Received 2 April 1971; revised 21 April 1971 9.8, 9.3

Speech Analysis and Synthesis by Linear Prediction of the Speech Wave

B. S. Atal and Suzanne L. Hanauer

Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey 07974

We describe a procedure for efficient encoding of the speech wave by representing it in terms of time-varying parameters related to the transfer function of the vocal tract and the characteristics of the excitation. The speech wave, sampled at 10 kHz, is analyzed by predicting the present speech sample as a linear combination of the 12 previous samples. The 12 predictor coefficients are determined by minimizing the mean-squared error between the actual and the predicted values of the speech samples. Fifteen parameters—namely, the 12 predictor coefficients, the pitch period, a binary parameter indicating whether the speech is voiced or unvoiced, and the rms value of the speech samples—are derived by analysis of the speech wave, encoded and transmitted to the synthesizer. The speech wave is synthesized as the output of a linear recursive filter excited by either a sequence of quasiperiodic pulses or a white-noise source. Application of this method for efficient transmission and storage of speech signals as well as procedures for determining other speech characteristics, such as formant frequencies and bandwidths, the spectral envelope, and the autocorrelation function, are discussed.

INTRODUCTION

Efficient representation of speech signals in terms of a small number of slowly varying parameters is a problem of considerable importance in speech research. Most methods for analyzing speech start by transforming the acoustic data into spectral form by performing a short-time Fourier analysis of the speech wave.[1] Although spectral analysis is a well-known technique for studying signals, its application to speech signals suffers from a number of serious limitations arising from the nonstationary as well as the quasiperiodic properties of the speech wave.[2] As a result, methods based on spectral analysis often do not provide a sufficiently accurate description of speech articulation. We present in this paper a new approach to speech analysis and synthesis in which we represent the speech waveform directly in terms of time-varying parameters related to the transfer function of the vocal tract and the characteristics of the source function.[3–5] By modeling the speech wave itself, rather than its spectrum, we avoid the problems inherent in frequency-domain methods. For instance, the traditional Fourier analysis methods require a relatively long speech segment to provide adequate spectral resolution. As a result, rapidly changing speech events cannot be accurately followed. Furthermore, because of the periodic nature of voiced speech, little information about the spectrum between pitch harmonics is available; consequently, the frequency-domain techniques do not perform satisfactorily for high-pitched voices such as the voices of women and children. Although pitch-synchronous analysis-by-synthesis techniques can provide a partial solution to the above difficulties, such techniques are extremely cumbersome and time consuming even for modern digital computers and are therefore unsuitable for automatic processing of large amounts of speech data.[6,7] In contrast, the techniques presented in this paper are shown to avoid these problems completely.

The speech analysis–synthesis technique described in this paper is applicable to a wide range of research problems in speech production and perception. One of the main objectives of our method is the synthesis of speech which is indistinguishable from normal human speech. Much can be learned about the information-carrying structure of speech by selectively altering the properties of the speech signal. These techniques can thus serve as a tool for modifying the acoustic properties of a given speech signal without degrading the speech quality. Some other potential applications of these techniques are in the areas of efficient storage and transmission of speech, automatic formant and pitch extraction, and speaker and speech recognition.

In the rest of the paper, we describe a parametric model for representing the speech signal in the time

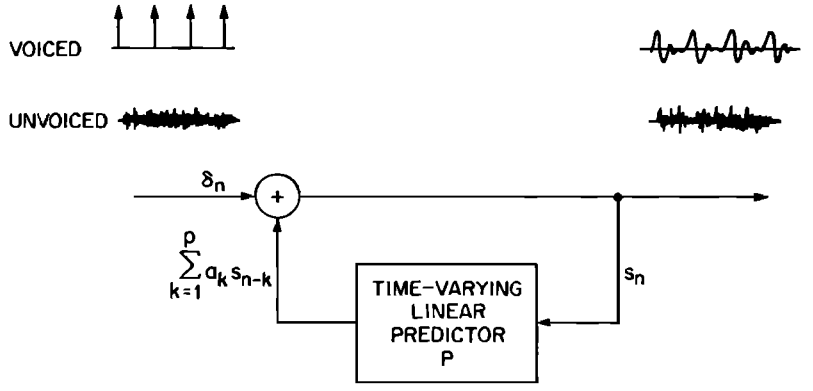


FIG. 1. Block diagram of a functional model of speech production based on the linear prediction representation of the speech wave.

domain; we discuss methods for analyzing the speech wave to obtain these parameters and for synthesizing the speech wave from them. Finally, we discuss applications for efficient coding of speech, estimation of the spectral envelope, formant analysis, and for modifying the acoustic properties of the speech signal.

The paper is organized such that most of the mathematical details are discussed in a set of appendixes. The main body of the paper is nearly complete in itself, and those readers who are not interested in the mathematical or computational aspects may skip the appendixes.

I. MODEL FOR PARAMETRIC REPRESENTATION OF THE SPEECH WAVE

In modern signal-processing techniques, the procedures for analyzing a signal make use of all the information that can be obtained in advance about the structure of that signal. The first step in signal analysis is thus to make a model of the signal.

Speech sounds are produced as a result of acoustical excitation of the human vocal tract. During the production of voiced sounds, the vocal tract is excited by a series of nearly periodic pulses generated by the vocal cords. In the case of unvoiced sounds, the excitation is provided by air passing turbulently through constrictions in the tract. A simple model of the vocal tract can be made by representing it as a discrete time-varying linear filter. If we assume that the variations with time of the vocal-tract shape can be approximated with sufficient accuracy by a succession of stationary shapes, it is possible to define a transfer function in the complex z domain for the vocal tract. The transfer function of a linear network can always be represented by its poles and zeros. It is well known that for nonnasal voiced speech sounds the transfer function of the vocal tract has no zeros.[8] For these sounds, the vocal tract can therefore be adequately represented by an all-pole (recursive) filter. A representation of the vocal tract for unvoiced and nasal sounds usually includes the antiresonances (zeros) as well as the resonances (poles) of the vocal tract. Since the zeros of the transfer function of the vocal tract for unvoiced and nasal sounds lie within the unit circle in the z plane,[9] each factor in the numerator of the transfer function can be approximated by multiple poles in the denominator of the transfer function.[10] In addition, the location of a pole is considerably more important perceptually than the location of a zero; the zeros in most cases contribute only to the spectral balance. Thus, an explicit representation of the antiresonances by zeros of the linear filter is not necessary. An all-pole model of the vocal tract can approximate the effect of antiresonances on the speech wave in the frequency range of interest to any desired accuracy.

The z transform of the glottal volume flow during a single pitch period can also be assumed to have poles only and no zeros. With this approximation, the z transform of the glottal flow can be represented by

$$U_g(z)=\frac{K_1}{(1-z_a z^{-1})(1-z_b z^{-1})}, \tag{1}$$

where K_1 is a constant related to the amplitude of the glottal flow and z_a, z_b are poles on the real axis inside the unit circle. In most cases, one of the poles is very close to the unit circle. If the radiation of sound from the mouth is approximated as radiation from a simple spherical source, then the ratio between the sound pressure at the microphone and the volume velocity at the lips is represented in the z-transform notation as $K_2(1-z^{-1})$, where K_2 is a constant related to the amplitude of the volume flow at the lips and the distance from the lips to the microphone.[11] The contribution of the glottal volume flow, together with the radiation, can thus be represented in the transfer function by the factor

$$\frac{K_1K_2(1-z^{-1})}{(1-z_a z^{-1})(1-z_b z^{-1})},$$

which, in turn, can be approximated as

$$\frac{K_1K_2}{[1+(1-z_a)z^{-1}](1-z_b z^{-1})}. \tag{2}$$

The error introduced by this approximation is given by

$$\frac{K_1K_2z^{-2}(1-z_a)}{(1-z_a z^{-1})[1+(1-z_a)z^{-1}](1-z_b z^{-1})}.$$

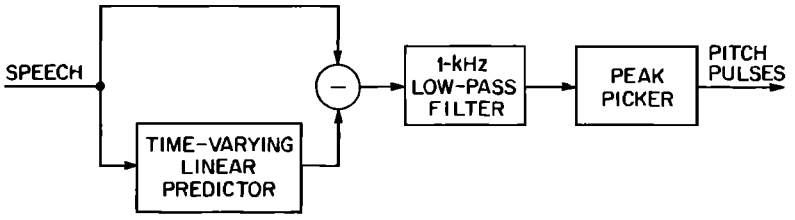


FIG. 2. Block diagram of the pitch pulse detector.

The contribution of this error to the transfer function in the frequency range of interest can be assumed to be small, since $z_a \approx 1$.

One of the important features of our model is that the combined contributions of the glottal flow, the vocal tract, and the radiation are represented by a single recursive filter. The difficult problem of separating the contribution of the source function from that of the vocal tract is thus completely avoided.

This representation of the speech signal is illustrated in sampled-data form in Fig. 1. The vocal-cord excitation for voiced sounds is produced by a pulse generator with adjustable period and amplitude. The noise-like excitation of unvoiced sounds is produced by a white-noise source. The linear predictor P, a transversal filter with p delays of one sample interval each, forms a weighted sum of the past p samples at the input of the predictor. The output of the linear filter at the nth sampling instant is given by

$$s_n = \sum_{k=1}^{p} a_k s_{n-k} + \delta_n, \tag{3}$$

where the "predictor coefficients" a_k account for the filtering action of the vocal tract, the radiation, and the glottal flow; and δ_n represents the nth sample of the excitation.

The transfer function of the linear filter of Fig. 1 is given by

$$T(z) = 1/(1 - \sum_{k=1}^{p} a_k z^{-k}). \tag{4}$$

The poles of $T(z)$ are the (reciprocal) zeros of the polynomial (in z^{-1}) in the denominator on the right side of Eq. 4. The linear filter thus has a total of p poles which are either real or occur in conjugate pairs. Moreover, for the linear filter to be stable, the poles must be inside the unit circle.

The number of coefficients p required to represent any speech segment adequately is determined by the number of resonances and antiresonances of the vocal tract in the frequency range of interest, the nature of the glottal volume flow function, and the radiation. As discussed earlier, two poles are usually adequate to represent the influence of the glottal flow and the radiation on the speech wave. It is shown in Appendix B that, in order to represent the poles of the vocal-tract transfer function adequately, the linear predictor memory must be equal to twice the time required for sound waves to travel from the glottis to the lips (nasal opening for nasal sounds). For example, if the vocal tract is 17 cm in length, the memory of the predictor should be roughly 1 msec in order to represent the poles of transfer function of the vocal tract. The corresponding value of p is then 10 for a sampling interval of 0.1 msec. With the two poles required for the glottal flow and the radiation added, p should be approximately 12. These calculations are meant to provide only a rough estimate of p and will depend to some extent on the speaker as well as on the spoken material. The results based on speech synthesis experiments (see Sec. IV) indicate that, in most cases, a value of p equal to 12 is adequate at a sampling frequency of 10 kHz. p is, naturally, a function of the sampling frequency f_s and is roughly proportional to f_s.

The predictor coefficients a_k, together with the pitch period, the rms value of the speech samples, and a binary parameter indicating whether the speech is voiced or unvoiced, provide a complete representation of the speech wave over a time interval during which the vocal-tract shape is assumed to be constant. During speech production, of course, the vocal-tract shape changes continuously in time. In most cases, it is sufficient to readjust these parameters periodically, for example, once every 5 or 10 msec.

II. SPEECH ANALYSIS

A. Determination of the Predictor Parameters

Going back to Fig. 1, we see that, except for one sample at the beginning of every pitch period, samples of voiced speech are linearly predictable in terms of the past p speech samples. We now use this property of the speech wave to determine the predictor coefficients. Let us define the prediction error E_n as the difference between the speech sample s_n and its predicted value $\hat{s}_n$ given by

$$\hat{s}_n = \sum_{k=1}^{p} a_k s_{n-k}. \tag{5}$$

E_n is then given by

$$E_n = s_n - \hat{s}_n = s_n - \sum_{k=1}^{p} a_k s_{n-k}. \tag{6}$$

We define the mean-squared prediction error $\langle E_n^2 \rangle_{av}$ as the average of E_n^2 over all the sampling instances n in the speech segment to be analyzed except those at the

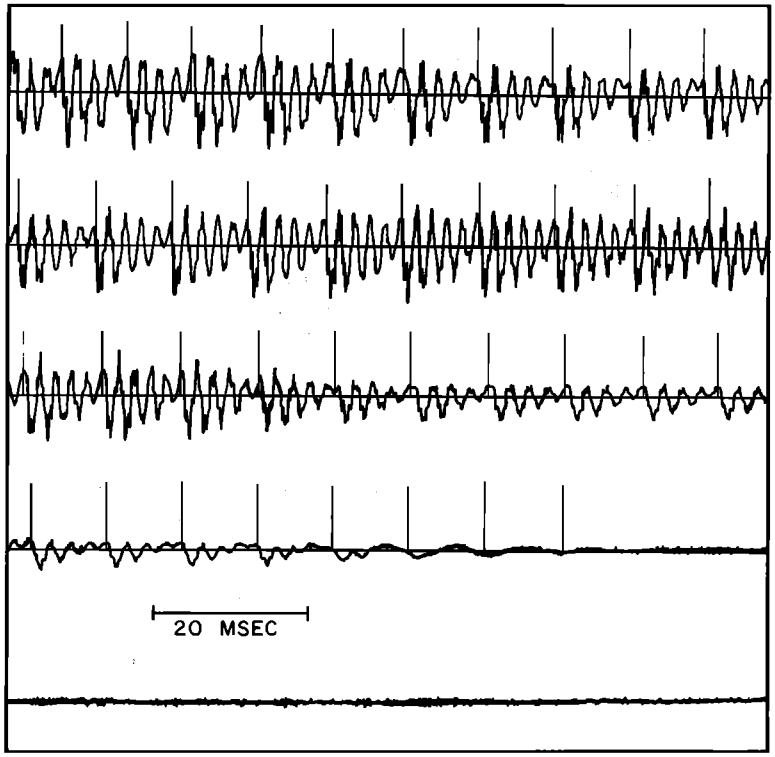


FIG. 3. Waveform of the speech signal together with the positions of the pitch pulses (shown by vertical lines).

beginning of each pitch period, i.e.,

$$\langle E_n^2\rangle_{\text{av}} = \langle (s_n - \sum_{k=1}^{p} a_k s_{n-k})^2\rangle_{\text{av}}. \tag{7}$$

The predictor coefficients a_k of Eq. 3 are chosen so as to minimize the mean-squared prediction error $\langle E_n^2\rangle_{\text{av}}$. The same procedure is used to determine the predictor parameters for unvoiced sounds, too.

The coefficients a_k which minimize the mean-squared prediction error are obtained by setting the partial derivative of $\langle E_n^2\rangle_{\text{av}}$ with respect to each a_k equal to zero. It can then be shown[3] that the coefficients a_k are obtained as solutions of the set of equations

$$\sum_{k=1}^{p} \varphi_{jk} a_k = \varphi_{j0}, \quad j=1, 2, \cdots, p, \tag{8}$$

where

$$\varphi_{jk} = \langle s_{n-j} s_{n-k}\rangle_{\text{av}}. \tag{9}$$

In general, the solution of a set of simultaneous linear equations requires a great deal of computation. However, the set of linear equations given by Eq. 8 is a special one, since the matrix of coefficients is symmetric and positive definite. There are several methods of solving such equations.[12,13] A computationally efficient method of solving Eq. 8 is outlined in Appendix C.

Occasionally, the coefficients a_k obtained by solving Eq. 8 produce poles in the transfer function which are outside the unit circle. This can happen whenever a pole of the transfer function near the unit circle appears outside the unit circle, owing to approximations in the model. The locations of all such poles must be corrected. A simple computational procedure to determine if any pole of the transfer function is outside the unit circle and a method for correcting the predictor coefficients are described in Appendix D.

B. Pitch Analysis

Although any reliable pitch-analysis method can be used to determine the pitch of the speech signal, we outline here briefly two methods of pitch analysis which are sufficiently reliable and accurate for our purpose.

In the first method,[14] the speech wave is filtered through a 1-kHz low-pass filter and each filtered speech sample is raised to the third power to emphasize the high-amplitude portions of the speech waveform. The duration of the pitch period is obtained by performing a pitch-synchronous correlation analysis of the cubed speech. The voiced–unvoiced decision is based on two factors, the density of zero crossings in the speech wave and the peak value of the correlation function. This method of pitch analysis is described in detail in Ref. 14.

The second method of pitch analysis is based on the linear prediction representation of the speech wave.[15] It follows from Fig. 1 that, except for a sample at the beginning of each pitch period, every sample of the voiced speech waveform can be predicted from the past sample values. Therefore, the positions of individual pitch pulses can be determined by computing the prediction error E_n given by Eq. 6 and then locating the

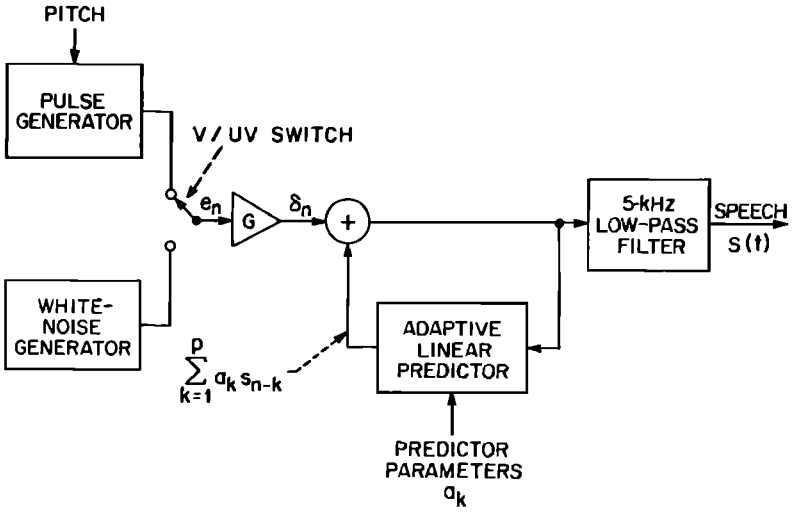


FIG. 4. Block diagram of the speech synthesizer.

samples for which the prediction error is large. The latter function is easily accomplished by a suitable peak-picking procedure. This procedure is illustrated in Fig. 2. In practice, the prediction error was found to be large at the beginning of the pitch periods and a relatively simple peak-picking procedure was found to be effective. The voiced–unvoiced decision is based on the ratio of the mean-squared value of the speech samples to the mean-squared value of the prediction error samples. This ratio is considerably smaller for unvoiced speech sounds than for voiced speech sounds—typically, by a factor of 10. The result of the pitch analysis on a short segment of the speech wave is illustrated in Fig. 3. The positions of the individual pitch pulses, shown by vertical lines, are superimposed on the speech waveform for easy comparison.

III. SPEECH SYNTHESIS

The speech signal is synthesized by means of the same parametric representation as was used in the analysis. A block diagram of the speech synthesizer is shown in Fig. 4. The control parameters supplied to the synthesizer are the pitch period, a binary voiced–unvoiced parameter, the rms value of the speech samples, and the p predictor coefficients. The pulse generator produces a pulse of unit amplitude at the beginning of each pitch period. The white-noise generator produces uncorrelated uniformly distributed random samples with standard deviation equal to 1 at each sampling instant. The selection between the pulse generator and the white-noise generator is made by the voiced–unvoiced switch. The amplitude of the excitation signal is adjusted by the amplifier G. The linearly predicted value $\hat{s}_n$ of the speech signal is combined with the excitation signal δ_n to form the nth sample of the synthesized speech signal. The speech samples are finally low-pass filtered to provide the continuous speech wave $s(t)$.

It may be pointed out here that, although for time-invariant networks the synthesizer of Fig. 4 will be equivalent to a traditional formant synthesizer with variable formant bandwidths, its operation for the time-varying case (which is true in speech synthesis) differs significantly from that of a formant synthesizer. For instance, a formant synthesizer has separate filters for each formant and, thus, a correct labeling of formant frequencies is essential for the proper functioning of a formant synthesizer. This is not necessary for the synthesizer of Fig. 4, since the formants are synthesized together by one recursive filter. Moreover, the amplitude of the pitch pulses as well as the white noise is adjusted to provide the correct rms value of the synthetic speech samples.

The synthesizer control parameters are reset to their new values at the beginning of every pitch period for voiced speech and once every 10 msec for unvoiced speech. If the control parameters are not determined pitch-synchronously in the analysis, new parameters are computed by suitable interpolation of the original parameters to allow pitch-synchronous resetting of the synthesizer. The pitch period and the rms value are interpolated "geometrically" (linear interpolation on a logarithmic scale). In interpolating the predictor coefficients, it is necessary to ensure the stability of the recursive filter in the synthesizer. The stability cannot, in general, be ensured by direct linear interpolation of the predictor parameters. One suitable method is to interpolate the first p samples of the autocorrelation function of the impulse response of the recursive filter. The autocorrelation function has the important advantage of having a one-to-one relationship with the predictor coefficients. Therefore, the predictor coefficients can be recomputed from the autocorrelation function. Moreover, the predictor coefficients derived from the autocorrelation function always result in a stable filter in the synthesizer.[16] The relationship between the predictor coefficients and the autocorrelation function can be derived as follows:

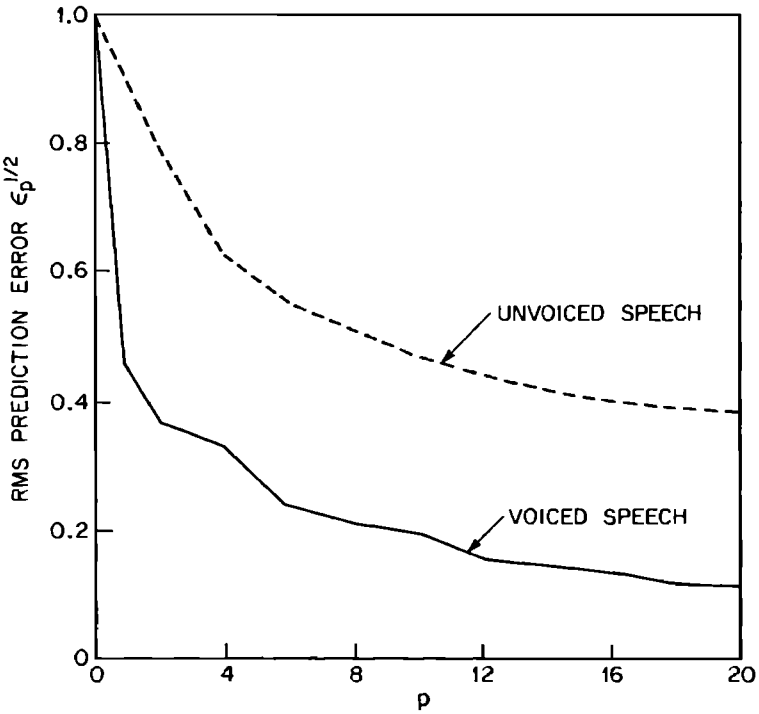


FIG. 5. Variation of the minimum value of the rms prediction error with p, the number of predictor coefficients. Solid line shows the curve for voiced speech. Dotted line shows the curve for unvoiced speech.

From Eq. 3, the impulse response of the linear recursive filter of Fig. 1 satisfies the equation

$$s_n = \sum_{k=1}^{p} a_k s_{n-k}, \quad n \geq 1, \tag{10}$$

with the initial conditions $s_0 = 1$ and $s_n = 0$ for $n < 0$. The autocorrelation function of the impulse response is, by definition, given by

$$r_{|i|} = \sum_{n=0}^{\infty} s_n s_{n+|i|}. \tag{11}$$

Let us multiply both sides of Eq. 10 by s_{n+i} and perform a sum over n from 0 to ∞. We then obtain

$$r_i = \sum_{k=1}^{p} a_k r_{|i-k|}, \quad i \geq 1, \tag{12}$$

and

$$r_0 = \sum_{k=1}^{p} a_k r_k + 1. \tag{13}$$

Equations 12 and 13 enable us to compute the samples of the autocorrelation function from the predictor coefficients, and the predictor coefficients from the autocorrelation function. A computational procedure for performing the above operations is outlined in Appendix E.

The gain of the amplifier G is adjusted to provide the correct power in the synthesized speech signal. In any speech segment, the amplitude of the nth synthesized speech sample s_n can be decomposed into two parts: one part q_n contributed by the memory of the linear predictor carried over from the previous speech segments and the other part v_n contributed by the excitation from the current speech segment. Thus, $s_n = q_n + v_n = q_n + g u_n$, where g is the gain of the amplifier G. Let us assume that $n = 1$ is the first sample and $n = M$ the last sample of the current speech segment. The first part q_n is given by

$$q_n = \sum_{k=1}^{p} a_k q_{n-k}, \quad 1 \leq n \leq M, \tag{14}$$

where $q_0, q_{-1}, \cdots, q_{1-p}$ represent the memory of the predictor carried over from the previous synthesized speech segments. In addition, u_n is given by

$$u_n = \sum_{k=1}^{p} a_k u_{n-k} + e_n, \quad 1 \leq n \leq M, \tag{15}$$

where $u_n = 0$ for nonpositive values of n, and e_n is the nth sample at the output of the voiced–unvoiced switch as shown in Fig. 4. Let P_s be the mean-squared value of the speech samples. Then P_s is given by

$$P_s = \frac{1}{M} \sum_{n=1}^{M} (q_n + g u_n)^2 = \overline{(q_n + g u_n)^2}. \tag{16}$$

On further rearrangement of terms, Eq. 16 is rewritten as

$$g^2 \overline{u_n^2} + 2g \overline{q_n u_n} + \overline{q_n^2} - P_s = 0. \tag{17}$$

Equation 17 is solved for g such that g is real and non-

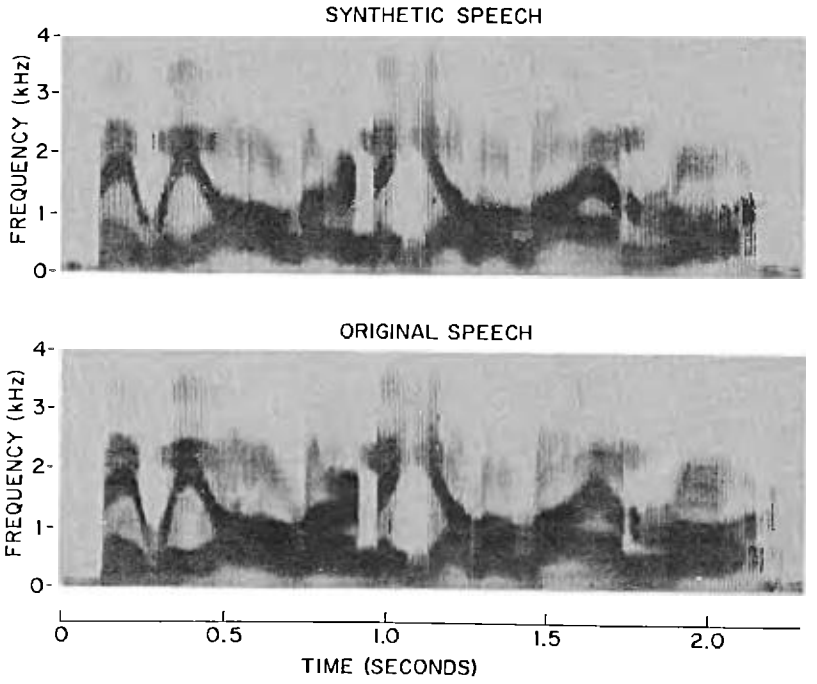


FIG. 6. Comparison of wide-band sound spectrograms for synthetic and original speech signals for the utterance "May we all learn a yellow lion roar," spoken by a male speaker: (a) synthetic speech, and (b) original speech.

negative. In case such a solution does not exist, g is set to zero. The nth sample of the synthesized wave is finally obtained by adding q_n to $g u_n$.

IV. COMPUTER SIMULATION OF THE ANALYSIS–SYNTHESIS SYSTEM

In order to assess the subjective quality of the synthesized speech, the speech analysis and synthesis system described above was simulated on a digital computer. The speech wave was first low-pass filtered to 5 kHz and then sampled at a frequency of 10 kHz. The analysis segment was set equal to a pitch period for voiced speech and equal to 10 msec for unvoiced speech. The various parameters were then determined for each analysis segment according to the procedure described in Sec. II. These parameters were finally used to control the speech synthesizer shown in Fig. 4.

The optimum value for the number of predictor parameters p was determined as follows: The speech wave was synthesized for various values of p between 2 and 18. Informal listening tests revealed no significant differences between synthetic speech samples for p larger than 12. There was slight degradation in speech quality at p equal to 8. However, even for p as low as 2, the synthetic speech was intelligible although poor in quality. The influence of decreasing p to values less than 10 was most noticeable on nasal consonants. Furthermore, the effect of decreasing p was less noticeable on female voices than on male voices. This could be expected in view of the fact that the length of the vocal tract for female speakers is generally shorter than for male speakers and that the nasal tract is slightly longer than the oral tract. From these results, it was concluded that a value of p equal to 12 was required to provide an adequate representation of the speech signal. It may be worthwhile at this point to compare these results with the objective results based on an examination of the variation of the prediction error as a function of p. In Fig. 5, we have plotted the minimum value of the rms prediction error as a function of several values of p. The speech power in each case was normalized to unity. The results are presented separately for voiced and unvoiced speech. As can be seen in the figure, the prediction error curve is relatively flat for values of p greater than 12 for voiced speech and for p greater than 6 for unvoiced speech. These results suggest again that p equal to 12 is adequate for voiced speech. For unvoiced speech, a lower value of p, e.g., p equal to 6, should be adequate. For those readers who wish to listen to the quality of synthesized speech at various values of p, a recording accompanies this article. Appendix A gives the contents of the record. The reader should listen at this point to the first section of the record.

In informal listening tests, the quality of the synthetic speech was found to be very close to that of the original speech for a wide range of speakers and spoken material. No significant differences were observed between the synthetic speech samples of male and female speakers. The second section of the record includes examples of synthesized speech for several utterances of different speakers. In each case, p was set to equal to 12. The spectrograms of the synthetic and the original speech for two of these utterances are compared in Figs. 6 and 7. As can be seen, the spectrogram of the synthetic speech closely resembles that of the original speech.

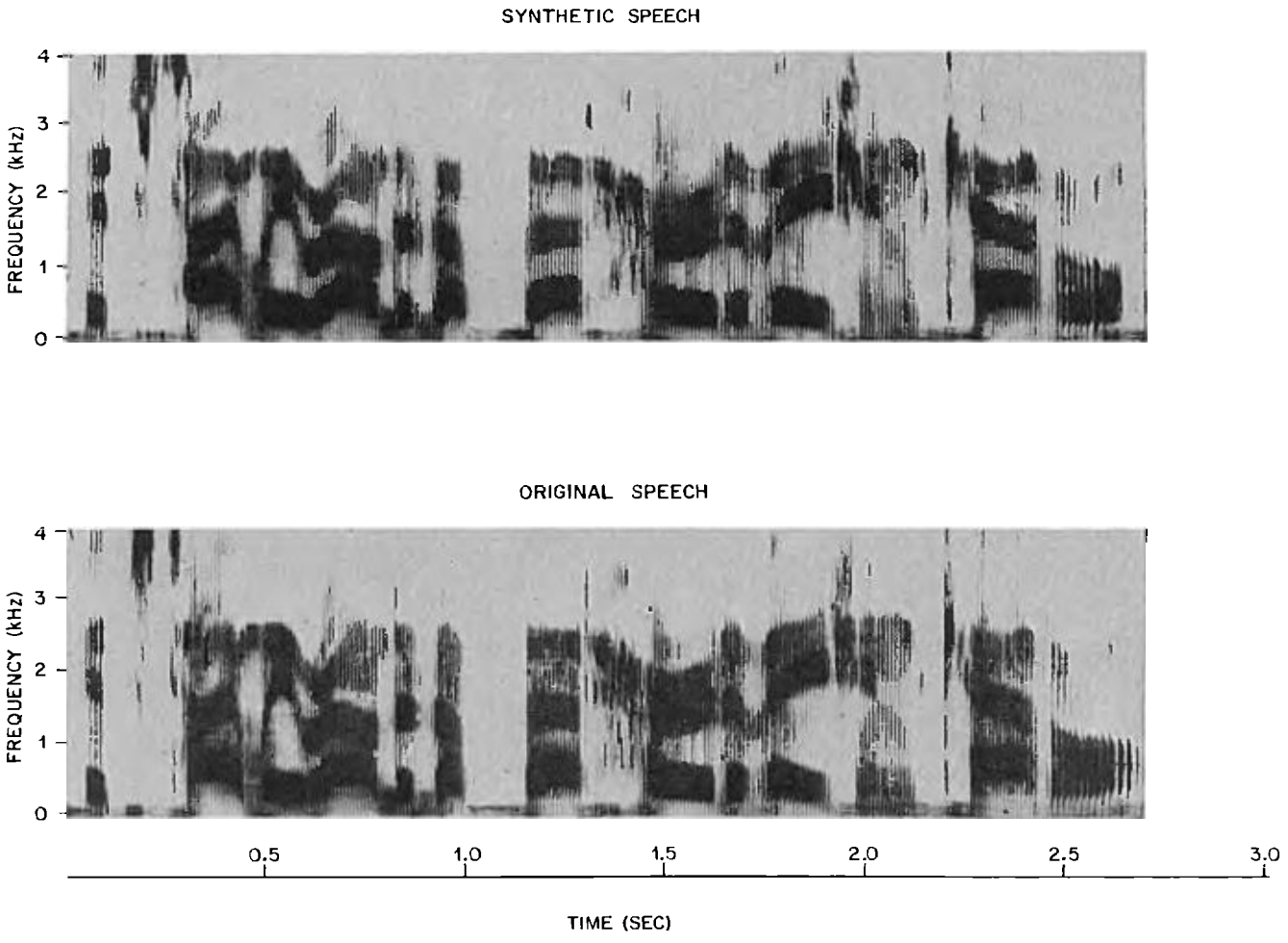


FIG. 7. Comparison of wide-band sound spectrograms for synthetic and original speech signals for the utterance "It's time we rounded up that herd of Asian cattle," spoken by a male speaker: (a) synthetic speech, and (b) original speech.

V. APPLICATIONS

A. Digital Storage and Transmission of Speech

Methods for encoding speech at data rates considerably smaller than those needed for PCM encoding are important in many practical applications. For example, automatic answerback services can be practical if a sufficiently large vocabulary of words and phrases can be stored economically in a digital computer. Efficient speech coding methods can reduce, by a factor of 30 or more, the space needed for storing the vocabulary. We discuss in this section several procedures for efficient coding of the synthesizer control information.

The synthesizer control information includes 15 parameters for every analysis interval, i.e., the twelve predictor coefficients, the pitch period, the voiced unvoiced parameter, and the rms value. The methods for proper encoding of this information, except the predictor coefficients, are relatively well understood.[17] On the other hand, the procedure for encoding the predictor coefficients must include provision for ensuring the stability of the linear filter in the synthesizer. In general, to ensure stability, relatively high accuracy (about 8–10 bits per coefficient) is required if the predictor coefficients are quantized directly. Moreover, the predictor coefficients are samples of the inverse Fourier transform of the reciprocal of the transfer function. The reciprocal of the transfer function has zeros precisely where the transfer function has poles. Therefore, small errors in the predictor coefficients often can result in large errors in the poles. The direct quantization of the predictor coefficients is thus not efficient. One suitable method is to convert the 12 predictor coefficients to another equivalent set of parameters which possess well-defined constraints for achieving the desired stability. For example, the poles of the linear filter can be computed from the predictor coefficients. For stability of the filter, it is sufficient that the poles be inside the unit circle. The stability is therefore easily ensured by quantizing the frequencies and the bandwidths of the poles. The poles of the transfer function are by definition the

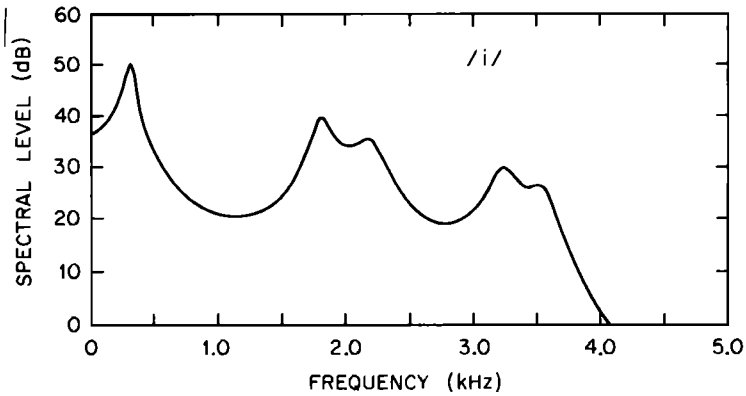


FIG. 8. Spectral envelope for the vowel /i/ in "we," spoken by a male speaker (F_0=120 Hz).

roots of the polynomial equation

$$\sum_{k=1}^{p} a_k z^{-k} = 1, \tag{18}$$

where, as before, a_k are the predictor coefficients. Table I shows the precision with which each of the parameters is quantized. It was found that the frequencies and the bandwidths of the poles can be quantized within 60 bits without producing any perceptible effect on the synthesized speech. Adding this value to the bits needed for the pitch (6 bits), the rms value (5 bits), and the voiced–unvoiced parameter (1 bit), one arrives at a value of 72 bits (60+6+5+1) for each frame of analyzed data. The data rate in bits/sec is obtained by multiplying the number of bits used to encode each frame of data by the number of frames of data stored or transmitted per second. Thus, a bit rate of 7200 bits/sec is achieved if the parameters are sampled at a rate of 100/sec. The bit rate is lowered to 2400 bits/sec at a sampling rate of 33/sec.

At this point, the reader can listen to recorded examples of synthesized speech encoded at three different data rates, namely, 7200, 4800, and 2400 bits/sec, respectively, in the third section of the enclosed record.

The quantizing of the frequencies and the bandwidths of the poles is not the only method of encoding the predictor coefficients. For example, it can be shown (see Appendix F) that a transfer function with p poles is always realizable as the transfer function of an acoustic tube consisting of p cylindrical sections of equal length with the last section terminated by a unit acoustic resistance. Moreover, the poles are always inside the unit circle if the cross-sectional area of each cylindrical section is positive. Thus, the stability of the synthesizer filter is easily achieved by quantizing the areas of the sections or any other suitable function of the areas.

TABLE I. Quantization of synthesizer control information.

Parameter	Number of levels	Bits
Pitch	64	6
V/UV	2	1
rms	32	5
Frequencies and bandwidths of the poles		60
	Total	72

No significant difference in speech quality was observed for the different quantizing methods outlined above at various bit rates above 2400 bits/sec. It is quite possible that at very low bit rates these different methods of coding may show appreciable differences. An example of speech synthesized using area quantization is presented in the fourth section of the record.

The data rates discussed in this paper are suitable for speech-transmission applications where large buffer storage is to be avoided. The efficiency of speech coding naturally can vary considerably from one application to another. For example, it has been assumed so far that the speech signal is analyzed at uniform time intervals. However, it may be more efficient to vary the analysis interval so that it is short during fast articulatory transitions and long during steady-state segments. Furthermore, in applications such as disk storage of voice messages, additional savings can be realized by choosing the quantization levels for each parameter around its mean value determined in advance over short time intervals. The mean value itself can be quantized separately.

B. Separation of Spectral Envelope and Fine Structure

It is often desirable to separate the envelope of the speech spectrum from its fine structure.[18] The representation of the speech signal shown in Fig. 1 is very suitable for achieving this decomposition. In this representation, the fine structure of the spectrum is contributed by the source while the envelope is contributed by the linear filter. Thus, the two are easily separated.[19] The spectral envelope is the power spectrum of the impulse response of the linear filter. In mathematical notation, the relationship between the spectral envelope $G(f)$ at the frequency f and the predictor coefficients is expressed by

$$G(f) = 1/|1 - \sum_{k=1}^{p} a_k e^{-2\pi j f k/f_s}|^2, \tag{19}$$

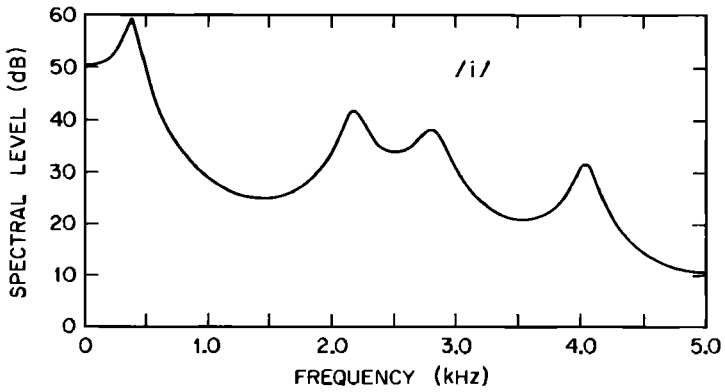


FIG. 9. Spectral envelope for the vowel /i/ in "we," spoken by a female speaker (F_0=200 Hz).

where a_k, as before, are the predictor coefficients and f_s is the sampling frequency. Two examples of the spectral envelope obtained in the above manner for the vowel /i/ belonging to the word "we" in the utterance "May we all learn a yellow lion roar" spoken by a male and a female speaker are illustrated in Figs. 8 and 9, respectively. We would like to add here that a spectral section obtained on a sound spectrograph failed to separate the third formant from the second formant for the female speaker both in the wide-band and the narrow-band analysis. The spectral section showed one broad peak for the two formants. On the other hand, the spectral envelope of Fig. 9 shows the two formants without any ambiguity. Of course, it is difficult to evaluate the accuracy of this method from results based on real speech alone. Results with synthetic speech, where the spectral envelope is known precisely, indicate that the spectral envelope is accurately determined over a wide range of pitch values (from 50 to 300 Hz).

It also follows from Eq. 19 that, although the Fourier transform of $G(f)$ is not time limited, the Fourier transform of $1/G(f)$ is time limited to $2p/f_s$ sec. Thus, spectral samples of $G(f)$, spaced $f_s/2p$ Hz apart, are sufficient for reconstruction of the spectral envelope. For p=12 and f_s=10 kHz, this means that a spacing of roughly 400 Hz between spectral samples is adequate.

In some applications, it may be desired to compute the Fourier transform of $G(f)$, namely, the autocorrelation function. The autocorrelation function can be determined directly from the predictor coefficients without computing $G(f)$. The relationship between the predictor coefficients and the autocorrelation function is given in Eqs. 12 and 13, and a computational method for performing these operations is outlined in Appendix E.

C. Formant Analysis

The objective of formant analysis is to determine the complex natural frequencies of the vocal tract as they change during speech production. If the vocal-tract configuration were known, these natural frequencies could be computed. However, the speech signal is influenced both by the properties of the source and by the vocal tract. For example, if the source spectrum has a zero close to one of the natural frequencies of the vocal tract, it will be extremely difficult, if not impossible, to determine the frequency or the bandwidth of that particular formant. A side-branch element such as the nasal cavity creates a similar problem. In determining formant frequencies and bandwidths from the speech signal, one can at best hope to obtain such information which is not obscured or lost owing to the influence of the source.

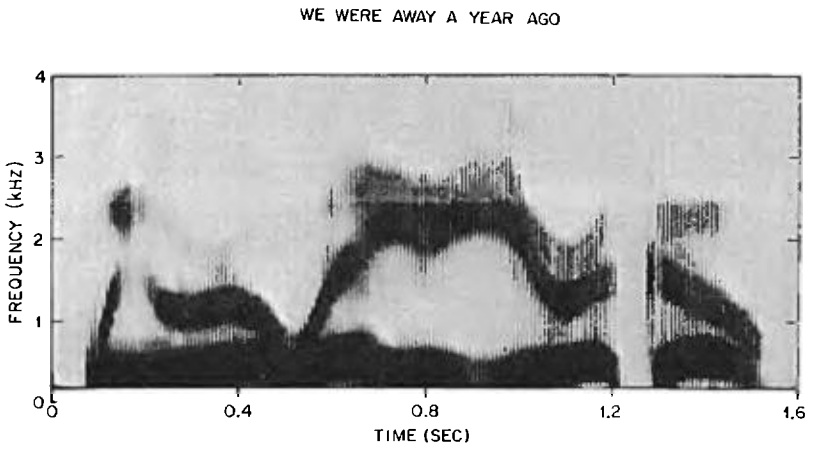
Present methods of formant analysis usually start by transforming the speech signal into a short-time Fourier spectrum, and consequently suffer from many additional problems which are inherent in short-time Fourier transform techniques.[5,6,20,21] Such problems, of course, can be completely avoided by determining the formant frequencies and bandwidths directly from the speech wave.[2]

In the representation of the speech wave shown in Fig. 1, the linear filter represents the combined contributions of the vocal tract and the source to the spectral envelope. Thus, the poles of the transfer function of the filter include the poles of the vocal tract as well as the source. So far, we have made no attempt to separate these two contributions. For formant analysis, however, it is necessary that the poles of the vocal tract be separated out from the transfer function. In general, it is our experience that the poles contributed by the source either fall on the real axis in the unit circle or produce a relatively small peak in the spectral envelope. The magnitude of the spectral peak produced by a pole can easily be computed and compared with a threshold to determine whether a pole of the transfer function is indeed a natural frequency of the vocal tract. This is accomplished as follows:

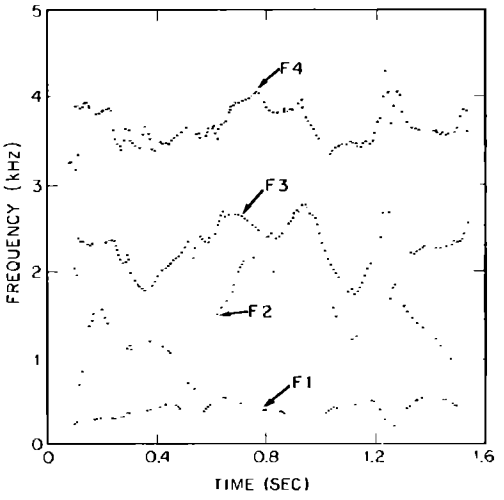
From Eq. 4, the poles of the transfer function are the roots of the polynomial equation

$$\sum_{k=1}^{p} a_k z^{-k} = 1. \tag{20}$$

Let there be n complex conjugate pairs of roots $z_1, z_1^*; z_2, z_2^*; \cdots; z_n, z_n^*$. The transfer function due to these



(a)



(b)

FIG. 10. Formant frequencies for the utterance "We were away a year ago," spoken by a male speaker ($F_0=120$ Hz). (a) Wide-band sound spectrogram for the above utterance, and (b) formants determined by the computer program.

roots is given by

$$V(z)=\prod_{i=1}^{n}(1-z_i)(1-z_i^*)/\prod_{i=1}^{n}(z-z_i)(z-z_i^*), \quad (21)$$

where the additional factors in the numerator set the transfer function at dc ($z=1$) equal to 1. The spectral peak produced by the kth complex conjugate pole pair is given by

$$A_k=\left|\frac{(1-z_k)(1-z_k^*)}{(z-z_k)(z-z_k^*)}\right|^2, \quad (22)$$

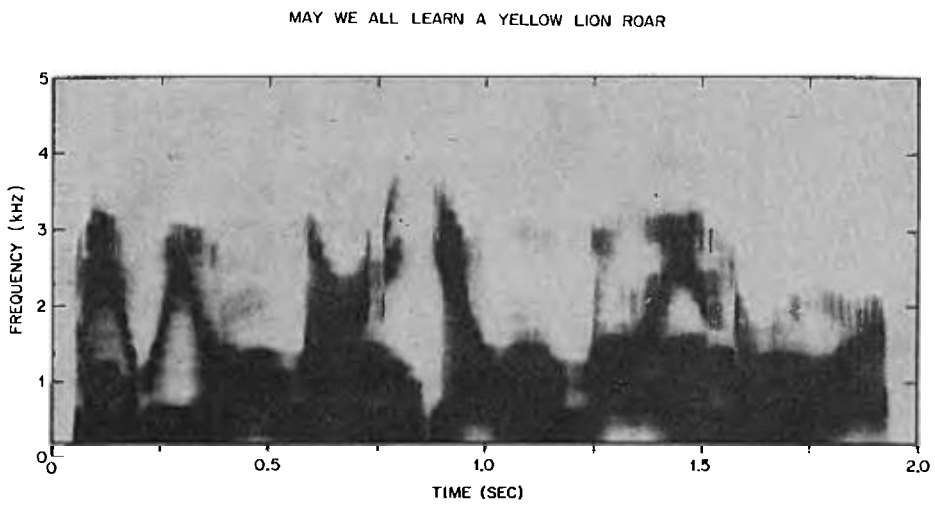
where $z=\exp(2\pi jf_kT)$, $z_k=|z_k|\exp(2\pi jf_kT)$, and T is the sampling interval. The threshold value of A_k was set equal to 1.7. Finally, the formant frequency F_k and the bandwidth (two-sided) B_k are related to the z-plane root z_k by

$$F_k=(1/2\pi T)\,\mathrm{Im}(\ln z_k), \quad (23)$$

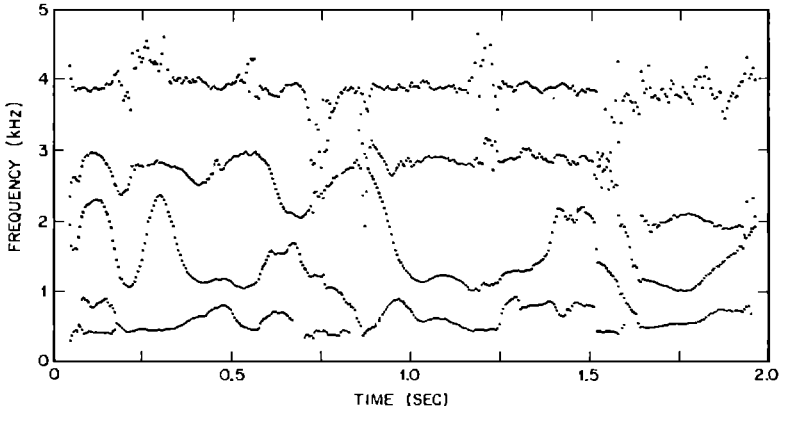
and

$$B_k=(1/\pi T)\,\mathrm{Re}\left(\frac{1}{\ln z_k}\right). \quad (24)$$

Examples of the formant frequencies determined according to the above procedure are illustrated in Figs. 10–12. Each figure consists of (a) a wide-band sound spectrogram of the utterance, and (b) formant data as determined by the above method. The results are presented for three different utterances. The first utterance, "We were away a year ago," was spoken by a male speaker (average fundamental frequency $F_0=120$ Hz).



(a)



(b)

FIG. 11. Formant frequencies for the utterance "May we all learn a yellow lion roar," spoken by a female speaker ($F_0=200$ Hz). (a) Wide-band sound spectrogram for the above utterance, and (b) formants determined by the computer program.

The second utterance, "May we all learn a yellow lion roar," was spoken by a female speaker ($F_0=200$ Hz). The third utterance, "Why do I owe you a letter?" was spoken by a male speaker ($F_0=125$ Hz). Each point in these plots represents the results from a single frame of the speech signal which was equal to a pitch period in Figs. 10 and 11 and equal to 10 msec in Fig. 12. No smoothing of the formant data over adjacent frames was done.

TABLE II. Factor by which each parameter was scaled for simulating a female voice from parameters derived from a male voice.

Parameter	Scaling factor
Pitch period T	0.58
Formant frequencies F_i	1.14
Formant bandwidths B_i	$2-F_i/5000$

Again, in order to obtain a better estimate of the accuracy of this method of formant analysis, speech was synthesized with a known formant structure. The correspondence between the actual formant frequencies and bandwidths and the computed ones was found to be extremely close.

D. Re-forming the Speech Signals

The ability to modify the acoustical characteristics of a speech signal without degrading its quality is important for a wide variety of applications. For example, information regarding the relative importance of various acoustic variables in speech perception can be obtained by listening to speech in which some particular

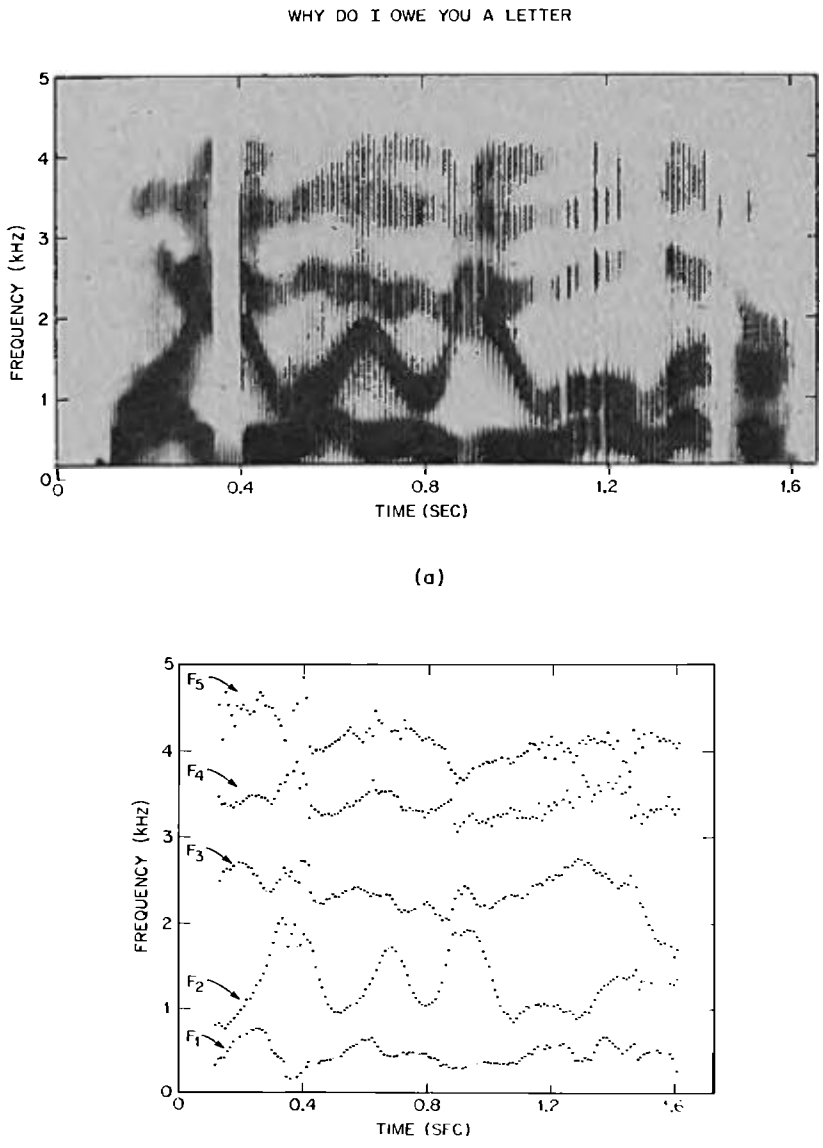


FIG. 12. Formant frequencies for the utterance "Why do I owe you a letter?" spoken by a male speaker (F_0=125 Hz). (a) Wide-band sound spectrogram for the above utterance, and (b) formants determined by the computer program.

acoustic variables have been altered in a controlled manner. The speech analysis and synthesis techniques described in this paper can be used as a flexible and convenient method for conducting such speech-perception experiments. We would like to point out here that the synthesis procedure allows independent control of such speech characteristics as spectral envelope, relative durations, pitch, and intensity. Thus, the speaking rate of a given speech signal may be altered, e.g., for producing fast speech for blind persons or for producing slow speech for learning foreign languages. Or, in an application such as the recovery of "helium speech," the frequencies of the spectral envelope can be scaled, leaving the fundamental frequency unchanged. Moreover, in synthesizing sentence-length utterances from stored data about individual words, the method can be used to reshape the intonation and stress contours so that the speech sounds natural.

Examples of speech in which selected acoustical characteristics have been altered are presented in the fifth section of the enclosed record. First, the listener can hear the utterance at the normal speaking rate. Next, the speaking rate is increased by a factor of 1.5. As the third item, the same utterance with the speaking rate reduced by a factor of 1.5 is presented. Finally, an example of a speech signal in which the pitch, the formant frequencies, and their bandwidths were changed from their original values, obtained from a male voice, to

TABLE III. Computation times needed to perform various operations discussed in the paper on the GE 635 ($p=10$, $f_s=10$ kHz).

Operation	Computation time
Predictor coefficients from speech samples (No. of samples=100)	75 msec/frame
Spectral envelope (500 spectral samples) from predictor coefficients	250 msec/frame
Formant frequencies and bandwidths from predictor coefficients	60 msec/frame
p samples of autocorrelation function from predictor coefficients	10 msec/frame
Speech from predictor coefficients	8 times real time
Pitch analysis	10 times real time

simulate a "female" voice is presented. The factor by which each parameter was changed from its original value is shown in Table II.

VI. COMPUTATIONAL EFFICIENCY

The computation times needed to perform several of the operations described in this paper are summarized in Table III. The programs were run on a GE 635 computer having a cycle time of 1 μsec. As can be seen, this method of speech analysis and synthesis is computationally efficient. In fact, the techniques are about five to 10 times faster than the ones needed to perform equivalent operations by fast-Fourier-transform methods. For instance, both the formant frequencies and their bandwidths are determined in 135 msec for each frame of the speech wave 10 msec long. Assuming that the formants are analyzed once every 10 msec, the program will run in about 13 times real time; by comparison, fast-Fourier-transform techniques need about 100 times real time. Even for computing the spectral envelope, the method based on predictor coefficients is at least three times faster than the fast-Fourier-transform methods. The complete analysis and synthesis procedure was found to run in approximately 25 times real time. Real-time operation could easily be achieved by using special hardware to perform some of the functions.

VII. CONCLUSIONS

We have presented a method for automatic analysis and synthesis of speech signals by representing them in terms of time-varying parameters related to the transfer function of the vocal tract and the characteristics of the excitation. An important property of the speech wave, namely, its linear predictability, forms the basis of both the analysis and synthesis procedures. Unlike past speech analysis methods based on Fourier analysis, the method described here derives the speech parameters from a direct analysis of the speech wave. Consequently, various problems encountered when Fourier analysis is applied to nonstationary and quasiperiodic signals like speech are avoided. One of the main advantages of this method is that the analysis procedure requires only a short segment of the speech wave to yield accurate results. This method is therefore very suitable for following rapidly changing speech events. It is also suitable for analyzing the speech of speakers with high-pitched voices, such as women or children. As an additional advantage, the analyzed parameters are rigorously related to other well-known speech characteristics. Thus, by first representing the speech signal in terms of the predictor coefficients, other speech characteristics can be determined as desired without much additional computation.

The speech signal is synthesized by a single recursive filter. The synthesizer, thus, does not require any information about the individual formants and the formants need not be determined explicitly during analysis. Moreover, the synthesizer makes use of the formant bandwidths of real speech, in contrast to formant synthesizers, which use fixed bandwidths for each formant. Informal listening tests show very little or no perceptible degradation in the quality of the synthesized speech. These results suggest that the analyzed parameters retain all the perceptually important features of the speech signal. Furthermore, the various parameters used for the synthesis can be encoded efficiently. It was found possible to reduce the data rate to approximately 2400 bits/sec without producing significant degradation in the speech quality. The above bit rate is smaller by a factor of about 30 than that for direct PCM encoding of the speech waveform. The latter bit rate is approximately 70 000 bits (70 000 bits=7 bits/sample$\times$10 000 samples/sec).

In addition to providing an efficient and accurate description of the speech signal, the method is computationally very fast. The entire analysis and synthesis procedure runs at about 25 times real time on a GE 635 digital computer. The method is thus well suited for analyzing large amounts of speech data automatically on the computer.

APPENDIX A: DESCRIPTION OF ENCLOSED RECORDED MATERIAL

Side 1

Section 1. Speech analysis and synthesis for various values of p, the number of predictor coefficients:

(*a*) $p=2$,
(*b*) $p=6$,
(*c*) $p=10$,
(*d*) $p=14$,
(*e*) $p=18$,
(*f*) original speech.

Section 2. Comparison of synthesized speech with the original, $p=12$. Synthetic—original. Five utterances.

Section 3. Synthesized speech encoded at different bit rates, the parameters quantized as shown in Table I, $p=12$. Original—unquantized—7200 bits/sec—4800 bits/sec—2400 bits/sec. Three utterances.

Section 4. Synthesized speech obtained by quantizing the areas of an acoustic tube, $p=12$. Bit rate=7200 bits/sec.

(1) Frequencies and bandwidths quantized into 60-bit frames.
(2) Areas quantized into 60-bit frames.

The rest of the parameters are quantized as shown in Table I.

Section 5. Fast and slow speech, $p=14$:

(*a*) Original speech.
(*b*) Speaking rate=1.5 times the original.
(*c*) Speaking rate=0.67 times the original.

Section 6. Manipulation of pitch, formant frequencies, and their bandwidths, $p=10$:

(*a*) Pitch, formant frequencies, and bandwidths altered as shown in Table II.
(*b*) Original voice.

APPENDIX B: RELATIONSHIP BETWEEN THE LENGTH OF THE VOCAL TRACT AND THE NUMBER OF PREDICTOR COEFFICIENTS

Below about 5000 Hz, the acoustic properties of the vocal tract can be determined by considering it as an acoustic tube of variable cross-sectional area. The relationship between sound pressure P_g and volume velocity U_g at the glottis and the corresponding quantities P_l, U_l at the lips is best described in terms of the *ABCD* matrix parameters (chain matrix) of the acoustic tube. These parameters are defined by the matrix equation (see Fig. B-1):

$$\begin{bmatrix} P_g \\ U_g \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} P_l \\ U_l \end{bmatrix}. \tag{B1}$$

We now prove that the inverse Fourier transforms of these parameters in the time domain have finite duration $\tau=2l/c$, where l is the length of the tube and c is the velocity of sound. Let $S(x)$ be the area function of the vocal tract, where x is the distance from the glottis to the point at which the cross-sectional area is specified. Consider a small element of tube of length dx at a distance x from the glottis. The *ABCD* matrix parameters of the tube element dx are given by[B1]

$$\begin{aligned} A &= D = \cosh \Gamma dx = 1/2(e^{\Gamma dx} + e^{-\Gamma dx}), \\ B &= -Z_0 \sinh \Gamma dx = -Z_0(e^{\Gamma dx} - e^{-\Gamma dx})/2, \\ C &= -\sinh \Gamma dx / Z_0 = -(e^{\Gamma dx} - e^{-\Gamma dx})/2Z_0, \end{aligned} \tag{B2}$$

where Z_0 is the characteristic impedance of the tube element $dx=\rho c/S(x)$, Γ is the propagation constant $=j\omega/c$, ρ is the density of air, c is the velocity of sound, and ω is the angular frequency in radians. The *ABCD* matrix of the complete tube is given by the product of the *ABCD* matrices of the individual tube elements of length dx spaced dx apart along the length of the tube. Let $l=ndx$. It is now easily verified that each of the *ABCD* parameters of the tube can be expressed as a power series in $e^{\Gamma dx}$ of the form

$$\sum_{k=-n}^{n} \alpha_k e^{k\Gamma dx}.$$

The *ABCD* parameters are thus Fourier transforms of functions of time each with duration $\tau=2n\cdot dx/c$. Taking the limit as $dx \to 0$, $n \to \infty$, and $n\cdot dx=l$, we obtain $\tau=2l/c$.

From Eq. B1, the relationship between the glottal and the lip volume velocities is expressed in terms of the *ABCD* parameters by

$$U_g = CP_l + DU_l. \tag{B3}$$

Since $P_l \cong j\omega K U_l$, K being a constant related to the mouth area, Eq. B3 is rewritten as

$$U_g \cong (j\omega KC + D)U_l. \tag{B4}$$

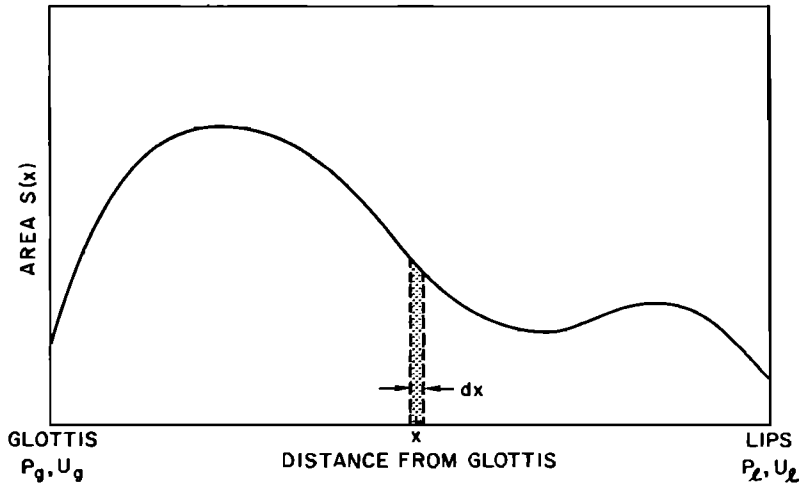


FIG. B-1. Nonuniform acoustic tube.

The memory of the linear predictor (see Fig. 1) is by definition equal to the duration of the inverse Fourier transform of the reciprocal of the transfer function between the lip and the glottal volume velocities. Therefore, from Eq. B4, the memory of the linear predictor is equal to $\tau=2l/c$.

APPENDIX C: DETERMINATION OF THE PREDICTOR PARAMETERS FROM THE COVARIANCE MATRIX Φ

Equation 8 can be written in matrix notation as

$$\Phi\mathbf{a}=\psi, \tag{C1}$$

where $\Phi=[(\varphi_{ij})]$ is a positive definite (or positive semidefinite) symmetric matrix, and $\mathbf{a}=[(a_j)]$ and $\psi=[(\varphi_{j0})]$ are column vectors. Since Φ is positive definite (or semidefinite) and symmetric, it can be expressed as the product of a triangular matrix V with real elements and its transpose V^t. Thus,

$$\Phi=VV^t. \tag{C2}$$

Equation C1 can now be resolved into two simpler equations:

$$V\mathbf{x}=\psi, \tag{C3}$$

$$V^t\mathbf{a}=\mathbf{x}. \tag{C4}$$

Since V is a triangular matrix, Eqs. C3 and C4 can be solved recursively.[C1]

Equations C3 and C4 provide a simple method of computing the minimum value of the prediction error $\langle E_n^2\rangle_{av}$ as a function of p, the number of predictor coefficients. It is easily verified from Eqs. 7–9 that the minimum value of the mean-squared prediction error is given by

$$e_p=\varphi_{00}-\mathbf{a}^t\psi. \tag{C5}$$

On substituting for $\mathbf{a}$ from Eq. C4 into Eq. C5, we obtain

$$\epsilon_p=\varphi_{00}-\mathbf{x}^tV^{-1}\psi,$$

which on substitution from Eq. C3 for ψ yields

$$\epsilon_p=\varphi_{00}-\mathbf{x}^t\mathbf{x}. \tag{C6}$$

Thus, the minimum value of the mean-squared prediction error is given as

$$\epsilon_p=\varphi_{00}-\sum_{k=1}^{p}x_k^2. \tag{C7}$$

The advantage of using Eq. C7 lies in the fact that a single computation of the vector x for one value of p is sufficient. After the vector $\mathbf{x}$ is determined for the largest value of p at which the error is desired, ϵ_p is calculated for smaller values of p from Eq. C7.

APPENDIX D: CORRECTION OF THE PREDICTOR COEFFICIENTS

Let us denote by $f(z)$ a polynomial defined by

$$f(z)=z^p-a_1z^{p-1}-\cdots-a_p, \tag{D1}$$

where the polynomial coefficients a_k are the predictor coefficients of Eq. 3. Associated with the polynomial $f(z)$, we define a reciprocal polynomial $f^*(z)$ by

$$f^*(z)=z^pf(z^{-1})=-a_pz^p-a_{p-1}z^{p-1}-\cdots-a_1z+1. \tag{D2}$$

Let us construct the sequence of polynomials $f_{p-1}(z)$, $f_{p-2}(z),\cdots,f_n(z),\cdots,f_1(z)$, where $f_n(z)$ is a polynomial of degree n, according to the formula

$$f_n(z)=k_{n+1}f_{n+1}^*(z)-l_{n+1}f_{n+1}(z), \tag{D3}$$

where $f_p(z)=f^*(z)$, k_n is the coefficient of z^n in $f_n(z)$, and l_n is the constant term in $f_n(z)$. It can then be shown that the polynomial $f(z)$ has all its zeros inside the unit circle if and only if $|l_n|>|k_n|$ for each $n\leq p$.[D1]

When one or more of the zeros of $f(z)$ are outside the unit circle, let us set

$$f(z)=\prod_{k=1}^{p}(z-z_k). \tag{D4}$$

For every k for which $|z_k|>1$, we replace z_k by $z_k/|z_k|$ in Eq. D4 and construct a new polynomial which has all of its zeros either inside or on the unit circle.

The above procedure is also suitable for testing that all the zeros of $f(z)$ are within any given circle $|z|=r$. In this case, we replace a_k by $r^{-k}a_k$ in Eq. D1 and proceed as before.

The roots of the polynomial $f(z)$ are determined by an iterative procedure based on the Newton–Raphson method.[D2] We start with a trial value of the root and then construct successively better approximations. The iteration formula has the form

$$z_k^{(n+1)}=z_k^{(n)}-\frac{f[z_k^{(n)}]}{f'[z_k^{(n)}]}, \tag{D5}$$

where $z_k^{(n)}$ is the approximation of the kth root at the nth iteration. The iteration process is continued until either $f(z)=0$ or the absolute difference between the roots in two successive iterations is less than a fixed threshold value. If convergence is not reached within 100 iterations, a new starting value is selected. The starting value is chosen randomly on the unit circle. Furthermore, the starting value never lies on the real axis.

APPENDIX E: DETERMINATION OF THE PREDICTOR COEFFICIENTS FROM THE AUTOCORRELATION FUNCTION AND THE AUTOCORRELATION FUNCTION FROM THE PREDICTOR COEFFICIENTS

We outline first a method of solving Eq. 12 for the predictor coefficients. Consider the set of equations

$$r_i=\sum_{k=1}^{n}a_k^{(n)}r_{|i-k|},\quad \text{for}\quad n\geq i\geq 1. \tag{E1}$$

Equation E1 is identical to Eq. 12 for $n=p$. In matrix form the above equation becomes

$$\begin{bmatrix} r_0 & r_1 & \cdots & r_{n-1} \\ r_1 & r_0 & \cdots & r_{n-2} \\ \vdots & \vdots & & \vdots \\ r_{n-2} & r_{n-3} & \cdots & r_1 \\ r_{n-1} & r_{n-2} & \cdots & r_0 \end{bmatrix} \begin{bmatrix} a_1^{(n)} \\ a_2^{(n)} \\ \vdots \\ a_{n-1}^{(n)} \\ a_n^{(n)} \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_{n-1} \\ r_n \end{bmatrix}. \quad \text{(E2)}$$

Let R_n be the $n\times n$ matrix on the left side of Eq. E2, $\mathbf{a}_m^{(n)}$ an m-dimensional vector whose kth component is $a_k^{(n)}$, and $\mathbf{r}_n$ an n-dimensional vector whose kth component is r_{n-k+1}. Let us define, for every vector, a reciprocal vector by the relationship

$$\mathbf{a}_m^{(n)*}|_k = \mathbf{a}_m^{(n)}|_{m-k+1}. \quad \text{(E3)}$$

Equation E2 can now be rewritten as

$$R_{n-1}\mathbf{a}_{n-1}^{(n)} + a_n^{(n)}\mathbf{r}_{n-1} = \mathbf{r}_{n-1}^*, \quad \text{(E4)}$$

and

$$\mathbf{r}_{n-1}^t\mathbf{a}_{n-1}^{(n)} + r_0 a_n^{(n)} = r_n. \quad \text{(E5)}$$

On multiplying Eq. E4 through by R_{n-1}^{-1} and rearranging terms, we obtain

$$\mathbf{a}_{n-1}^{(n)} = R_{n-1}^{-1}\mathbf{r}_{n-1}^* - a_n^{(n)}R_{n-1}^{-1}\mathbf{r}_{n-1}. \quad \text{(E6)}$$

It is easily verified from Eq. E2 that

$$R_{n-1}^{-1}\mathbf{r}_{n-1}^* = \mathbf{a}_{n-1}^{(n-1)}. \quad \text{(E7)}$$

Inserting Eq. E7 into Eq. E6 gives

$$\mathbf{a}_{n-1}^{(n)} = \mathbf{a}_{n-1}^{(n-1)} - a_n^{(n)}[\mathbf{a}_{n-1}^{(n-1)}]^*. \quad \text{(E8)}$$

Next, we multiply Eq. E8 through by $\mathbf{r}_{n-1}^t$ and insert the result in Eq. E5. After rearrangement of the terms, we find that

$$a_n^{(n)}[r_0 - \sum_{k=1}^{n-1} r_k a_k^{(n-1)}] = r_n - \sum_{k=1}^{n-1} r_{n-k}a_k^{(n-1)}. \quad \text{(E9)}$$

Equations E8 and E9 provide a complete recursive solution of Eq. E1. We start with $n=1$. The solution is obviously

$$a_1^{(1)} = r_1/r_0. \quad \text{(E10)}$$

Next, $a_n^{(n)}$ and $\mathbf{a}_{n-1}^{(n)}$ are computed for successively increasing values of n until $n=p$. Furthermore, if R_n is nonsingular, the expression inside the brackets on the left side of Eq. E9 is always positive. Therefore, $a_n^{(n)}$ is always finite.

To determine the autocorrelation function from the predictor coefficients, we proceed as follows: From Eq. E8,

$$[\mathbf{a}_{n-1}^{(n)}]^* = [\mathbf{a}_{n-1}^{(n-1)}]^* - a_n^{(n)}\mathbf{a}_{n-1}^{(n-1)}. \quad \text{(E11)}$$

Therefore, after eliminating $[a_{n-1}^{(n-1)}]^*$ from Eqs. E8 and E11, one obtains

$$\mathbf{a}_{n-1}^{(n-1)} = \{\mathbf{a}_{n-1}^{(n)} + a_n^{(n)}[\mathbf{a}_{n-1}^{(n)}]^*\}/\{1-[a_n^{(n)}]^2\}. \quad \text{(E12)}$$

Starting with $n=p$, we compute $\mathbf{a}_n^{(n)}$ for successively smaller values of n until $n=1$. The autocorrelation function at the nth sampling instant is given from Eq. E1 by

$$r_n = \sum_{k=1}^{n} a_k^{(n)} r_{n-k}, \quad \text{for} \quad 1\le n\le p. \quad \text{(E13)}$$

The samples of the autocorrelation function are computed recursively for successively larger values of n starting with $n=1$. Note that, on the right side of Eq. E13, only samples up to r_{n-1} appear in the sum. Thus, r_n can be computed recursively. Equation E13 is used to determine all the samples of the autocorrelation function with r_0 normalized to 1. The value of r_0 is finally determined from Eq. 13.

APPENDIX F: TRANSFER FUNCTION OF A NONUNIFORM ACOUSTIC TUBE

We consider sound transmission in an acoustic tube formed by cascading N uniform cylindrical sections, each of length Δ as shown in Fig. F-1. Let the cross-sectional area of the nth section be S_n. Let $p_n(t)$ and $v_n(t)$ be the components of the volume velocity due to the forward- and the backward-traveling waves, respectively, at the input of the nth section. Let us assume that there is a sound source of constant volume velocity at the input of the first section. Consider now the sound transmission between two adjacent sections, e.g., n and $n+1$. On applying the boundary conditions for the continuity of volume velocity and sound pressure across the junction, we obtain

$$p_n\left(t-\frac{\Delta}{c}\right) - v_n\left(t+\frac{\Delta}{c}\right) = p_{n+1}(t) - v_{n+1}(t), \quad \text{(F1)}$$

$$\left[p_n\left(t-\frac{\Delta}{c}\right) + v_n\left(t+\frac{\Delta}{c}\right)\right]\frac{\rho c}{S_n} = [p_{n+1}(t) + v_{n+1}(t)]\frac{\rho c}{S_{n+1}}, \quad \text{(F2)}$$

where ρc is the characteristic impedance of air. Equations F1 and F2 can be solved for $p_{n+1}(t)$ and $v_{n+1}(t)$. We then obtain

$$p_{n+1}(t) = \frac{1}{1+r_n}\left[p_n\left(t-\frac{\Delta}{c}\right) - r_n v_n\left(t+\frac{\Delta}{c}\right)\right], \quad \text{(F3)}$$

$$v_{n+1}(t) = \frac{1}{1+r_n}\left[-r_n p_n\left(t-\frac{\Delta}{c}\right) + v_n\left(t+\frac{\Delta}{c}\right)\right], \quad \text{(F4)}$$

where

$$r_n = \frac{S_n - S_{n+1}}{S_n + S_{n+1}}. \quad \text{(F5)}$$

It is convenient to write Eqs. F3 and F4 in the z-trans-

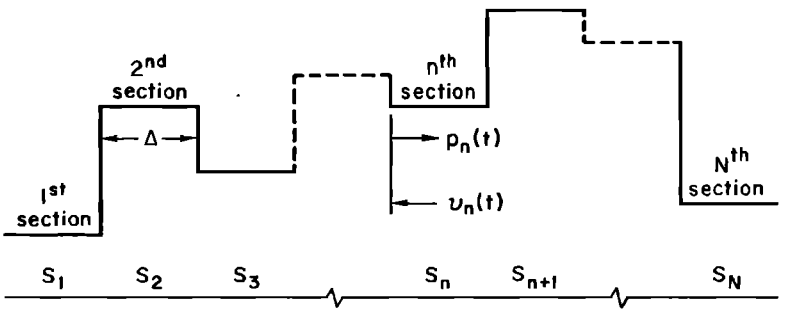


FIG. F-1. A nonuniform acoustic tube formed by cascading uniform cylindrical sections.

form notation as

$$\begin{bmatrix}\mathcal{P}_{n+1}(z)\\ \mathcal{U}_{n+1}(z)\end{bmatrix}=\frac{1}{1+r_n}\begin{bmatrix}z^{-\frac{1}{2}} & -r_n z^{\frac{1}{2}}\\ -r_n z^{-\frac{1}{2}} & z^{\frac{1}{2}}\end{bmatrix}\begin{bmatrix}\mathcal{P}_n(z)\\ \mathcal{U}_n(z)\end{bmatrix}, \quad \text{(F6)}$$

where $\mathcal{P}_n(z)$ and $\mathcal{U}_n(z)$ are the z transforms of $p_n(t)$ and $v_n(t)$, respectively, with $z=\exp[j\omega(2\Delta/c)]$. Similarly as in Eq. F6, we have also the inverse relationship

$$\begin{bmatrix}\mathcal{P}_n(z)\\ \mathcal{U}_n(z)\end{bmatrix}=\frac{1}{1-r_n}\begin{bmatrix}z^{\frac{1}{2}} & r_n z^{\frac{1}{2}}\\ r_n z^{-\frac{1}{2}} & z^{-\frac{1}{2}}\end{bmatrix}\begin{bmatrix}\mathcal{P}_{n+1}(z)\\ \mathcal{U}_{n+1}(z)\end{bmatrix}, \quad \text{(F7)}$$

which can be written in matrix notation as

$$H_n(z)=Q_n(z)H_{n+1}(z). \quad \text{(F8)}$$

Moreover, from Eq. F8,

$$H_1(z)=\prod_{k=1}^{N} Q_k(z)H_{N+1}(z), \quad \text{(F9)}$$

$$=W_N(z)H_{N+1}(z). \quad \text{(F10)}$$

Let

$$W_n(z)=\prod_{k=1}^{n} Q_k(z)=\begin{bmatrix}w_{11}^{(n)}(z) & w_{12}^{(n)}(z)\\ w_{21}^{(n)}(z) & w_{22}^{(n)}(z)\end{bmatrix}. \quad \text{(F11)}$$

The matrix $W_n(z)$ satisfies the equation

$$W_{n+1}(z)=W_n(z)Q_{n+1}(z). \quad \text{(F12)}$$

It can be verified from Eq. F11 that

$$J[W_n^t(z^{-1})]^{-1}J=W_n(z), \quad \text{(F13)}$$

where

$$J=\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}.$$

Equation F13 implies that

$$\begin{bmatrix}w_{11}(z) & w_{12}(z)\\ w_{21}(z) & w_{22}(z)\end{bmatrix}=\begin{bmatrix}w_{22}(z^{-1}) & w_{21}(z^{-1})\\ w_{12}(z^{-1}) & w_{11}(z^{-1})\end{bmatrix}, \quad \text{(F14)}$$

or

$$W_n(z)=\begin{bmatrix}w_{11}^{(n)}(z) & w_{21}^{(n)}(z^{-1})\\ w_{21}^{(n)}(z) & w_{11}^{(n)}(z^{-1})\end{bmatrix}. \quad \text{(F15)}$$

Let us assume that the tube is terminated in a unit acoustic resistance. We then have the terminal boundary condition

$$p_{N+1}(t)+v_{N+1}(t)=p_{N+1}(t)-v_{N+1}(t), \quad \text{(F16)}$$

from which it follows that $v_{N+1}(t)=0$. The volume velocity at the input of the tube is given by $p_1(t)-v_1(t)$. Let

$$C_N(z)=\frac{\text{volume velocity at the input}}{\text{volume velocity at the output}}$$

$$=\frac{\mathcal{P}_1(z)-\mathcal{U}_1(z)}{\mathcal{P}_{N+1}(z)}. \quad \text{(F17)}$$

It can now be easily verified from Eq. F10 that

$$C_N(z)=w_{11}^{(N)}(z)-w_{21}^{(N)}(z). \quad \text{(F18)}$$

Let us define for each n, between 1 and N,

$$C_n(z)=w_{11}^{(n)}(z)-w_{21}^{(n)}(z). \quad \text{(F19)}$$

On multiplying Eq. F12 by a vector $[1 \quad -1]$, and substituting for $W_n(z)$ from Eq. F15, we find that

$$[C_{n+1}(z) \quad -C_{n+1}(z^{-1})]=\frac{1}{1-r_{n+1}}[C_n(z) \quad -C_n(z^{-1})]\times\begin{bmatrix}z^{\frac{1}{2}} & r_{n+1}z^{\frac{1}{2}}\\ r_{n+1}z^{-\frac{1}{2}} & z^{-\frac{1}{2}}\end{bmatrix}. \quad \text{(F20)}$$

Hence,

$$C_{n+1}(z)=\frac{1}{1-r_{n+1}}[z^{\frac{1}{2}}C_n(z)-z^{-\frac{1}{2}}r_{n+1}C_n(z^{-1})]$$

$$=\frac{z^{\frac{1}{2}}}{1-r_{n+1}}[C_n(z)-r_{n+1}C_n(z^{-1})z^{-1}]. \quad \text{(F21)}$$

Except for the factor $z^{n/2}$, each $C_n(z)$ is a polynomial of degree n. Thus, the transfer function, which is the reciprocal of $C_N(z)$, consists of a factor $z^{-N/2}$ divided by a polynomial of degree N. The factor $z^{-N/2}$ represents, of course, the transmission delay in the tube. The transfer function has N poles which are the zeros of $C_N(z)$. Furthermore, the poles are inside the unit circle, pro-

vided that r_n satisfies the condition[D1]

$$|r_n|<1, \quad 1\leq n\leq N, \tag{F22}$$

which, together with Eq. F5, implies that

$$S_n>0, \quad 1\leq n\leq N. \tag{F23}$$

We now show that every all-pole transfer function having poles inside the unit circle is always realizable, except for a constant multiplying factor and a delay, as the transfer function of an acoustic tube.

It follows from Eq. F20 that

$$C_n(z)=\frac{z^{-\frac{1}{2}}}{1+r_{n+1}}[C_{n+1}(z)+r_{n+1}C_{n+1}(z^{-1})]. \tag{F24}$$

Furthermore, for each $C_n(z)$, the ratio of the coefficients of $z^{-n/2}$ and $z^{n/2}$ is r_n. Given $C_N(z)$, one can compute $C_n(z)$ for successively decreasing values of n starting with $n=N$ from Eq. F24. In each case, the coefficient r_n is always defined as the ratio of the coefficients of $z^{-n/2}$ and $z^{n/2}$. A sequence of numbers, $r_1, r_2, \cdots, r_N$, obtained in the above manner, defines a tube with areas $S_1, S_2, \cdots, S_N$ according to Eq. F5, provided that each of the areas is positive, i.e., $|r_n|<1$ for $1\leq n\leq N$. This is, however, assured if the original polynomial $C_N(z)$ has all its roots inside the unit circle (see Appendix D). Since the poles of the transfer function are inside the unit circle, it is indeed true.

[1] J. L. Flanagan, *Speech Analysis Synthesis and Perception* (Academic, New York, 1965), p. 119.

[2] E. N. Pinson, "Pitch-Synchronous Time-Domain Estimation of Formant Frequencies and Bandwidths," J. Acoust. Soc. Amer. **35**, 1264–1273 (1963).

[3] B. S. Atal and M. R. Schroeder, "Adaptive Predictive Coding of Speech Signals," Bell System Tech. J. **49**, 1973–1986 (1970).

[4] B. S. Atal, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Amer. **47**, 65(A) (1970).

[5] B. S. Atal, "Characterization of Speech Signals by Linear Prediction of the Speech Wave," Proc. IEEE Symp. on Feature Extraction and Selection in Pattern Recognition, Argonne, Ill. (Oct. 1970), pp. 202–209.

[6] M. V. Mathews, J. E. Miller, and E. E. David, Jr., "Pitch Synchronous Analysis of Voiced Sounds," J. Acoust. Soc. Amer. **33**, 179–186 (1961).

[7] C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens, and A. S. House, "Reduction of Speech Spectra by Analysis-by-Synthesis Techniques," J. Acoust. Soc. Amer. **33**, 1725–1736 (1961).

[8] G. Fant, *Acoustic Theory of Speech Production* (Mouton, The Hague, 1960), p. 42.

[9] B. S. Atal, "Sound Transmission in the Vocal Tract with Applications to Speech Analysis and Synthesis," Proc. Int. Congr. Acoust., 7th, Budapest, Hungary (Aug. 1971).

[10] Each factor of the form $(1-az^{-1})$ can be approximated by $[1/(1+az^{-1}+a^2z^{-2}+\cdots)]$ if $|a|<1$, which is the case if the zeros are inside the unit circle.

[11] Ref. 1, p. 33.

[12] C. E. Fröberg, *Introduction to Numerical Analysis* (Addison-Wesley, Reading, Mass., 1969), 2nd ed., pp. 81–101.

[13] J. P. Ellington and H. McCallion, "The Determination of Control System Characteristics from a Transient Response," Proc. IEE **105**, Part C, 370–373 (1958).

[14] B. S. Atal, "Automatic Speaker Recognition Based on Pitch Contours," PhD thesis, Polytech. Inst. Brooklyn (1968).

[15] B. S. Atal, "Pitch-Period Analysis by Inverse Filtering" (to be published).

[16] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications* (Univ. California Press, Berkeley, 1958), p. 40.

[17] L. G. Stead and R. C. Weston, "Sampling and Quantizing the Parameters of a Formant-Tracking Vocoder System," Proc. Speech Commun. Seminar, R.I.T., Stockholm (29 Aug.–1 Sept. 1962).

[18] M. R. Schroeder, "Vocoders: Analysis and Synthesis of Speech," Proc. IEEE **54**, 720–734 (1966).

[19] The problem of separating the spectral envelope from the fine structure of the speech spectrum should be distinguished from the problem of separating the influence of the source from the speech spectrum. The latter problem is far more difficult and is discussed partially in the next subsection.

[20] R. W. Schafer and L. R. Rabiner, "System for Automatic Formant Analysis of Voiced Speech," J. Acoust. Soc. Amer. **47**, 634–648 (1970).

[21] J. P. Olive, "Automatic Formant Tracking by a Newton–Raphson Technique," J. Acoust. Soc. Amer. **50**, 661–670 (1971).

[B1] I. Malecki, *Physical Foundations of Technical Acoustics*, English transl. by I. Bellert (Pergamon, Oxford, England, 1969), p. 475.

[C1] D. K. Faddeev and V. N. Faddeeva, *Computational Methods of Linear Algebra*, English transl. by R. C. Williams (W. H. Freeman, San Francisco, 1963), pp. 144–147.

[D1] Ref. 16, pp. 40–41. See also L. Ya. Geronimus, *Orthogonal Polynomials* (Consultants Bureau, New York, 1961), p. 156.

[D2] Ref. 12, pp. 21–28.