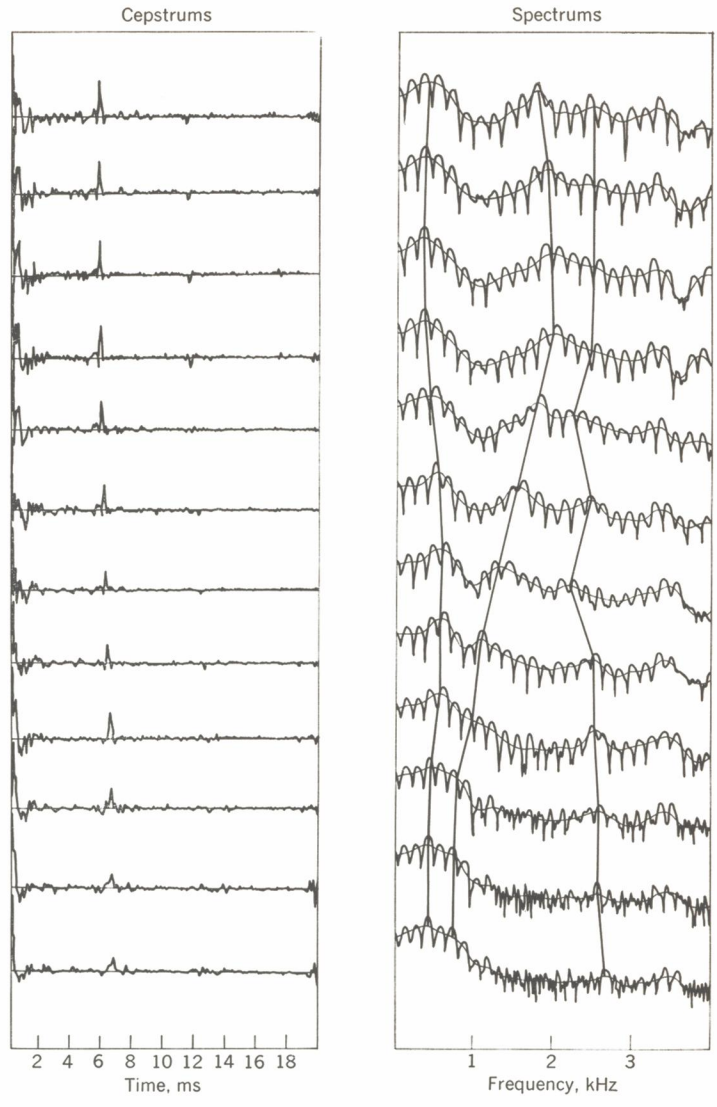
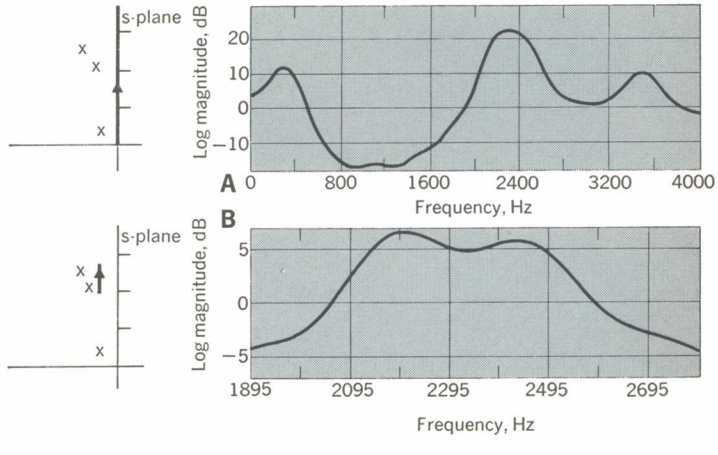


FIGURE 9. Cepstrum analysis of continuous speech. The left row shows cepstrums of consecutive segments of speech separated by 20 ms. The right row shows the corresponding short-time and cepstrally smoothed spectrums.

FIGURE 10. Illustration of the enhancement of formant resonances. A—Cepstrally smoothed spectrum in which F_2 and F_3 are not resolved. B—Narrow-band analysis along a contour passing closer to the poles.



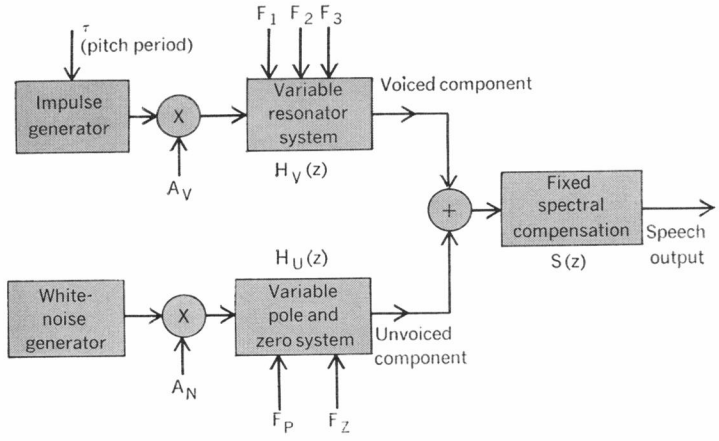


FIGURE 11. Digital formant synthesizer.

second, the required information rate can be shown to be 4600 b/s. (See Appendix A for a derivation of this figure.) Although this represents a considerable saving over the PCM representation, it is important to determine how much the information rate can be lowered before there is a significant perceptual difference between speech synthesized at 4600 b/s and speech synthesized at a lower rate.

In investigating this question, two considerations are important. The first is the bandwidth required to preserve the essential variations of the parameters. The second is the degree to which the synthesis parameters may be quantized. Sampling the parameters 100 times per second implies that they occupy a bandwidth of 50 Hz. In fact, their bandwidth occupancy is much smaller.

Figure 14(A) shows the formant frequencies as automatically estimated from natural speech at a rate of 100 Hz — that is, allowing a 50-Hz bandwidth for each parameter. Figure 14(B) shows the same data smoothed by a 16-Hz low-pass filter. Similarly, Fig. 14(C) shows the results for a 4-Hz low-pass cutoff. An auditory comparison of conditions represented by Fig. 14(A) and (B) shows that 16-Hz smoothing has negligible effect on the parameter variations. However, Fig. 14(C) shows that the 4-Hz filter has smeared out the waveforms significantly. Preliminary perceptual experiments indicate that the parameters can be band-limited to approximately 12 to 16 Hz with no noticeable degradation.[23] These results imply that the required sampling rate may be no higher than about 35 times per second. [Section 2 of the record illustrates the effect of low-pass-filtering the speech parameters.]

The second consideration involves the quantization of the parameter samples. Experiments indicate that pitch must be quantized to approximately 6 bits, whereas the remaining parameters each require 4 bits or less. Figure 15 shows a comparison between formant functions determined at 100 Hz and quantized to 11 bits and those sampled at 35 Hz and quantized to 3, 4, and 2 bits, respectively. The total bit rates for the syntheses are 4600 and 600 b/s, respectively. The two results are nearly indistinguishable. [Sections 3 and 4 of the record illustrate quantizing effects and give an example of speech synthesized at 600 b/s.]

It should be pointed out that these results are not meant to be taken as completely general. The exact bit rate depends on many factors, such as speaking rate and the nature of a particular speech utterance, which can vary widely. The results do give an idea of how low the information rate can become, and demonstrate that intelligible and natural-sounding synthetic speech can be generated from data rates of the order of 1000 b/s.

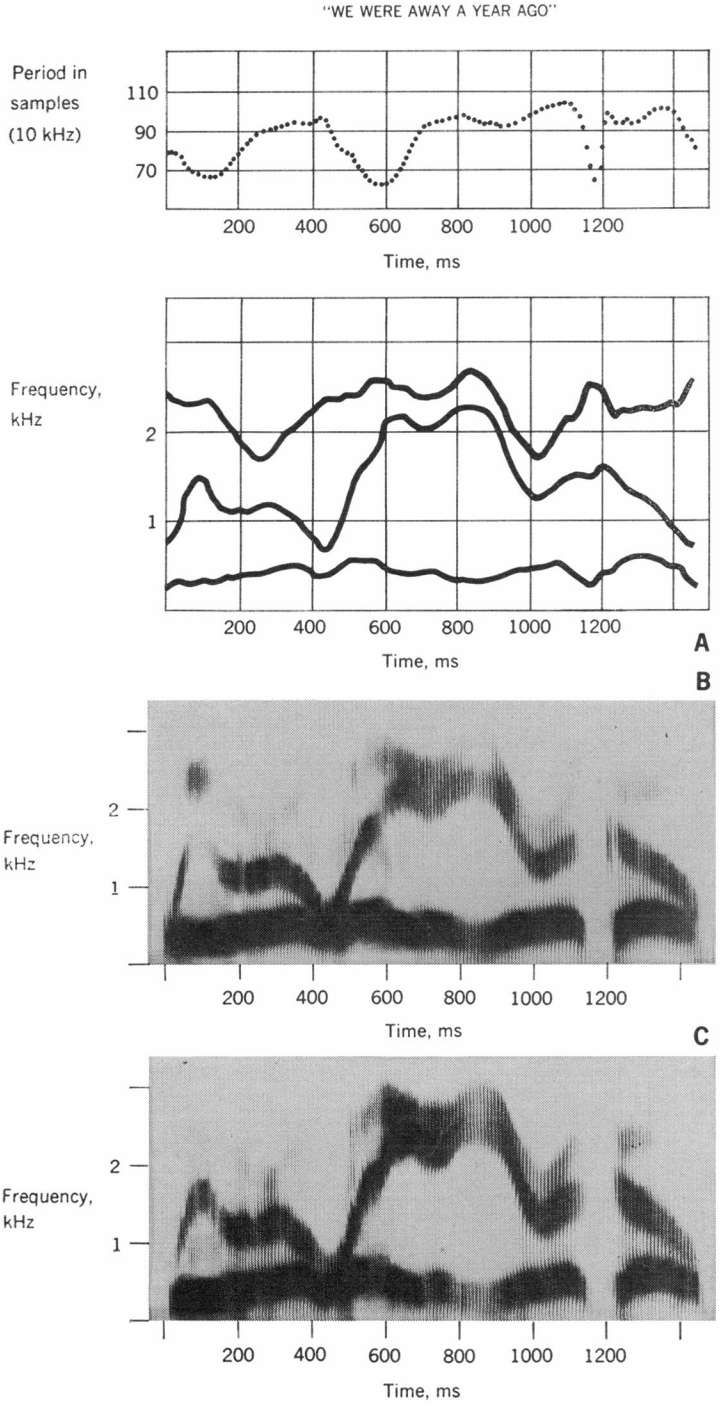


FIGURE 12. Automatic analysis and synthesis. A—Pitch period and formant frequencies estimated from natural speech. B—Spectrogram of the original speech. C—Spectrogram of synthetic speech.

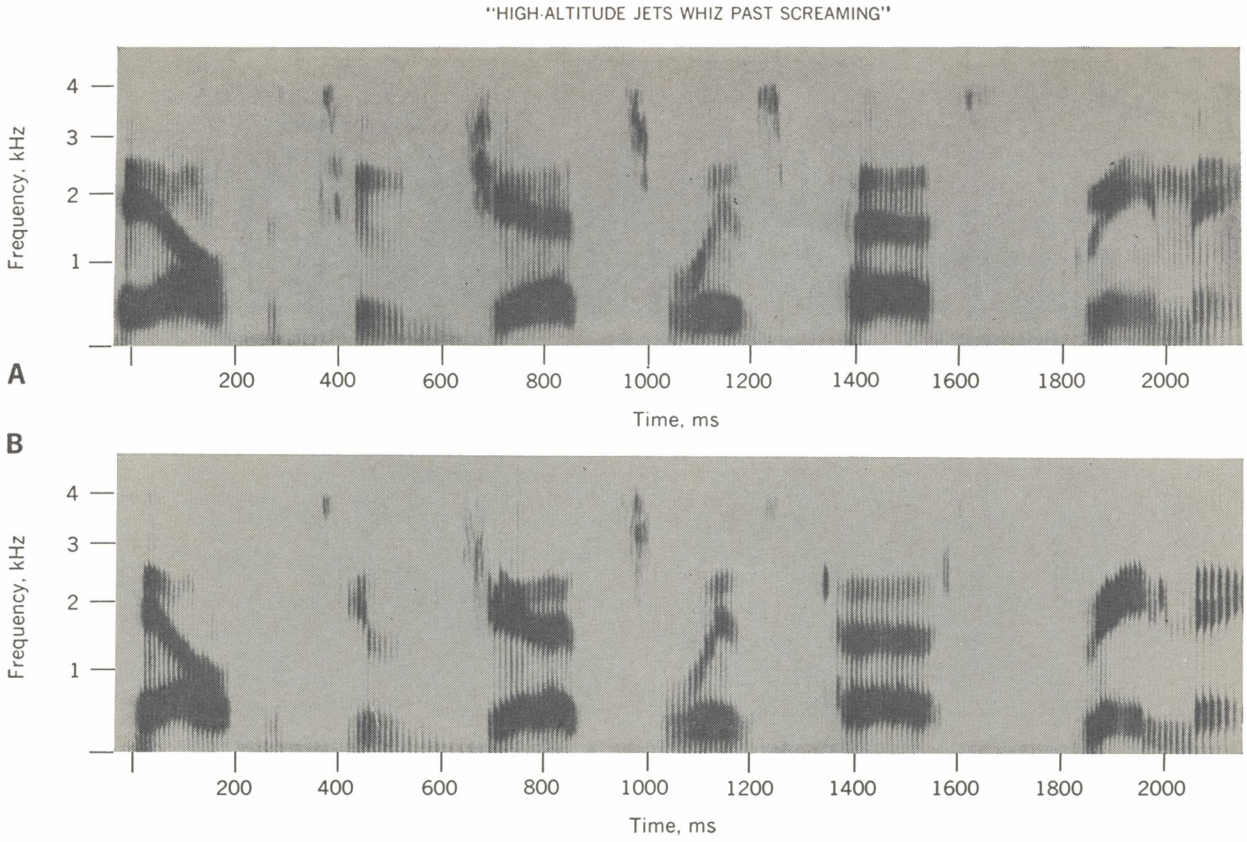
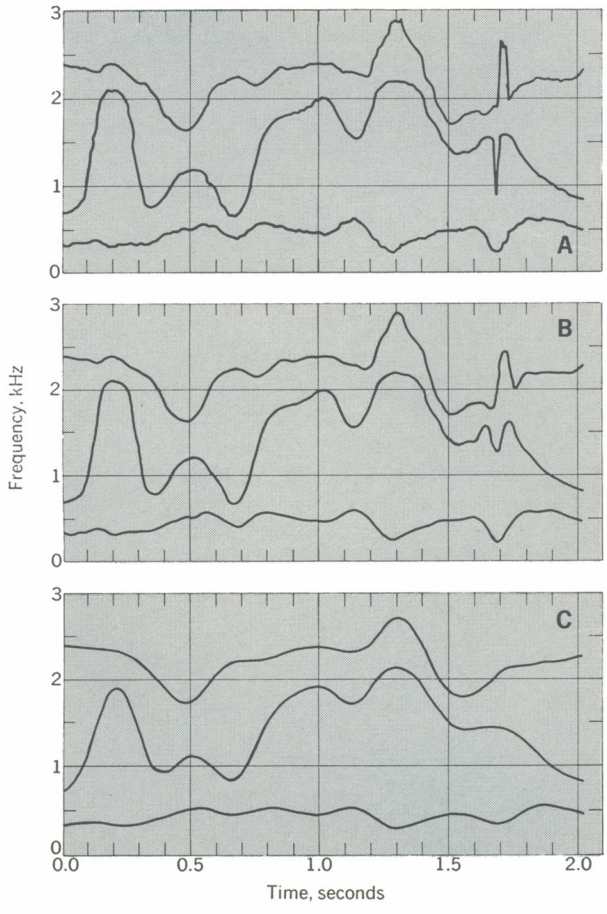


FIGURE 13. Automatic analysis and synthesis. A—Spectrogram of original speech. B—Spectrogram of synthetic speech.

FIGURE 14. Smoothing of formant signals. A—50-Hz bandwidth (no smoothing). B—16-Hz bandwidth. C—4-Hz bandwidth. (Sampling rate 100 Hz throughout.)



Speech synthesis by concatenation of formant data. The previous discussion emphasized the storage economy that can be realized by coding speech in terms of formant and excitation data. Another advantage of formant-coded speech is equally important—namely, its flexibility for fabricating messages from preanalyzed, naturally spoken, isolated words.

In the formant representation of an utterance, formant frequencies, voice pitch, amplitude, and timing can all be manipulated independently. Thus in synthesizing an utterance one can substitute an artificial pitch contour* for the natural contour. A steady-state sound can be lengthened or shortened, and even the entire utterance can be speeded up or slowed down with little or no loss in intelligibility. Formants can be locally distorted, and the entire formant contour can be uniformly raised or lowered to alter voice quality. [Section 5 of the record demonstrates synthetic speech where one or more of the basic parameters have been manipulated.]

A concatenation program that uses the flexibility of formant-coded speech to synthesize message-length utterances is shown in Fig. 16. Words spoken in isolation are automatically analyzed to estimate the parameters required by the synthesizer of Fig. 11. Quantized Nyquist rate samples of these control signals are stored in a word catalog. To synthesize a message composed of words from the catalog, the printed word string is supplied to the concatenation program. The program uses separate strategies for deriving the "segmental features" of the

* "Contour" is used to mean the time course of the relevant parameter.

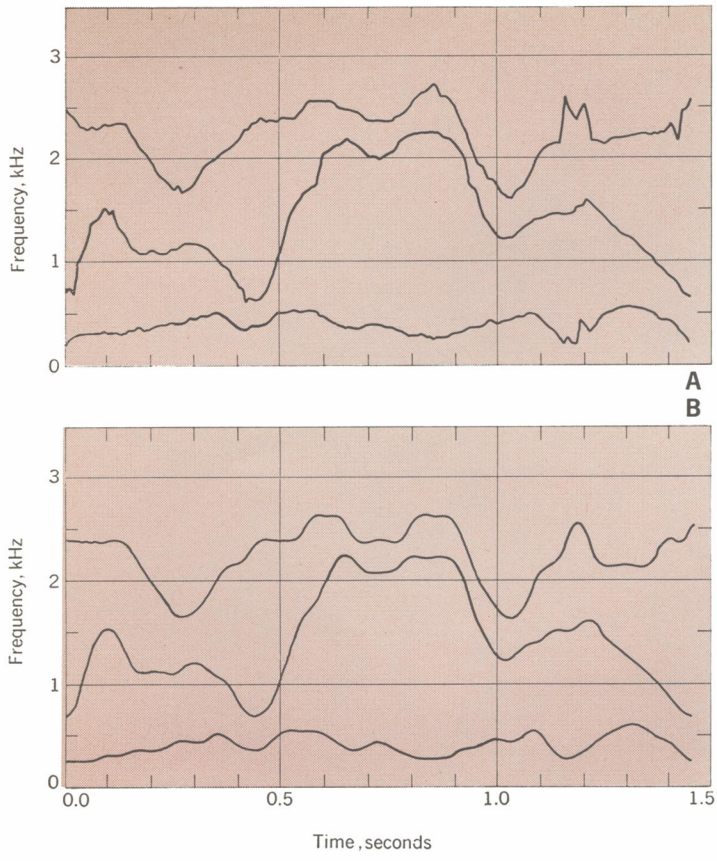
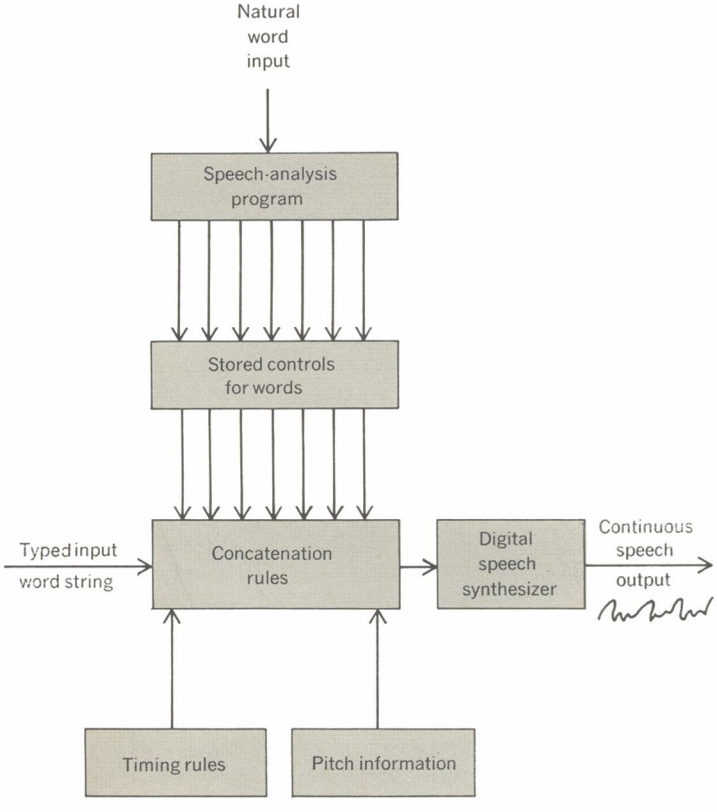


FIGURE 15. Quantization and smoothing of formant signals. A—Unquantized formants (4600 b/s). B—Quantized and smoothed formants (600 b/s).

FIGURE 16. Block-diagram representation of concatenation program.



message (formant frequencies and unvoiced pole and zero frequencies) and the "prosodic features" (timing, amplitude, and pitch). Figure 16 shows separate inputs for timing rules and pitch information. The program strategy for deriving segmental features is included in the box labeled "concatenation rules." It is the flexibility in manipulating formant-coded speech that permits breaking the synthesis problem into two parts. The output of the concatenation program is a set of smoothly varying formant-synthesis parameters, which are supplied to a digital synthesizer of the type shown in Fig. 11.*

To synthesize a continuous message, timing, pitch, and formant information must be generated. Timing information is derived in several ways. The techniques employed include

1. External specification of the duration of each word in the input string to be synthesized. In this case, word duration is chosen according to some external criterion (e.g., it can be measured from a naturally spoken version of the message to be synthesized) and in no way is meant to be a typical duration, independent of context.

2. Calculation of word duration by rules based on models of English language. Rules of this type are described in the next section and are used for synthesis from printed text.

3. Specification of word duration from tables of stored data. For limited-context messages, such as sequences of digits, such specification of word duration often is acceptable.

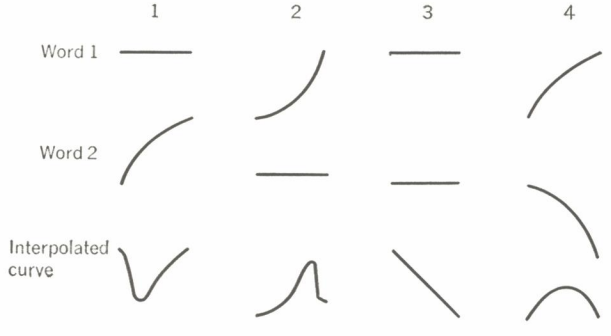
The next step in the synthesis procedure is to determine the pitch contour (e.g., the pitch period as a function of time) for the message to be synthesized. Pitch information can be obtained in several ways. The use of the pitch information measured from the originally spoken words and the use of monotone pitch are possibilities, but both give unacceptably unnatural results. Three remaining alternatives are

1. Supplying a pitch contour extracted from a naturally spoken version of the message. Data of this type would normally be used when word durations have been obtained in a similar manner and supplied externally. Pitch and timing information obtained in this manner are the most natural.

2. Calculating a pitch contour by rule, as discussed in the next section.

* Both software and hardware synthesizers are available with the DDP-516 computers employed.

FIGURE 17. Interpolation of formant contours.



3. Using an archetypal pitch contour. For limited-context messages, this pitch contour is modified (i.e., locally shortened or lengthened) to match the word durations as determined from the timing rules. There are obviously many ways in which the prosodic information for the message can be obtained and the choice of the foregoing alternatives depends strongly on the desired quality of the synthetic speech and the specific application in mind.

Once the timing pattern for the message is established, the segmental data in the word catalog must be altered to match the timing. This means that the formant data for a word in the catalog must be either lengthened or shortened. The formant contours in successive voiced words must also be merged to form smooth, continuous transitions.

The choice of place in a word to alter duration is based on the dynamics of the formant contours. For every 10-ms voiced interval of each word in the catalog, a measure of the rate of change of the formant contours is computed. This measure is called the "spectral derivative." Regions of the word in which the spectral derivative is small are regions wherein the word can be shortened or lengthened with the least effect on the word intelligibility. Thus to shorten a word by a given amount, an appropriate number of 10-ms intervals are deleted in the region of the smallest spectral derivative. To lengthen a word, the region of the lowest spectral derivative is lengthened by adding an appropriate number of 10-ms intervals. Unvoiced regions of words are never modified.

Whenever the end of one word is voiced and the beginning of the next word is also voiced, a smooth transition must be made from the formants of the first word to those of the second word. This transition is made over the last 100 ms of the first word and the first 100 ms of the second. The transition rate depends on the relative rates of spectrum change of the two words, over the merging region. Figure 17 gives examples of interpolations for typical shapes of formant curves. (In the figure it is assumed that, for both words 1 and 2, all three formants have identical shapes; only one formant is shown in the illustration.)

In case 1, word 1 has a very small spectrum change over its last 100 ms of voicing, but word 2 has a very large spectrum change. The interpolated curve shown at the bottom of the first column, though beginning at the formants of word 1, rapidly makes a transition and follows the formants of word 2. Case 2 shows the reverse of case 1; word 2 has little spectrum change, whereas word 1 has a large spectrum change. The interpolated curve now follows the formants of word 1 for most of the merging region, making the transition to the formants of word 2 at the end of the region. Cases 3 and 4 show examples where spectrum changes in both words are relatively the same. When they are small, as in case 3, the interpolated curve becomes essentially linear. When they are large (case 4), the interpolated curve tends to follow the formants of the first word for half the merging region, and the formants of the second word for the other half of the merging region.*

The final step in producing the message is to synthesize the speech using the chosen prosodic features and segmental features generated by the preceding rules. A hardware digital speech synthesizer performs this task in real time at a sampling rate of 10 kHz. Its control signals are updated pitch-synchronously from the interpolated stored data.

Figure 18 is a spectrographic illustration of the synthesis technique. At the top is a spectrogram of a natural utterance of "We were away a year ago." At the bottom is a spectrogram of the same sentence produced by abutting individually spoken, formant-synthesized words with no alteration of pitch or timing of the isolated words. The lack of continuity of formants is obvious. The resulting synthetic output is choppy and completely unacceptable. At the center of Fig. 18 is a spectrogram of the individual words concatenated according to the rules discussed previously. For this example the pitch and timing were obtained from the natural version of the utterance. It can be seen that the rule-interpolated formant functions of the middle spectrogram closely resemble the natural transition of the top spectrogram.* [Examples are on section 6 of the record.]

Figure 19 compares original and synthesized versions of the limited-context message "The number is 836-1246." Pitch and timing for the synthetic signal were calculated completely by rule. Timing for the complete message was derived from a stored table and is specific to sequences of 7-digit telephone numbers. An archetypal pitch contour was calculated and fitted to the timing of the word elements.† [Several examples of this type are given on section 7 of the record.]

Present success with rule-concatenated, word-length formant data makes it appear attractive for computer voice response. Simple rules for timing and pitch appear to suffice for certain limited-context applications. Broader application will depend upon the success of general prosodic analysis, as described in some detail in the next section.

Synthesis from printed text

Synthesis of speech from printed text offers virtually unlimited message capacity. The expense is the large storage capacity required for a pronouncing dictionary of word data.

A typical desk-size abridged dictionary lists 130 000 words. If the storage form is to be PCM or formant-encoded words, the dictionary must be expanded to include variations of most entries as pronounced with different common endings: plurals for nouns; *-s*, *-ed*, *-ing*, and *-er* for verbs; *-er*, *-est*, and *-ly* for adjectives; and as well, verbs with a number of prefixes, such as *re-*, *de-*, and *un-*, and nouns with invented word endings, such as *-ize*, *-ish*, and *-y*, and, as well, *-ized*, *-izes*, *-izing*, *-izer*, etc.

The dictionary would contain many infrequent words of long duration. Such words contain, on the average, 3½ syllables per word, one of which is generally stressed, and may be up to 1 second per word in length. If we assume that the dictionary contains from 500 000 to

* This interpolation function is a preliminary one, chosen for initial experiments. The ideal interpolation is likely to be context-dependent. This question is the subject of continuing research.

* Note particularly the way the sound "*a*" at about 1400 ms in the bottom spectrogram is smoothly interpolated into the formant tracks at about 1000 ms in the middle spectrogram.

† The ideal pitch contour, like the formant interpolation function, is likely to be context-dependent. The archetypal contour used here represents a preliminary function for experiment.

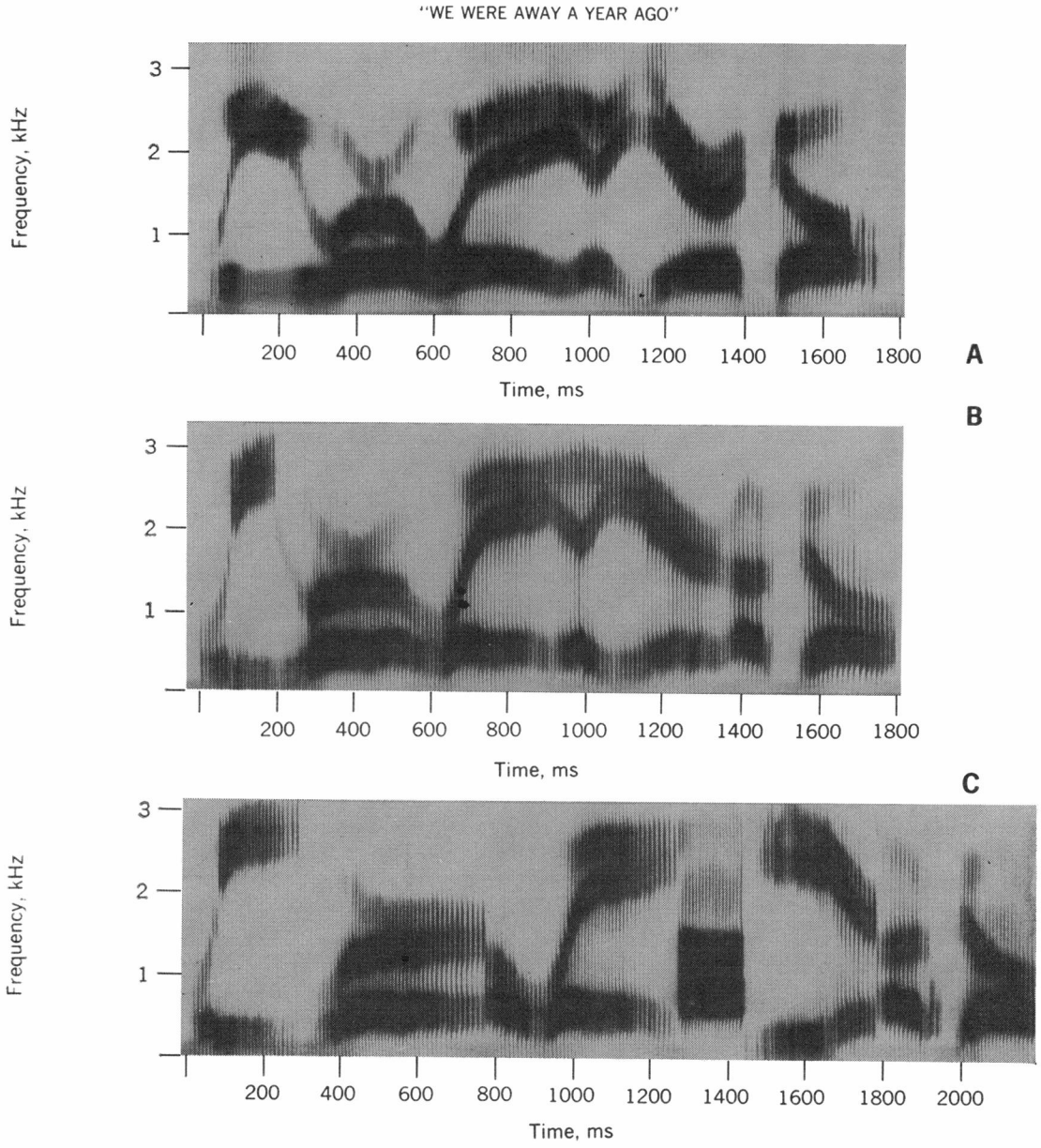
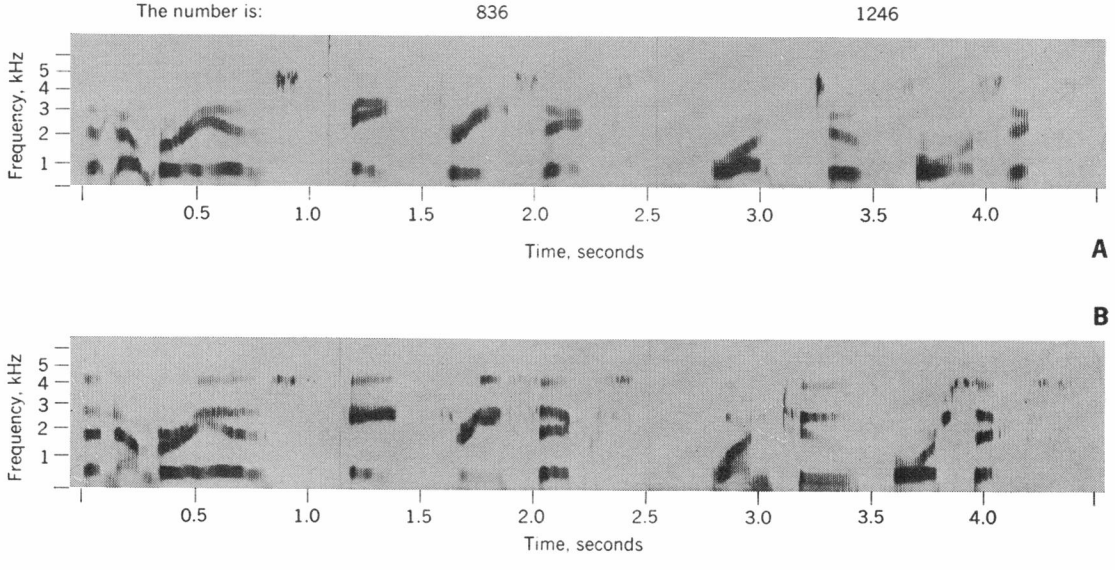


FIGURE 18. Sound spectrograms of (A) natural speech, (B) concatenated speech, and (C) isolated words.

FIGURE 19. Sound spectrograms of (A) concatenated digits and (B) naturally spoken digit string.



700 000 words including derived forms, we can project a PCM pronouncing dictionary to require around 100 hours of speech or, equivalently, 25 to 37 billion bits! With formant storage, the estimate drops to perhaps 500 million bits, which is still rather large but necessary for talking encyclopedias. A more compact form is obviously desirable; the most compact form is probably phonemic transcription.

Phonemic transcriptions of single-word entries may be encoded with about 72 bits per second (6 bits per phoneme; 12 phonemes and lexical stress marks per second). Thus the phonemic dictionary offers a 13:1 direct saving over formant storage and a 700:1 saving over PCM. In addition, it permits simple means for generation, rather than storage, of derived word forms, usually by simple concatenation of phonemes.* Using such schemes to derive words, rather than store them, the saving might approach 100:1 over formant data, or about 5000:1 over PCM.

We shall describe a working experimental system for synthesis from text, using a phoneme dictionary. The system differs from other work in two significant ways: (1) It contains fully automatic rules for conversion from English text (without special marks) to speech with reasonably natural timing and intonation. (2) It is a synthesis of unrestricted speech from a dynamic characterization of the human articulatory system.

A block diagram of the text to speech conversion program is shown in Fig. 20.† Blocks A through D convert from printed text to discrete symbols representing phonemes with detailed data for pitch and duration. Blocks E through I convert the discrete phonemic symbols to sequences of articulatory motion and thence to sound.

Articulatory synthesis. Figure 21 is a diagram of the vocal-tract model used for phoneme synthesis. Seven parameters are used to describe the cross-sectional area of the vocal tract as a function of distance along it. The parameters are: two coordinates each for the configuration of the lips (W, L), the position of the tongue tip (R, B), and the position of the tongue body (X, Y); and one coordinate for the position of the velum and uvula (N). The width of the pharynx and position of the teeth (jaw angle) are dependent variables inferred from the position of the tongue body.

The design of such a model seeks to satisfy two somewhat conflicting goals: (1) The model should be very general; that is, it should approximate well all of the vocal-tract shapes that occur in normal speech. (2) The model should be strongly constrained. It should exhibit such natural constraints of the vocal mechanism as continuity of the tongue surface, curvature of the vocal tract, and discontinuities of the tract at the teeth, velum, and esophagus. Most important, the model should incorporate the temporal constraints and dynamic behavior of the speech mechanism.

In the present model, there exists a good balance between constraints to exclude unnatural vocal-tract shapes, and flexibility to match the shapes that do occur in speech. Figure 22 shows comparisons between actual human vocal-tract shapes and those of the model,

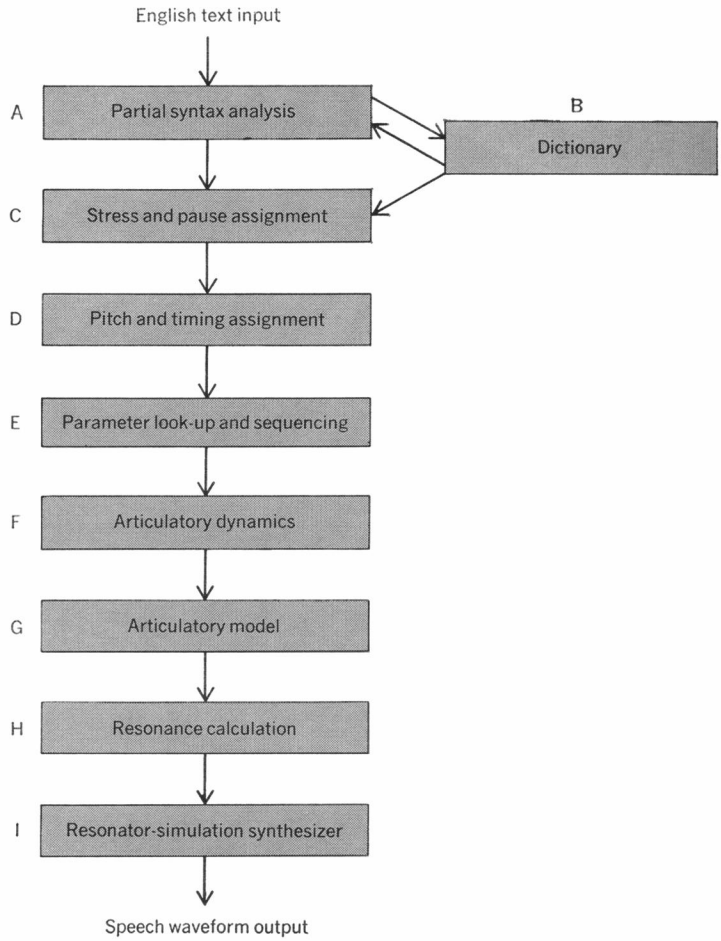
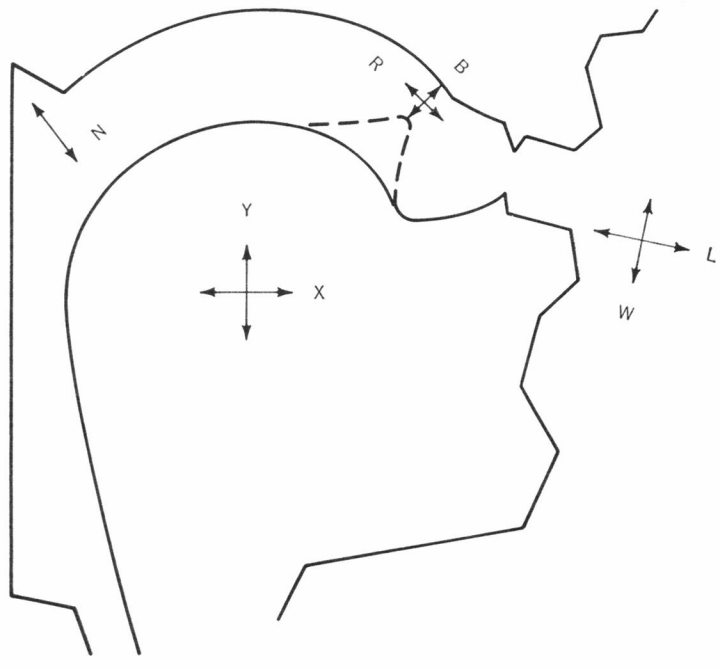


FIGURE 20. Block diagram of a complete system for synthesis of speech from English text. Blocks A through D convert from conventional text to detailed phonetic text. Blocks E through I achieve phoneme synthesis by modeling human articulation.

FIGURE 21. Schematic diagram of the computational articulatory model used in phoneme synthesis.



* This subject was studied extensively by Lee.[24]

† Work on this system began and developed on an older computer, a DDP-24. Continuing research is implemented on a DDP-516, described in Appendix B.

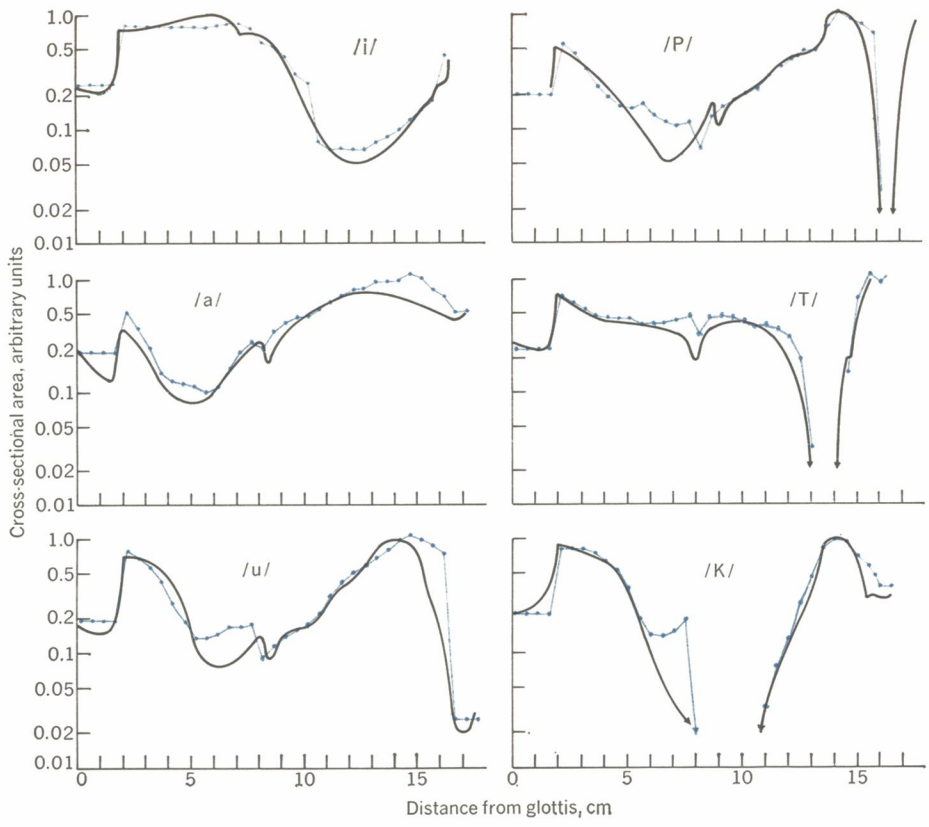
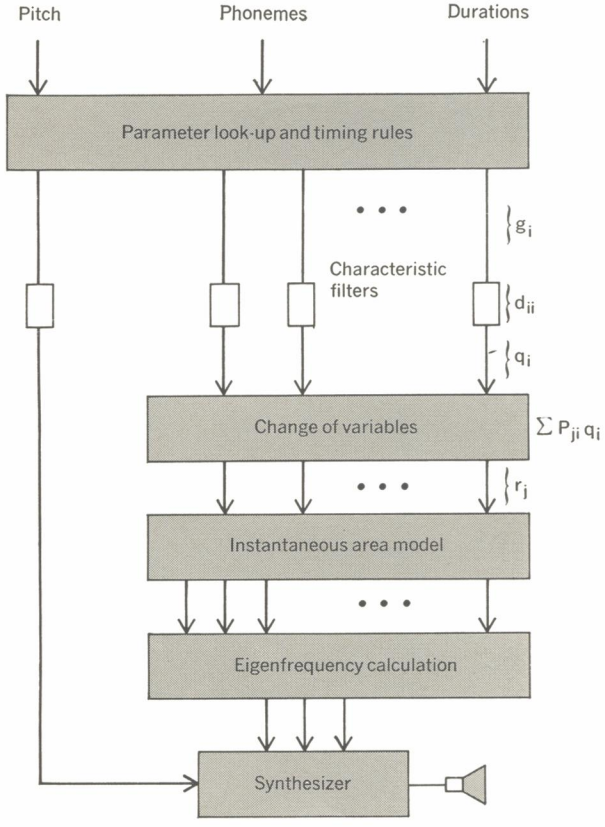


FIGURE 22. Examples of the ability of the articulatory model to fit human articulatory data as taken from X rays.

FIGURE 23. Block diagram of the synthesis system showing the method of supplying dynamics for the articulatory model. The variables g_i identify closely with "command variables" for independent rounding or closing the lips, motion of the tongue tip and tongue body. "Characteristic filters" simulate dynamics, and a change of variables accounts for relatively minor control interactions. (See Fig. 25.)



for a representative group of speech sounds. The solid lines are plots of cross-sectional area against distance along the vocal tract. The data are derived from multiple X-ray photographs of a human. The dotted points are corresponding functions computed from the vocal-tract model.

An important property of the model is its ability to represent accurately the shape of the vocal tract where area is small. It provides specific coordinates for control of place and degree of constriction in most of the consonants of English. This leads to a direct way of establishing the dynamic behavior of the model or, at least, of optimizing the dynamics of each part of the model for the specific phonemes in which that part is most significant. For example, control of lip protrusion (L) is optimized for the rounded vowels and the sound /w/; control of lip opening (W) is chosen for good reproduction of the consonants /p/, /b/, and /m/.

Figure 23 shows how the model is used to synthesize connected speech from sequences of discrete symbols. The computation of the time-dependence of the model coordinates is logically divisible into two separate parts: the sequencing of "position commands" to the model, and the simulation of responses to these various commands.

Basic speech sound elements of the synthesizer vocabulary, most of them corresponding to conventional phonemes, are stored for the model as sets of "idealized position coordinates" or target positions. These are presented as sequences of step functions (g_i) to a set of "characteristic filters" (d_{ii}), which simulate the dynamics of the system. These filters have second-order overdamped responses, with rise times ranging from 50 to 200 ms.

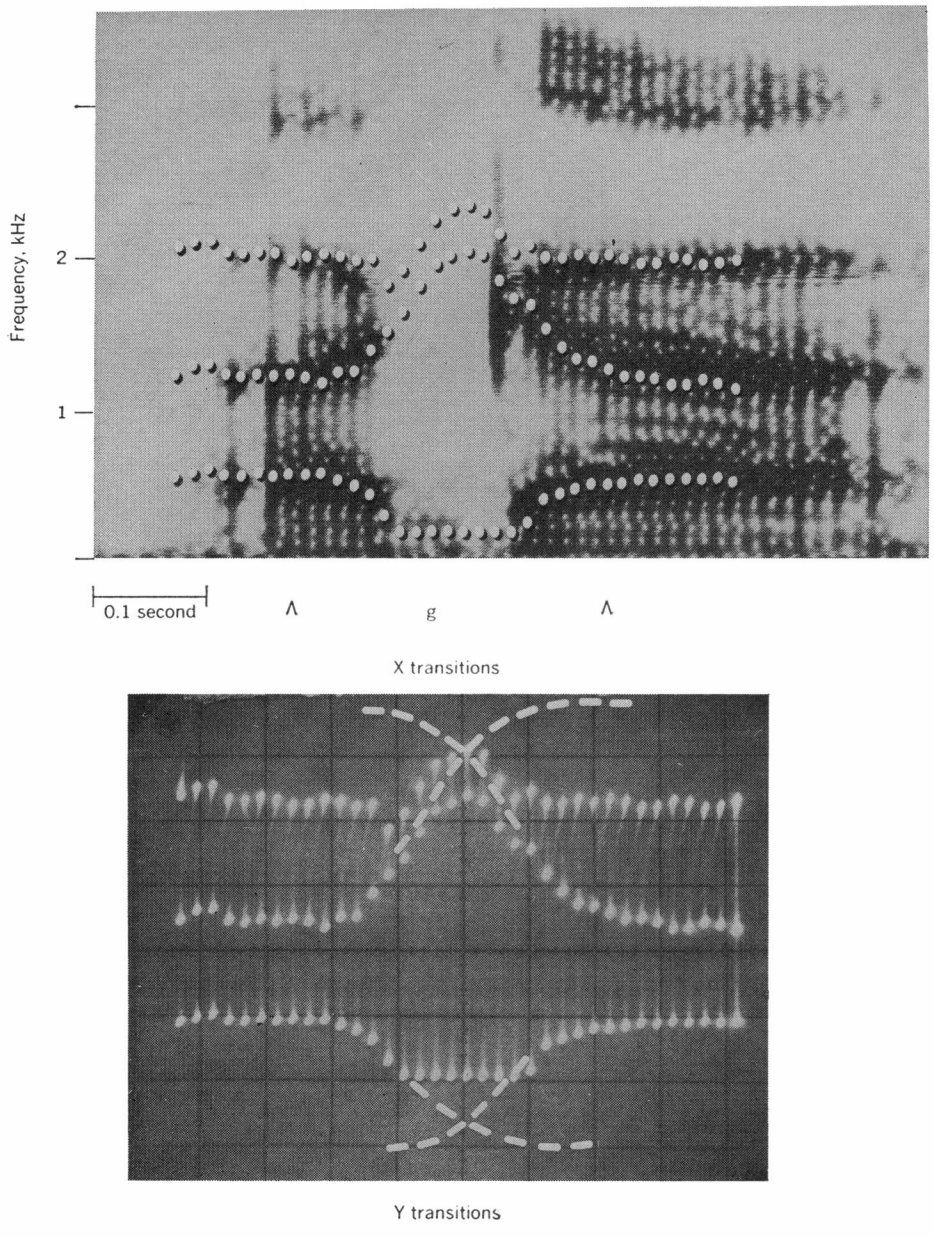


FIGURE 24. Comparisons of formants calculated from the articulatory model with a spectrogram of natural speech for the same utterance. Dots superimposed on the spectrogram were traced by the computer scope display. Sketched on the lower photograph are transitions in control variables to the articulatory model.

The responses for commands for lip rounding, velar motion, and front-to-back motion of the tongue tip are slowest; those for simple lip closure and raising of the tongue tip are fastest. The outputs q_i of the "characteristic filters" are converted to the position coordinates of the model (r_j) by a transformation of variables. The transformation accounts for components of tongue tip and lip motion due to tongue-body motion, and for a component of lip closure due to lip rounding.

The position coordinates r_j specify the tract area function at 10-ms intervals of time. These coordinates are applied to the computational model to derive sets of data representing area as a function of distance along the tract. These area functions are incorporated into a one-dimensional lossless solution of Webster's horn equation. The equation is iterated to solve for the first three eigenfrequencies of the tract for each 10-ms interval of time. The data at this point refer to the variables of formant synthesis. These computed formant frequencies are outputted to a hardware synthesizer, of the type shown in Fig. 11, to generate the sound.

Figure 24 is a comparison of a natural-speech utterance of a nonsense syllable */igagi/* (*eégahgee*) and the formant frequencies for a similar synthetic sequence computed from the vocal-tract model. The figure also illustrates the basic approach to establishing dynamic response of the model. The response of the lowest resonance is found to depend primarily on the control variable dominating the size of the constriction for the consonant */g/*; the transients in the second and third formants are found to depend mainly on the parameter affecting the place of the constriction. Similar relationships are found for parameters controlling the lips and tongue tip in other consonants.

Besides the dynamics of articulatory motion, another important feature is incorporated in control of the model. The fixed-target, step-function method of control matches real speech only if some asynchrony is introduced into

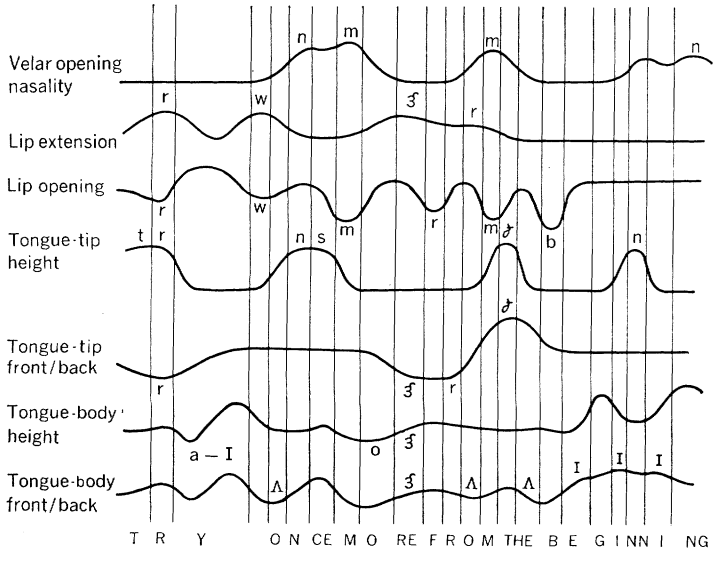
the timing of different articulatory parameters. Consider, for example, the transition from a vowel to a lip or tongue-tip consonant. The specific gesture of the lips or tongue tip to form the consonant occurs primarily before any significant motion of the tongue body. In a transition between two consonants, the specific gesture for each consonant overlaps that for the other. The result is that a constriction for one or the other of the consonants is consistently maintained. At the "midpoint" of the transition, the significant gestures of both consonants are almost fully articulated.

This phenomenon is incorporated in the model by a priority strategy. Each phoneme has a table of factors designating the relative importance of each articulatory parameter to the formation of the phoneme. The timing of step-function changes in the control variables g_i is staggered according to comparisons of priorities between adjacent phonemes. Transitions to a critical value therefore occur early, and those away from a critical value occur late.

Figure 25 illustrates results of this process. The plotted data are the filtered variables q_i, before the change of variables to actual model coordinates r_j. Transitions in the nasality, tongue, and lip variables are caused to occur primarily during the neighboring phonemes. Nasality, for example, is unimportant for the phoneme /s/ in "once more." The action of the priority algorithm allows the velum target positions of the preceding phoneme /n/ and following phoneme /m/ to dominate during /s/. A similar example occurs for the tongue-tip front/back control in the sequence "more from." In the tongue-tip front/back control for "the beginning" and in the lip extension control for "once," the control transitions can be seen to extend across not one but two adjacent phonemes.

Although the action of the model must be considered only approximate to that of a real vocal tract, Fig. 25 illustrates a rather remarkable property of speech. Discrete symbols (phonemes) are transmitted through a multivariable articulatory system whose parts are too sluggish individually to resolve discrete values at the natural rate of phoneme production. The sequential constraints of the language and the control strategies of articulation, however, allow sufficient time for each individual articulator to reach its goal *when it becomes necessary*.

FIGURE 25. Outputs of "characteristic filters" for an example utterance. The variables B and R for the articulatory model are B′ and R′ above, plus components of tongue-body position, X and Y. Actual lip opening W is the sum of W′ above, a component of lip extension L and components of X and Y representing jaw angle.



Section 8 of the recording is an example of the capabilities of the articulatory synthesizer when controlled with hand-supplied data. For this example, discrete phonemes with additional symbols for timing and pitch control were fed into the computer by typewriter. Timing and pitch were selected by repeated listening and retyping as necessary to make the synthesized sound more natural. Experiments of this type have been the primary source of data upon which rules for the prosodic features of English have been developed. Additional resources include the dynamic matching of X-ray motion pictures and the visual matching of sound spectrograms.[25]

Conversion of text to detailed phonetic transcription. The articulatory model just described (blocks E through I of Fig. 20) requires discrete phonetic symbols and pitch and duration data to effect synthesis. Printed alphabetic text must therefore be transformed into this phonetic form (blocks A through D of Fig. 20).

Automatic conversion of printed English text into discrete phonetic symbols must provide sufficient information about the prosodic features of speech to effect a natural-sounding synthesis. A human speaker never gives the same importance to every word. Some words are made prominent by giving them higher pitch, increased intensity, and lengthened duration. Some words are so reduced (shortened) that phonemes become very short and weak, and may even be dropped completely. Pauses may be inserted in a sentence at places not marked by punctuation.

Recently, with the development of digital computers, several attempts have been made to program systems for the automatic conversion of printed text to synthetic speech.* These efforts depend upon an analysis of the surface structure of sentences to derive information about grammatical boundaries and about the patterns of stress within sentences.

The first part of the system described here (blocks A through C of Fig. 20) generates the minimum information for pause and stress assignment. The remaining part of the system (block D) generates pitch and timing assignments for each phoneme and produces discrete symbols that are used directly as the input to the articulatory model described in the previous section. The rules for pitch and timing assignment, however, are also applicable to the formant synthesis technique mentioned previously, as well as to similar synthesis systems.

Syntax analyzer and dictionary. The dictionary (block B, Fig. 20) provides a phonemic transcription of each word with lexical stress marks, and the possible usages

* Holmes, Mattingly, and Shearme developed semiautomatic rules in which the phonemic transcriptions and stress and pause assignments were made by a human operator.[9,10] A fully automatic system for accomplishing the conversion was first developed by Teranishi and Umeda.[26] The syntax analyzer reported here is an outgrowth of the latter work. Treatments of stress and pause assignment were made by Vanderslice.[27] Also, the work of Lee[24] and Allen[28] is contemporary with the work reported here.

of the word, such as noun, adjective, or verb. It also provides a content–function distinction.* Content words convey substantial meaning in the sentence. They relate to things, actions, or attributes. Function words, usually monosyllabic, serve mainly to establish grammatical relationships. Function words include articles, prepositions, conjunctions, and personal pronouns. There are also "intermediate" words, which are neither function nor content. Polysyllabic, less-frequent prepositions and conjunctions, and frequently used verbs (such as "get," "take," "give," etc.) fall into this category, as do some pronouns, adverbs, and adjectives.

The role of the syntax analyzer (shown as block A in Fig. 20) is multiple. For each sentence it must (1) assign the probability of a break (or more exactly, the potential of a pause) at every grammatical boundary (such as a phrase or clause), (2) select alternative pronunciation of certain words according to their usage, and (3) change the function–content distinction for specific conditions to assign proper stress to important words in the sentence (e.g., an auxiliary verb at the end of a clause has to be changed to a content word). Using the information on word class contained in the dictionary, the analyzer groups words into phrases. For each word, it assigns a phrase category such as introductory modifier, subject, verb, object (or complement), or tail modifier, according to the possible order of occurrence of the phrase categories. The potential of a pause or of a weaker break is assigned between all words. Words in the same phrase have zero probability of break. Break probability is higher between subject and predicate than between verb and object. Break probability is relatively high between an introductory prepositional phrase and the subject. Any reverse order of occurrence among phrase categories indicates a clause boundary at the reverse point. Clause boundaries have higher probability of a break than any phrase boundary within the clause. Punctuation marks require the highest probabilities of a break.

Stress-pause assignment. Using the information on probabilities of a break obtained in the syntax analyzer, the program element of this stage (block C, Fig. 20) decides what kind of break and pitch contour is to be assigned at the end of the grammatical unit. The threshold for putting an actual pause or pseudo-pause† in the synthetic speech shifts depending on the length of the sentence, and on the speech rate. The unit separated by a pause or break with one focus stress will be called a "pause group." Full stops indicate the longest possible pause and have a falling pitch contour at the end. Commas indicate a pause and a rising pitch contour, implying continuation. For deliberate speech and for sentences of reasonable size, clause boundaries without punctuation are terminated without pause and with rising pitch. Phrase boundaries with higher probability of a break are accompanied by the elongation of the last phoneme in the phrase.

Stress levels are assigned to words mainly according to the content-function distinction. The stress levels are used in computing pitch and duration information. The stress assignments are:

1. No stress: function words
2. Weak stress: intermediate words
3. Primary stress: content words
4. Focus stress: focus of the sentence☆

A typical result of the entire conversion of printed text to discrete symbols for the articulatory synthesizer input is shown in Fig. 26. In the left column the input English text is shown word by word. (In this case, the text is the first line of Aesop's fable of the North Wind and the Sun.) Column 2, labeled *cat*, represents the phrase category of each word. Column 3, labeled *wc*, gives word class. Column 4, labeled *pp*, represents the probability of break. Column 5, labeled *cf*, makes a content–function distinction in the role of each word. Column 6, labeled *ic*, indicates the shape of the pitch contour for each word. Column 7, the rightmost column, shows the form of the input information to the synthesizer. These symbols are the result of pitch and timing assignments applied to data determined from the syntax analysis, dictionary look-up, and stress-pause assignment. These output symbols are described next.

Pitch-timing assignment. From the previously derived information, timing-control marks and pitch marks are assigned to each phoneme. In the rightmost column in Fig. 26, numerals and minus signs are timing controls, and the special marks *, $, and & are pitch controls that increase voice pitch from a nominal value. *q* is a pitch mark that lowers the pitch in a prescribed decrement. All alphabetic characters, except *q*, specify English phonemes. The symbols 〈 and 〉 indicate the second part of diphthongs—front and back, respectively.

The program has a timing table for all phonemes. Each phoneme has a fixed minimum duration and an additive variable duration. Numbers for timing control represent the duration in terms of the sums of the fixed portion and some multiple of the variable portion. Duration values are specific to individual phonemes. A given number can therefore play the same role in prosodic rules independent of absolute duration.

Two major rules are used to determine the timing and pitch controls: (1) a word boundary rule, which determines consonant durations, and (2) a stress and termination rule, which determines pitch marks and vowel durations.

Word boundary. Consonant durations at word boundaries are adjusted according to the relations of Fig. 27. Function words are merged together (like unstressed syllables inside a polysyllabic word) and take minimum duration of consonants at their boundaries (0-boundary). Average durations are assigned to consonants at the boundaries of words with intermediate stress (1-boundary). Content words always receive long consonants at the boundary so that they might be prominently separated in the stream of the utterance. Content-word boundaries (2-boundary) produce consonants about 24 ms longer than average (seen as the numeral 6 in Fig. 26). Intermediate boundaries (seen as 1-boundary in Fig. 27 and as the numeral 4 in the initial and/or final consonant in Fig. 26) produce consonants of average duration, and occur between intermediate and compound content words. Function word boundaries (0-boundary in Fig.

* This distinction is given intuitively by Pike.[29]

† A kind of break with rising pitch contour for termination without an actual pause following.

☆ In actual speech, any word could be the focus. In our system, however, the last content word in the pause group is assigned focus stress, unless the focus word is especially marked in the input text.

```
English
 Text          cat     wc       pp  cf  ic
the            s       tce      0           4dh 4a
north          s       a        0   ++  -   6n $4aw 2er 6th
wind           s       n        0   ++  -   6w *qq5i 4n 4d
and            s       cla      0           4aa -n -d
the            s       tce      0           -dh 4a
sun            s       n        0   ++  -   6s *qq5uh 6n
were           v       vbp      5           4w 4er
arguing        v       vg       0   ++  -   4: $q6ah -r -g -y 4uu 4i 6ng
one            o       aq       2   +       4w &5uh 4n
day            o       n        0   ++  /   6d *q9ay qq9<
,              p       comm     11  ++  /   $,
when           i       whn      0   +       2h 2w &5eh 4n
a              s       tca      6           4a
traveler       s       n        0   ++  *   4t 4tr *q7aa -v 4o -l 4er
came           v       vp0      5   +       4k &4ay 4< 4m
along          t       p2       6   +   /   4a 41 8aw 4ng
,              p       comm     11  ++  /   $,
wrapped        v       vp0      0   ++  -   6r $q8aa 4p 4t
in             t       p1       6           4i -n
a              t       tca      0           4a
warm           t       a        0   ++  -   6w $5ah 2er 6m
coat           t       n        0   ++      6k *q2oh qq20h 6t
               p       peri     12  ++
```

FIGURE 26. Printout of a program that converts from English text to discrete phonetic symbols. The leftmost column is the input. The columns cat (phrase category) and wc (word class or "part of speech") are internal decisions of the syntax analyzer; the columns pp (pause probability) and cf (content/functional distinction) are its output. The column ic (intonation contour) describes the stress pattern assignments: rising, sustained, or falling pitch. Data in the rightmost column, phonemic symbols, pitch, and timing, control the articulatory synthesizer to produce speech.

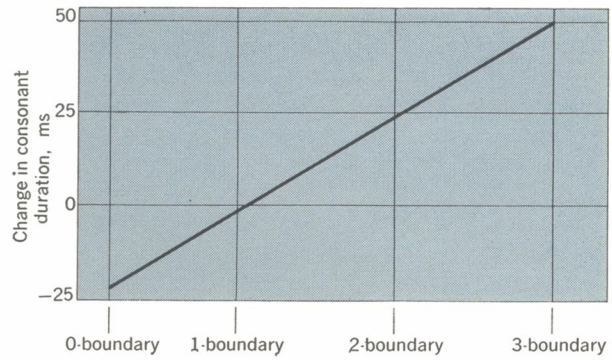


FIGURE 27. Consonant duration in terms of degree of word separation. 0-boundary is a weak separation typical of the juncture of two monosyllabic functional words. The 2-boundary is typical of the juncture of two important content words.

27) produce the minimum consonant duration at the boundary, about 18 ms shorter than average (shown as minus sign in Fig. 26). A more prominent boundary (3-boundary, Fig. 27) is applied to a phrase boundary where greater separation is needed.

Inside polysyllabic words the initial consonant (or consonant cluster) of a stressed syllable is assigned 1-boundary duration, and any other consonant gets 0-boundary duration.

Stress and termination. In accordance with the stress levels assigned to words, pitch marks are put on the stressed vowels as follows:

No stress (every unstressed vowel)		no mark
Weak stress		&
Primary stress:	reduced (verb)	$q
	normal (adj. and adv.)	$
	high (noun)	*qq
Focus stress		*q

The higher the pitch, the longer the vowel is made.

Two kinds of termination are assumed. Termination is used here to denote the pitch contour used to terminate a pause group.

1. Falling pitch: sentence end (unless yes–no question)
2. Rising pitch: middle of the sentence, followed by a comma, or at a point at which the break probability is high

Combinations of stress and termination assignment, used with the phonetic context, provide a range of timing control for each vowel. The continuation pitch rise forces a vowel to be quite elongated. Focus stress also produces a

long vowel. When these two factors fall on the same vowel, it is forced to have a steep falling pitch followed by a rising pitch; consequently the vowel becomes very long. Voiced consonants are also a factor in elongating a preceding vowel. When all three factors occur on the same vowel at the end of the pause group, the vowel is lengthened greatly, typically three times as long as its normal duration in ordinary context.

Figure 28 shows duration modification of vowels. Each line represents a different phonetic condition of stress, with or without termination. Different lines represent different conditions of stress and termination as a function of the phonetic contexts. Figure 28 is specific for the vowels /*a*/ and /*c*/. Other vowels have different slopes and ranges of variation for their duration increments. The pitch and timing assignments complete the information derived by the text-conversion program. The resulting data of the rightmost column in Fig. 26 are then supplied to the articulatory model and converted to connected speech. A spectrogram of the output synthetic speech is shown in Fig. 29. For comparison, a spectrogram of a corresponding natural utterance is also shown. [The reader is directed finally to sections 9 and 10 of the recording. These passages demonstrate speech synthesized automatically from printed-text input.]

Summary

The methods of speech synthesis from stored formant data and from printed text have advanced to a promising point. Their potential for automatic information services and for computer voice response appears good. At this point in time the quality of formant-synthesized speech is generally better than that of text synthesis. In the former the so-called segmental information is derived from naturally spoken speech. The suprasegmental (prosodic) information is calculated either by rule or from stored measurements on real speech. The saving in required storage is approximately 50:1 compared with digitized speech waveform. In text synthesis, both segmental and suprasegmental information are calculated, and no shred of naturally derived data is relied on. The economy in storage is concomitantly greater, being of the order of 700:1 compared with the digitized speech waveform.

Both techniques appear to have places in the range

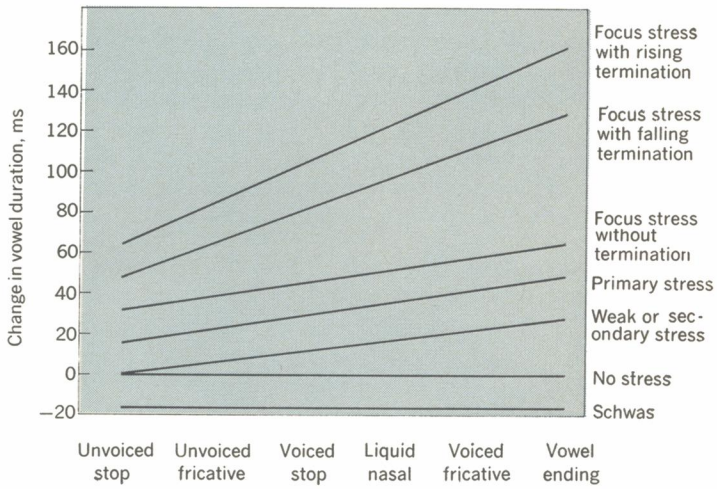
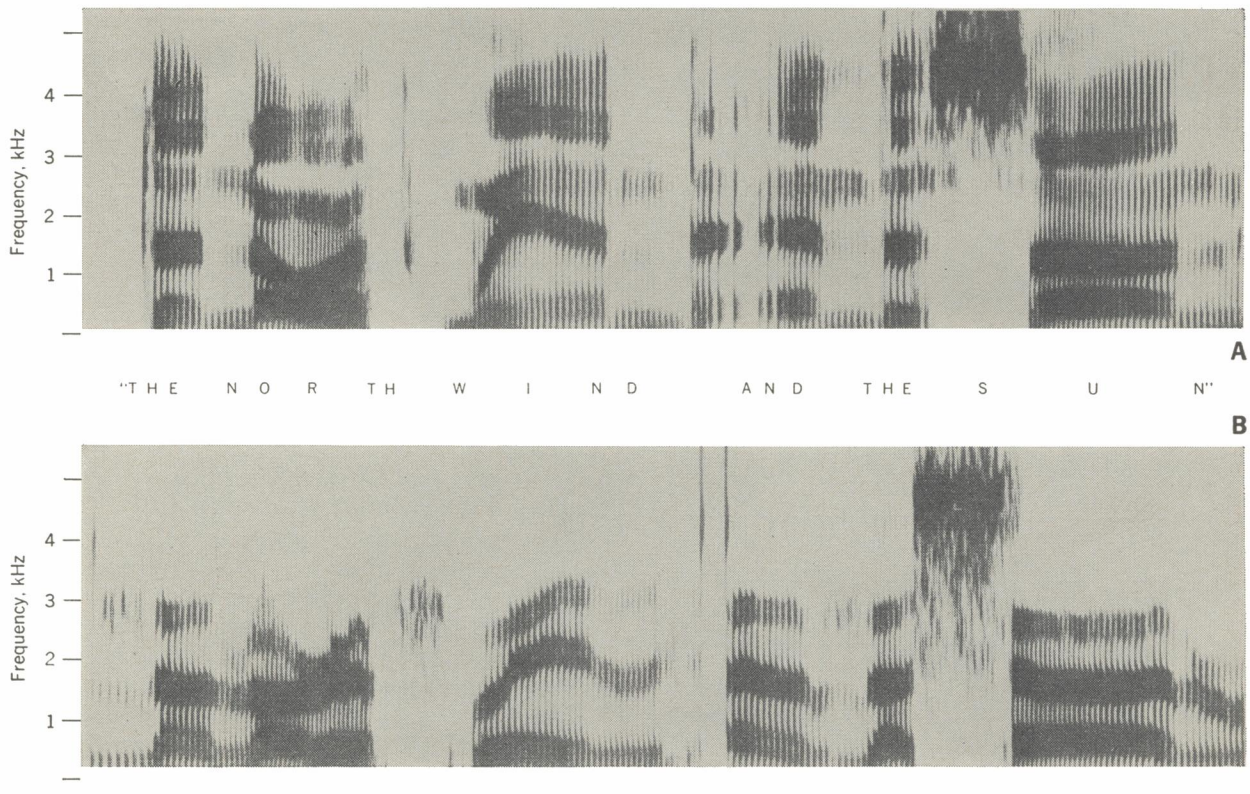


FIGURE 28. Vowel duration as a function of stress and phonetic context. The term "focus stress" indicates the major stressed word in a pause group.

FIGURE 29. Spectrogram of (A) natural speech and (B) speech synthesized from phoneme rules, for the same sentence.



of computer answer-back applications envisioned. Formant synthesis is a natural for cases requiring moderately large vocabularies of high information content (low redundancy) and good intelligibility. Typical applications might be automatic intercept and information operators, inventory reporting, weather forecasts, and stock market quotations. In contrast, automatic information services such as the talking encyclopedia would clearly depend upon text synthesis, as would uses such as reading machines for the blind and certain computer-based teaching techniques.

The synthesis methods described in this article are being researched using two laboratory computers especially tailored for this type of acoustic problem. Questions of feasibility in computer implementation therefore constitute an integral part of the work. Computation times, access times, and methods of computer control of digital hardware for synthesis are all relevant questions. For the reader interested in the computer techniques being used, an outline description of the computer facilities we have constructed is given in Appendix B.

Appendix A
Description of record

The recording mentioned in this article is an 8-mil-thick, 7-inch, $33\frac{1}{3}$-r/min microgroove disk containing the demonstration items listed below.

To purchase the record, please complete the address label below and forward it with 50 cents in coin to IEEE, RC Unit, 345 East 47 Street, New York, N.Y. 10017.

Side 1—Formant analysis and synthesis

1. Automatic analysis and synthesis—3 utterances: synthetic—original—synthetic
2. Smoothing of pitch and formant controls—1 utterance, 4 smoothing filters; 16-Hz, 12-Hz, 8-Hz, 4-Hz low-pass cutoff
3. Quantization and smoothing of pitch and formant controls—1 utterance, 16-Hz cutoff low-pass filter used on all versions
   a. Pitch quantizing: 7, 4, 3, 2, 1 bits
   b. Formant 1 quantizing: 4, 3, 2, 1 bits
   c. Formant 2 quantizing: 4, 3, 2, 1 bits
   d. Formant 3 quantizing: 3, 2, 1 bits
4. Bit rate comparison—1 utterance

	4600 b/s	600 b/s
Pitch	7 bits	6 bits
F_1	10 bits	3 bits
F_2	11 bits	4 bits
F_3	11 bits	3 bits
A_V	7 bits	2 bits
Sampling rate	100 per second	32 per second

   4600 b/s—original—4600 b/s—600 b/s
5. Manipulation of pitch, timing, and formant controls—1 utterance
   a. Pitch manipulations
      (1) Synthetic unmanipulated
      (2) Monotone pitch—100 Hz
      (3) Pitch doubled
      (4) Pitch squared
   b. Timing manipulations
      (1) Synthetic unmanipulated
      (2) Vowels in "we" and "year" lengthened by 100 ms
      (3) Synthetic unmanipulated
      (4) Total duration = 75% natural duration
      (5) Total duration = 50% natural duration
      (6) Synthetic unmanipulated
      (7) Total duration = 150% natural duration
      (8) Total duration = 200% natural duration
   c. Formant manipulations
      (1) Synthetic unmanipulated
      (2) Formants raised by 10%, pitch raised
      (3) Formants raised by 20%, pitch raised
      (4) Formants raised by 30%, pitch raised by 50%
      (5) Synthetic unmanipulated
      (6) Formants lowered by 10%
      (7) Formants lowered by 20%
6. Concatenation of words
   a. Isolated words in sequence
   b. Words concatenated by rule; natural pitch and timing from speaker 1
   c. Words concatenated by rule; natural pitch and timing from speaker 2
   d. Words concatenated by rule; natural pitch and timing from speaker 3
7. Concatenated digit strings—four comparisons of strings of isolated digits followed by concatenated digits

Side 2—Synthesis from printed text

8. Articulatory synthesis from manual phonetic input
9. Automatic synthesis from printed text, "Parable of the North Wind and the Sun"
10. Synthesis from printed text

IEEE
Attention: RC Unit
345 East 47 Street
New York, N.Y. 10017

Please send me a copy of the synthesis demonstration record. I enclose 50 cents in coin.

Name

Address

City State Zip Code

Appendix B
Laboratory computer facility for interactive studies of speech analysis and synthesis

The speech research described in this article is being carried out on an interactive laboratory computer especially configured for problems in acoustic signal processing. Because some of the capabilities are unique, engineering details are outlined for the reader interested in computer implementation.

The facility employed in these investigations includes two Honeywell DDP-516s. The machines communicate with each other and with a central GE-635 computer via

data-phone connections. The two machines and their software systems are identical; programs are completely interchangeable. One machine is normally dedicated to problems in speech analysis and synthesis and digital filtering, and it typically serves nine research staff members. The second machine is presently dedicated to perceptual experiments on synthetic speech, auditory acuity, and acoustic signal processing. It serves about eight staff members.

The Honeywell DDP-516 is an integrated circuit machine with a 0.96-μs cycle time and 16-bit word length. As shown in Fig. 30, our configurations include 16 k of core memory, hardware multiply and divide, direct multiplex control (DMC) with 16 data channels (0.25 MHz each), and direct memory access (DMA) channel (1.0 MHz). An ASR-33 teletypewriter is standard (and was the only peripheral delivered with the machine). Fortran IV compiler, DAP-16 machine-language assembler, math libraries, and various utility software are supplied by the manufacturer.

For our range of problems we have interfaced the peripherals shown in Fig. 30:

1. Two fixed-head disks for each machine. Each disk provides 394 k words of storage with a 33-ms maximum access time and 180-kHz word transfer rate.

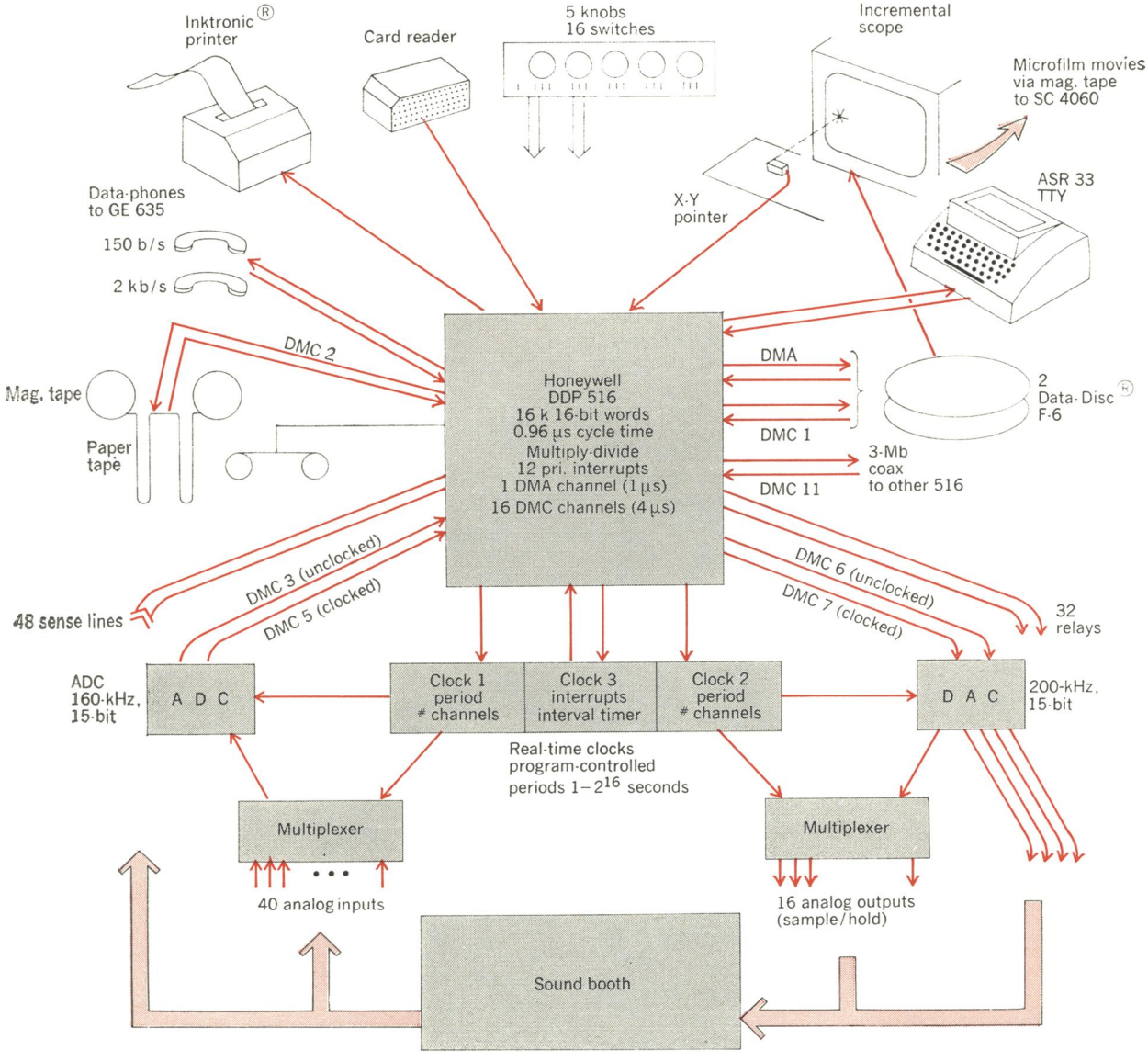
2. D/A converters. Four independent channels with program-controlled clocks for rates from dc to 180 kHz; 16 analog sample/hold channels with practical sampling rates from 50 Hz to 20 kHz.

3. A/D converters. 180-kHz 15-bit A/D converter; 40-channel 100-kHz multiplexer. The converter can operate through two data channels at the same time, sampling one group of channels synchronously at rates set by a program-controllable clock, and a different group asynchronously under direct program control.

4. Display scope. An incremental scope driven from a track on the disk, not from memory. The scope can reproduce 24 000 dots at an 825-kHz rate, thus providing a full page of text or about 10 meters of almost continuous line. An *x-y* pointer gives graphic input and manipulation.

5. Card reader. A 300-per-minute card reader is in-

FIGURE 30. Laboratory computer facility for speech analysis and synthesis.



corporated into the console. A keypunch is located beside the console. Card equipment and character set are compatible with the central GE-635 computer, and with nearby card editing and off-line listing facilities.

6. Inktronic® 120-character-per-second printer. Software allows the user to scan a listing with the display scope. With the press of a switch selected portions of the listing (such as the data-location map) are printed at a rate of 2 to 5 pages per minute.

7. Paper tape reader (150 characters per second). Used as a backup and for compatibility with manufacturer checkout software.

8. Magnetic tape drive (64 k characters per second). This nine-track drive provides selection of read and write densities. Tapes are compatible with nine-track drives on the central GE-635.

9. Three 16-bit input registers and two 16-bit output registers are used to control external electronic equipment.

10. A double-wall sound booth is installed adjacent to each machine for high-quality recording and for listening tests.

The operating system is based on the fast disk store, and the system programs reside permanently in 48-k write-protected words of the lower disk (⅛ of the total disk capacity). Users may write on all remaining disk and normally retain no permanent disk storage. Walk-away storage is through magnetic tape. Compiled programs may be conveniently saved on tape. Compiler, assembler, linking relocating loader, and all libraries are immediately available from permanent disks. Loading a Fortran source deck, compiling, listing on scope and/or line printer, loading compilation, and library all proceed expeditiously, with the disk serving as the storage medium at every step.

The scope system allows high-resolution line displays, and printing via Fortran-formatted I/O statements. The display is unique in that it provides a continually replenished display from the disk. The display consumes no core space and is operative even when the computer contains a different program or is halted. A feature of the display program causes, at the programmer's option, all display data to be echoed onto magnetic tape. A companion program in the local central computer allows regeneration of identical display frames via Xerox quick hard copy and SC4060 microfilm or 16-mm movies.

A display utility allows paging or scrolling through assembly and compilation listings (automatically written on disk), octal and decimal display of disk and core, quick hard copy of any printed display via the Inktronic printer or magnetic tape for microfilm, waveform display of disk or core, and simultaneous audio output.

A simple overlay procedure obviates serious limitations of the 16-k memory size. Through a subroutine call, a program can save itself onto disk, bring in another program, and transfer to it in about 0.1 second. Overlays may be nested to any depth and returns are made successively as with simple subroutine calls. A breakpoint and debugging package links the general display program as an overlay. A user can quickly alternate between running his program; looking at or listening to data on disk or inside his program; repositioning or printing selected portions of his listing; restoring his program to core, while keeping a page of listing on the scope; altering locations in his program or data; restoring registers; and running to another breakpoint or looping through the same one.

In evolving these laboratory systems, we followed a philosophy which to us seems important. Engineering of peripherals was done on a schedule that permitted maximum use of the machines at any stage of development. The first peripherals added were D/A converters. These required trivial effort and immediately allowed us to carry out three planned research projects. At the same time work on disks and tape proceeded. As a result it was possible to complete three separate research problems during the first year of operation of the first machine. By adding peripherals only when we were certain they were what we needed and would work, we insured that the facility was always a low-risk investment. Expenditures at each stage have been accompanied by productive output. This strategy also minimizes vulnerability to manufacturer delays or to new software difficulties. Considering the rate of depreciation of computer equipment, the overhead for space and support, this kind of planning, we feel, is valuable in obtaining the maximum return per computer dollar invested.

Our two DDP-516 systems are being applied to a range of problems in acoustic signal processing. Typical projects, besides the speech analysis and synthesis discussed here, include speech quality studies,[23,30,31] auditory detection,[32] vocal-cord modeling,[33] adaptive delta modulation, deconvolution of acoustic reverberation, and interactive design of digital filters.

REFERENCES

1. Flanagan, J. L., *Speech Analysis, Synthesis, and Perception.* New York: Academic Press, 1965.

2. Fant, C. G. M., *Acoustic Theory of Speech Production.* The Hague: Mouton and Co., 1960.

3. Flanagan, J. L., and Landgraf, L. L., "Self-oscillating source for vocal-tract synthesizers," *IEEE Trans. Audio and Electroacoustics*, vol. AU-16, pp. 57–64, Mar. 1968.

4. Chapman, W. D., "Prospectives in voice response from computers," *Proc. Internatl. Conf. Commun.*, 1970.

5. Dixon, N. R., and Maxey, H. D., "Terminal analog synthesis of continuous speech using the diphone method of segment assembly," *IEEE Trans. Audio and Electroacoustics*, vol. AU-16, pp. 40–50, Mar. 1968.

6. Rabiner, L. R., "Speech synthesis by rule: an acoustic domain approach," *Bell System Tech. J.*, vol. 47, pp. 17–37, Jan. 1968.

7. Liberman, A. M., Ingemann, F., Lisker, L., DeLattre, P., and Cooper, F. S., "Minimal rules for synthesizing speech," *J. Acoust. Soc. Am.*, vol. 31, pp. 1490–1499, 1959.

8. Kelly, J. L., Jr., and Gerstman, L. J., "An artificial talker driven from a phonetic input," *J. Acoust. Soc. Am.*, vol. 33, p. 835 (A), 1961.

9. Holmes, J. N., Mattingly, I. G., and Shearme, J. N., "Speech synthesis by rule," *Language and Speech*, vol. 7, pt. 3, pp. 127–143, July–Sept. 1964.

10. Mattingly, I. G., "Synthesis by rule of prosodic features," *Language and Speech*, vol. 9, pt. 1, pp. 1–13, Jan.–Mar. 1966.

11. Cooper, F. S., Gaitenby, J. H., Mattingly, I. G., and Umeda, N., "Reading aids for the blind: a special case of machine-to-man communication," *IEEE Trans. Audio and Electroacoustics*, vol. AU-17, pp. 266–270, Dec. 1969.

12. Schafer, R. W., and Rabiner, L. R., "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Am.*, vol. 47, pt. 2, pp. 634–648, Feb. 1970.

13. Bogert, B. P., Healy, M. J. R., and Tukey, J. W., "The quefrency alanysis of time-series for echoes," *Proc. Symp. Time Series Analysis*, M. Rosenblatt, ed., chap. 15, pp. 209–243, 1963.

14. Oppenheim, A. V., Schafer, R. W., and Stockham, T. G., "Nonlinear filtering of multiplied and convolved signals," *Proc. IEEE*, vol. 56, pp. 1264–1291, Aug. 1968.

15. Noll, A. M., "Cepstrum pitch determination," *J. Acoust. Soc. Am.*, vol. 41, pp. 293–309, Feb. 1967.

16. Rabiner, L. R., Schafer, R. W., and Rader, C. M., "The Chirp

z-transform algorithm and its application," *Bell System Tech. J.*, vol. 48, pp. 1249–1292, May–June 1969.

17. Tomlinson, R. S., "SPASS—an improved terminal-analog speech synthesizer," *J. Acoust. Soc. Am.*, vol. 38, p. 940(A), 1965.

18. Fant, G., Martony, J., Rengman, U., and Risberg, A., "OVE II synthesis strategy," Paper F5, Speech Commun. Seminar, Stockholm, 1962.

19. Flanagan, J. L., "Note on the design of terminal-analog speech synthesizers," *J. Acoust. Soc. Am.*, vol. 29, pp. 306–310, 1957.

20. Coker, C. H., and Cummiskey, P., "On-line computer control of a formant synthesizer," *J. Acoust. Soc. Am.*, vol. 38, p. 940(A), 1965.

21. Holmes, J. N., "Notes on synthesis work," Speech Transmission Lab. Quart. Progr. Rept., Stockholm, Apr. 1961.

22. Rabiner, L. R., "Digital-formant synthesizer for speech synthesis studies," *J. Acoust. Soc. Am.*, vol. 43, pp. 822–828, Apr. 1968.

23. Rosenberg, A. E., Schafer, R. W., and Rabiner, L. R., "An investigation of the effects of smoothing and quantization of the parameters of formant coded speech," to be presented at 80th Meeting of the Acoustical Society of America.

24. Lee, F. F., "Reading machine: from text to speech," *IEEE Trans. Audio and Electroacoustics*, vol. AU-17, pp. 275–282, Dec. 1969.

25. Houde, R. A., "A study of tongue body motion during selected speech sounds," Doctoral dissertation, Univ. of Michigan, 1967.

26. Teranishi, R., and Umeda, N., "Use of pronouncing dictionary in speech synthesis experiments," *Reports of the 6th ICA*, vol. II, pp. 155–158, 1968.

27. Vanderslice, R., "Synthetic elocution—consideration in automatic orthographic to phonetic conversion of English with special reference to prosodic features," Working Papers in Phonetics 8, U.C.L.A., Feb. 1968.

28. Allen, J., "Machine-to-man communication by speech. Part 2: Synthesis of prosodic features of speech by rule," *1968 Spring Joint Comput. Conf., AFIPS Proc.*, vol. 32. Washington, D.C.: Thompson, 1968, pp. 339–344.

29. Pike, K. L., *The Intonation of American English.* Ann Arbor: Univ. of Michigan Press, 1945.

30. Rosenberg, A. E., "A computer-controlled system for the subjective evaluation of speech samples," *IEEE Trans. Audio and Electroacoustics*, vol. AU-17, pp. 216–221, Sept. 1969.

31. Rosenberg, A. E., "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Am.*, in press.

32. Hall, J. L., "Maximum-likelihood sequential procedure for estimation of psychometric functions," *J. Acoust. Soc. Am.*, vol. 44, p. 370(A), 1968.

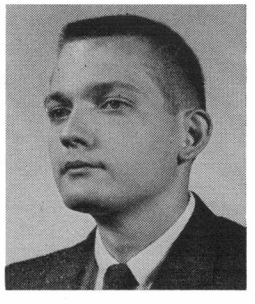
33. Flanagan, J. L., "Use of an interactive laboratory computer to study an acoustic-oscillator model of the vocal cords," *IEEE Trans. Audio and Electroacoustics*, vol. AU-17, pp. 2–6, Mar. 1969.

James L. Flanagan (F) received the B.S. degree from Mississippi State University in 1948 and the S.M. and Sc.D. degrees from the Massachusetts Institute of Technology in 1950 and 1955 respectively, all in electrical engineering. He was a member of the electrical engineering faculty at Mississippi State from 1950 to 1952, returning to M.I.T. for doctoral study as a Rockefeller Foundation Fellow. After joining Bell Laboratories in 1957, he became head of its Speech and Auditory Research Department in 1961 and of the Acoustics Research Department in 1967. His interests have centered on voice communication and digital techniques for signal analysis and transmission. He holds patents in speech coding, digital processing, and underwater acoustics.

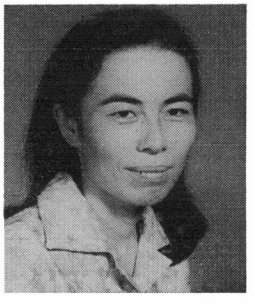
Dr. Flanagan has written a number of technical papers and a book, "Speech Analysis, Synthesis and Perception." He is chairman of the IEEE Audio and Electroacoustics Group and is a Fellow of the Acoustical Society of America. He is a member of the Committee on Hearing and Bioacoustics of the National Academy of Sciences, the Sensory Aids Subcommittee of the National Academy of Engineering, Tau Beta Pi, and Sigma Xi.

Cecil H. Coker (SM) is a supervisor in the Acoustics Research Department at Bell Laboratories. He received the B.S. and M.S. degrees in electrical engineering from Mississippi State University in 1954 and 1956 respectively, and in 1960 he received the Ph.D. degree in electrical engineering, with a minor in physics, from the University of Wisconsin. He then taught for a year in the university's Electrical Engineering Department. Dr. Coker joined Bell Laboratories in 1961 and worked for several years on formant analysis and synthesis of speech. Subsequent work included supervision of the development of two laboratory computer facilities. His work on modeling of the articulatory process was begun in 1966. He holds several patents and has written a number of papers on speech analysis and synthesis.

Lawrence Rabiner (M) received the S.B. and S.M. degrees simultaneously in 1964 and the Ph.D. degree in electrical engineering in 1967, all from the Massachusetts Institute of Technology. From 1962 to 1964 he participated in the cooperative plan in electrical engineering at Bell Laboratories, in Murray Hill and Whippany, N.J. He worked on digital circuitry, military communications problems, and problems in binaural hearing. At present he is engaged in research on speech communications and digital signal-processing techniques at Bell Laboratories. He is a member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, and the Acoustical Society of America.



Ronald W. Schafer (M) received the B.Sc.E.E. and the M.Sc.E.E. degrees in 1961 and 1962 from the University of Nebraska, Lincoln. In 1968 he earned the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology. While attending M.I.T. he was associated with its Electronic Systems Laboratory in 1965 and, between 1966 and 1968, he was a member of its Research Laboratory of Electronics. He joined the Acoustics Research Department of Bell Laboratories in 1968 and is now engaged there in research on speech analysis and synthesis, and digital signal-processing techniques.



Noriko Umeda received the B.S. and the M.S. degrees in linguistics from the University of Tokyo, Japan, in 1957 and 1959. She completed the doctorate course requirements in the same field and at the same university in 1962. That year, she joined the research staff of the Electrotechnical Laboratory, Ministry of International Trade and Industry, Japan, and there worked on speech analysis and synthesis—which included the problem of converting printed text to speech. She joined Bell Laboratories after taking a leave of absence from Electrotechnical in 1969. Her primary interest has been to study speech phenomenons (articulatory, prosodic, and syntactic) using the technique of synthesis by rule.