Digital Representations of Speech Signals

RONALD W. SCHAFER, SENIOR MEMBER, IEEE, AND LAWRENCE R. RABINER, MEMBER, IEEE

Invited Paper

Abstract—This paper presents several digital signal processing methods for representing speech. Included among the representations are simple waveform coding methods; time domain techniques; frequency domain representations; nonlinear or homomorphic methods; and finally linear predictive coding techniques. The advantages and disadvantages of each of these representations for various speech processing applications are discussed.

I. Introduction

THE NOTION of a *representation* of a speech signal is central to almost every area of speech communication research. Often the form of representation of the speech signal is not singled out for special attention or concern but yet it is implicit in the formulation of a problem or in the design of a system. A good example of this situation is in telephony, where speech is, in fact, represented by fluctuations in electrical current for purposes of long distance transmission. In other situations, however, we must often pay strict attention to the choice and method of implementation of the representation of the speech signal. This is true, for example, in such diverse areas as speech transmission, computer storage of speech and computer voice response, speech synthesis, speech aids for the handicapped, speaker verification and identification, and speech recognition. In all of these areas, digital representations; i.e., representations as sequences of numbers, are becoming increasingly dominant. There are two basic reasons for this. First, through the use of small general purpose digital computers, speech researchers have been able to apply a wide variety of digital signal processing techniques to speech communication problems. These techniques cover a range of complexity and sophistication that is impossible to match with analog methods. Second, the recent and predicted future developments in integrated circuit technology make it possible to realize digital speech processing schemes economically as hardware devices having the same sophistication and flexibility as a computer program implementation.

The purpose of this paper is to survey the important and most useful methods for obtaining digital representations of speech signals. This is a formidable task since the number and variety of such methods is great. Thus we must begin by disclaiming any pretentions to completeness; we shall only try to point out the methods that in our view are the most useful in the technical and research areas of speech communication.

The organization of this paper is as follows. In Section II, we briefly review the speech production process and show how it can be modeled with a simple digital representation. We then discuss a class of waveform coding methods for representing speech in Section III. Included in this class are linear pulse-code modulation (PCM), delta modulation (DM), differential PCM, adaptive delta modulation, and finally adaptive differential PCM (DPCM). It is shown at the end of this section that if an adaptive predictor is incorporated in these models, the waveform coding technique becomes quite similar to the linear predictive coding method to be discussed in Section VII.

In Section IV, we discuss various time-domain representations of speech. Included in this section are the concepts of zero crossing analysis, autocorrelation functions, "peak-to-peak" type estimations, and the use of "energy" functions. In Section V, we discuss frequency domain representations of speech for which the concept of short-time spectrum analysis is dominant. Several examples of systems based on short-time spectrum analysis are given in this section.

In Section VI, we discuss the topic of homomorphic analysis of speech. In this section the concept of the cepstrum is introduced. Finally, in Section VII, we discuss the two basic methods of linear prediction analysis, explain their similarities and differences and discuss the basic concepts which are derivable from them including the spectrum, cepstrum, and autocorrelation function.

II. A Digital Model for Production of the Speech Signal [1]-[3]

A schematic diagram of the human vocal apparatus is shown in Fig. 1. The vocal tract is an acoustic tube that is terminated at one end by the vocal cords and at the other end by the lips. An ancillary tube, the nasal tract, can be connected or disconnected by the movement of the velum. The shape of the vocal tract is determined by the position of the lips, jaw, tongue, and velum.

Sound is generated in this system in three ways. Voiced sounds are produced by exciting the vocal tract with quasiperiodic pulses of air pressure caused by vibration of the vocal cords. Fricative sounds are produced by forming a constriction somewhere in the vocal tract, and forcing air through the constriction, thereby creating turbulence which produces a source of noise to excite the vocal tract. Plosive sounds are created by completely closing off the vocal tract, building up pressure, and then quickly releasing it. All these sources create a wide-band excitation of the vocal tract which in turn acts as a linear time-varying filter which imposes its transmission properties on the frequency spectra of the sources. The vocal tract can be characterized by its natural frequencies (or formants) which correspond to resonances in the sound transmission characteristics of the vocal tract.

A typical speech waveform is shown in Fig. 2, which illustrates some of the basic properties of the speech signal. We see, for example, that although the properties of the waveform change with time, it is reasonable to view the speech waveform as being composed of segments during which the signal properties remain

Manuscript received September 18, 1974; revised November 25, 1974.

R. W. Schafer was with the Bell Laboratories, Inc., Murray Hill, N.J. He is now with the Department of Electrical Engineering, Georgia Institute of Technology, Atlanta, Ga. 30332.

L. R. Rabiner is with the Bell Laboratories, Inc., Murray Hill, N.J. 07974.

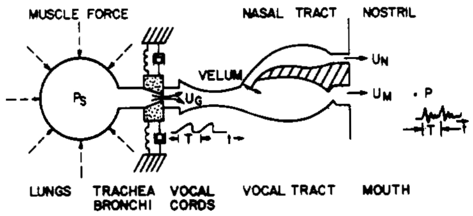


Fig. 1. Schematic diagram of mechanism of speech production. (After Flanagan *et al.* [2].)

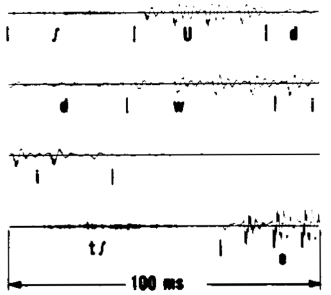


Fig. 2. An illustration of a speech waveform, corresponding to the utterance "*Should we chase*".

rather constant. Such segments are demarked in Fig. 2 below the waveform. These sample segments have the appearance either of a low-level random (unvoiced) signal (as in ʃ or tʃ in Fig. 2) or a high-level quasi-periodic (voiced signal) (as in *U* or *w* or *i*) with each period displaying the exponentially decaying response properties of an acoustic transmission system. We note that the dynamic range of the waveform is large; i.e., the peak amplitude of a voiced segment is much larger than the peak amplitude of an unvoiced segment.

Because the sound sources and vocal tract shape are relatively independent, a reasonable approximation is to model them separately, as shown in Fig. 3. In this digital model, samples of the speech waveform are assumed to be the output of a time-varying digital filter that approximates the transmission properties of the vocal tract and the spectral properties of the glottal pulse shape. Since, as is clear from Fig. 2, the vocal tract changes shape rather slowly in continuous speech (likewise its sound transmission properties) it is reasonable to assume that the digital filter in Fig. 3 has fixed characteristics over a time interval of on the order of 10 ms. Thus the digital filter may be characterized in each such interval by an impulse response or a set of coefficients for a digital filter. For voiced speech, the digital filter is excited by an impulse train generator that creates a quasi-periodic impulse train in which the spacing between impulses corresponds to the fundamental period of the glottal excitation.[1] For unvoiced speech, the filter is excited by a random number generator that produces flat spectrum noise. In both cases, an amplitude control regulates the intensity of the input to the digital filter.

This model is the basis of a wide variety of representations of speech signals. These are conveniently classified as either waveform representations or parametric representations depending upon whether the speech waveform is represented directly or whether the representation is in terms of time-varying parameters of the basic speech model. These representations range in complexity from simply samples of the speech waveform taken periodically in time to estimates of the parameters of the model in Fig. 3. The choice of the digital representation is governed by three major considerations: processing complexity, information (bit) rate, and flexibility. By complexity, we mean the amount of processing required to obtain the chosen representation. In many cases processing complexity is a measure of cost of implementation of the system in hardware. A simple representation can generally be implemented more economically than a complex representation. Thus complexity is often the overriding consideration in some transmission applications where low terminal cost is crucial. Information or bit rate is a measure of the redundancy in the speech signal which has been removed by the processing. A low bit rate means that the digital representation of the speech signal can be transmitted over a low capacity channel, or stored efficiently in digital memory. Finally flexibility is a measure of how the speech can be manipulated or altered for applications other than transmission, e.g., voice response, speech recognition, or speaker verification. In general, greater complexity is the price paid to lower the bit rate and increase the flexibility. However, tradeoffs can generally be made among these three factors. In transmission and voice response applications the quality and intelligibility of the reconstituted speech are also prime considerations. Most of the techniques we will discuss are capable of producing good quality, highly intelligible speech, although some of the techniques are primarily analysis methods, and as such are limited to applications where the speech signal need not be reconstructed.

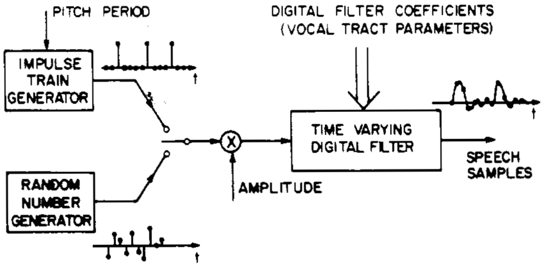


Fig. 3. Digital processing model for production of speech signals.

In the remainder of this paper, we will discuss a number of of digital representations that span the spectrum of possibilities in each of the above areas of concern. We shall begin with the simplest, least efficient and least flexible representation of speech and progress to more complex ones which have the greatest flexibility and lowest bit rate.

III. Digital Waveform Coding

Conceptually, the simplest digital representations of speech are concerned with direct representation of the speech waveform. Such schemes as PCM, DM, and DPCM are all based on Shannon's sampling theorem, which says that any bandlimited signal can be exactly reconstructed from samples taken periodically in time if the sampling rate is twice the highest frequency of the signal. We begin with a discussion of the simplest waveform coding technique; i.e., PCM.

A. PCM

In applying the sampling theorem to a digital representation of speech there are two main concerns. These are depicted in Fig. 4. If the signal bandwidth is W hertz, then the sampling period must be $T \leqslant 1/(2W)$. Since the samples $x(nT)$ of the signal generally take on a continuous range of values, they must be quantized for transmission or digital storage. If we repre-

[1] It is assumed that the effects of the glottal pulse shape are included in the digital filter.

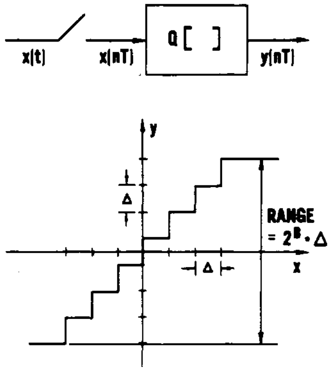


Fig. 4. Sampling and quantizing of an analog signal.

sent the samples as B-bit binary words, then the bit rate is $2BW$ bits/s. The value of W required for speech signals depends on the ultimate use of the samples. We know from measurements and theoretical studies that speech sounds such as fricatives have rather wide bandwidths (on the order of 10 kHz). On the other hand much of the information required for speech intelligibility is contained in the variation of the first three formant frequencies of voiced speech and these are typically below 3 kHz. Thus, a sampling rate between 6 kHz and 20 kHz is generally used. No matter what the sampling rate is, the speech signal must be suitably low-pass filtered prior to the sampling process to eliminate undesired high frequencies of the speech and high frequency noise.

The choice of the number of bits per sample B is also dependent upon the intended use of the samples. If our purpose is transmission or computer storage followed by conversion back to an analog signal, we are only concerned that the resulting analog signal be perceptually acceptable. Also, the sampling process just described is generally the first step in any digital speech analysis techniques. Since errors incurred in the sampling process will propagate to more refined digital representations, we are often justified in a very generous allotment of bits and sampling rate if the sampled speech wave is to undergo further processing. However it should be noted that the amount of processing required to implement most systems is proportional to sampling rate. Thus we should try to keep the sampling rate as low as possible, consistent with other objectives.

One objective measure of the fidelity of the PCM representation is the ratio of the signal power to the quantization noise power. If we define the quantization noise in Fig. 4 as the following:

$$e(nT) = x(nT) - Q[x(nT)] = x(nT) - y(nT)$$

then it can be shown [4], [7] that about 11 bits are required in order that the signal-to-noise ratio (SNR) be 60 dB. (This is often referred to as "toll quantity".) It is easily shown that the addition of one bit changes the SNR by 6 dB.

The preceding discussion can be summarized by stating that an adequate PCM representation for most purposes requires from 66 000 bits/s (11 bits × 6 kHz) to 220 000 bits/s (11 bits × 20 kHz). This is a very significant consideration in transmission or storage for processing on a computer.

Since we generally have little flexibility in lowering the sampling rate, as this is governed by other considerations, the main hope for lowering the overall bit rate is in reducing the number of bits/sample. The key to such reductions lies in considering one of the basic properties of the speech signal; namely, that speech has a wide dynamic range. We see from Fig. 4 that if B is fixed, then the step size Δ must be chosen so that $\Delta \cdot 2^B$ spans the maximum peak-to-peak range of the signal. Thus the quantizer step size is determined by the amplitude of the voiced segments of speech whereas a good representation of unvoiced segments requires a much smaller step size.

One solution to this problem is to use a nonlinear quantizer characteristic which distributes the quantization levels more densely for lower amplitudes than for high amplitudes. Based on empirical determinations of the amplitude distribution of speech signals, a logarithmic quantizer characteristic has been found to be nearly optimum [9]. Using a logarithmic quantizer, 7 bits/sample are sufficient to obtain toll quality. An alternative approach is the use of a time varying step size [5]-[7], i.e., an adaptive quantizer. When the signal level is low, a small step size is used; and when the signal amplitude is large, an appropriate large step size is used. The adjustment of the step size may be done by logical operations on the sequence of samples arising from adaptive quantization process [5]-[7].

B. Differential Quantization

Further reductions in bit rate for waveform quantization methods can be obtained by considering more of the detailed properties of the speech signal. Specifically, it is clear from Fig. 2 that there is a great deal of redundancy in the speech signal. Removal of some of this redundancy can yield a concomitant reduction in bit rate, at the expense of increased complexity in the signal processing algorithms. Fig. 5 depicts a general differential quantization scheme. The scheme is based on the fact that even for sampling at just the Nyquist rate ($T = 1/(2W)$), the correlation between successive samples is quite high and, as the sampling rate increases, the sample-to-sample correlation increases, approaching unity for very high sampling rates.

In the system of Fig. 5, let us assume that $\tilde{x}(n)$ is an estimate of the value of the speech sample $x(n) = x(nT)$. Then if the estimate is good, the variance of the difference $\delta(n) = x(n) - \tilde{x}(n)$ should be small, and thus the variance of the quantization error should be smaller than that incurred in quantizing the speech samples $x(n)$. The quantized difference signal $\hat{\delta}(n)$ when added to $\tilde{x}(n)$ produces a reconstructed signal $\hat{x}(n)$ which differs from $x(n)$ by only the quantization error of the difference signal; i.e.,

$$\begin{aligned} e(n) &= \delta(n) - \hat{\delta}(n) \\ &= [x(n) - \tilde{x}(n)] - [\hat{x}(n) - \tilde{x}(n)] \\ &= x(n) - \hat{x}(n). \end{aligned}$$

Due to the redundancy in the speech signal, it seems plausible that a given sample could be predicted as a linear combination of previous samples. In fact even the simplest linear combination may suffice; i.e., a constant times the previous sample. Therefore if the quantization error is small, $\hat{x}(n)$ will be a good approximation to $x(n)$ and

$$\tilde{x}(n) = a\hat{x}(n-1), \qquad a \approx 1 \tag{1}$$

will be a good estimate of $x(n)$. The z transform of (1) is

$$\tilde{X}(z) = az^{-1}\hat{X}(z).$$

Thus the predictor is characterized by the polynomial

$$P(z) = \frac{\tilde{X}(z)}{\hat{X}(z)} = az^{-1}. \tag{2}$$

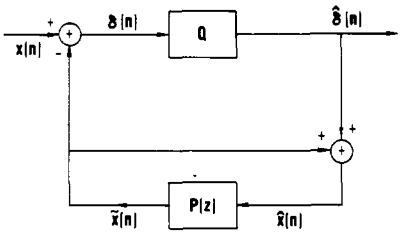
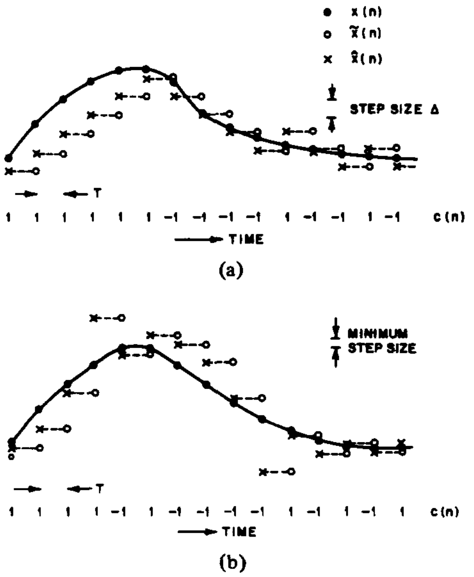


Fig. 5. General differential quantization scheme.



(a)

(b)

Fig. 6. Illustration of delta modulation. (a) Fixed step size. (b) Adaptive step size.

A more general predictor polynomial is of the form

$$P(z) = \sum_{k=1}^{p} a_k z^{-k}. \tag{3}$$

The basic principle of linear prediction is applied in more generality in Section VII.

In using differential quantization, we are free to choose the sampling rate, the quantizer and the predictor so as to reduce the bit rate. If the sampling rate is much higher than the Nyquist rate, the correlation between adjacent samples is very close to one and it is possible to use a 1-bit quantizer to obtain a good approximation to the input samples. This case, illustrated in Fig. 6(a), is called DM. In Fig. 6(a), we have illustrated how $x(n)$, $\tilde{x}(n)$, and $\hat{x}(n)$ vary with time. (We have shown the case where $a = 1$.) The quantized difference signal has the form

$$\hat{\delta}(n) = \Delta \cdot c(n)$$

where

$$c(n) = \begin{cases} +1, & \text{if } \delta(n) \geqslant 0 \\ -1, & \text{if } \delta(n) < 0 \end{cases}$$

and Δ is the fixed step size. Fig. 6(a) shows the two types of errors that are inherent in differential quantization schemes. On the left of the figure, the slope of the waveform is greater than the maximum rate of increase of the staircase approximation; i.e., for this choice of sampling period, Δ is too small to follow rapid changes in the waveform. This is called slope overload. On the right side of the figure, we see that in slowly varying parts of the waveform there is a tendency to oscillate up and down about the waveform. This is called granular distortion. In such regions we would like to have a smaller step size to reduce the magnitude of the quantization error.

One solution to this dilemma is to let the step size vary so that Δ becomes large during slope overload and small during granular distortion. This can be done by searching for patterns in the code word sequence $c(n)$. For example a run of +1's or −1's means slope overload, while an alternating pattern means granularity. A simple logic for varying the step size is [6]

$$\Delta(n) = \begin{cases} P\Delta(n-1), & \text{if } c(n) = c(n-1) \\ Q\Delta(n-1), & \text{if } c(n) \neq c(n-1). \end{cases}$$

The quantized difference signal is now

$$\hat{\delta}(n) = \Delta(n) \cdot c(n).$$

An optimum choice of the parameters is [6]

$$P = 1.5, \qquad Q = 1/P.$$

This scheme is illustrated by Fig. 6(b). (Here, for simplicity we have assumed $P = 2$ and $a = 1$.) It can be seen that this adaptive delta modulator (ADM) is able to follow rapid increases in slope and also it is able to use a smaller step size in regions of granularity. In practice, limits are placed on the step size variation so that $\Delta_{\min} \leqslant \Delta(n) \leqslant \Delta_{\max}$. This prevents the step size both from becoming unreasonably large and from being driven to zero when the input to the differential quantizer is zero.

If we use a multibit quantizer in Fig. 5, then a lower sampling rate can be used. This case is DPCM. If the sampling rate is the Nyquist rate, then we can use two bits less in the quantizer than required for straight PCM for the same SNR [8]. Furthermore, we can adapt the quantizer step size to obtain further improvements. Schemes similar to the ADM system just described have been implemented for multi-bit quantizers. These are called adaptive DPCM (ADPCM) systems [5].

Such a representation has been used for storage of speech at 24 kbits/s for a computer voice response system [10], [11]. An interesting result of this work is the observation that the adaptive quantizing provides a simple means of finding the beginning and end of a speech utterance [10]. This is a problem that arises in many situations, including speech recognition, speaker verification and computer voice response.

IV. Time Domain Analysis Methods

The objective of digital waveform coding is to represent the speech waveform as accurately as possible so that an acoustic signal can be reconstructed from the digital representation. In many speech processing problems, however, we are not interested in reconstructing an acoustic signal but rather we are concerned with representing the speech signal in terms of a set of properties or parameters of the model discussed in Section II. Some rather simple, but useful, characterizations can be derived by simple measurements on the waveform itself; i.e., upon a PCM representation of the waveform.

The key to these, and, indeed, the key to all parametric representations, is the concept of short-time analysis. We note from Fig. 2 that if we select an arbitrary segment of the speech waveform of about 10- to 30-ms duration, then it is quite probable that the properties of the waveform remain roughly invariant over that interval. For example, we may select a voiced interval in which the speech signal is characterized by the fundamental period and the amplitude of each basic period. On the other hand, we may select an unvoiced segment where the signal is characterized by the lack of periodicity and the amplitude of the waveform. Since these properties vary from segment-to-segment, it is common to analyze speech on a time-varying basis by carrying out an analysis on short segments of speech selected at uniformly spaced time intervals.

A. Peak Measurements

It is only necessary to glance at Fig. 2 to see that during voiced intervals, the speech signal is characterized by a sequence of peaks that occur periodically at the fundamental frequency of the speech signal. In contrast, during unvoiced intervals the peaks are relatively smaller and do not occur in any discernible pattern. Thus the maximum peak amplitude during an analysis interval can serve as a simple indication of the amplitude of the signal and as an aid in distinguishing between voiced and unvoiced speech segments.

The time between corresponding peaks is, of course, equal to the fundamental period for voiced speech. This principle has been used in a number of schemes for determining the fundamental period or pitch period. A difficulty with this approach is that even over a short analysis interval, the speech signal is not exactly periodic. Since each period has a number of peaks, it is possible to make several different estimates of the period. A method for logically combining the results of several simple measurements of this kind to improve accuracy has been discussed by Gold and Rabiner [12], [13]. By careful choice of the basic measurements and careful design of the logic, the accuracy of the combined results is much greater than the accuracy of any of the individual estimates.

B. Energy Measurements

One of the simplest representations of a signal is its energy. In the case of a real discrete-time signal $x(n)$, the energy is defined in general as

$$E = \sum_{n=-\infty}^{\infty} x^2(n). \tag{4}$$

For nonstationary signals such as speech, it is often more appropriate to consider a time-varying energy calculation such as the following:

$$E(n) = \sum_{m=0}^{N-1} [w(m)x(n-m)]^2 \tag{5}$$

where $w(m)$ is a weighting sequence or window which selects a segment of $x(n)$, and N is the number of samples in the window. For the simple case of $w(m) = 1$, $E(n)$ is the sum of the squares of the N most recent values of $x(n)$. Fig. 7(a) shows how the energy measurement of (5) can be viewed in terms of filtering the sequence $x^2(n)$ by a finite impulse response (FIR) filter with impulse response $w^2(n)$.

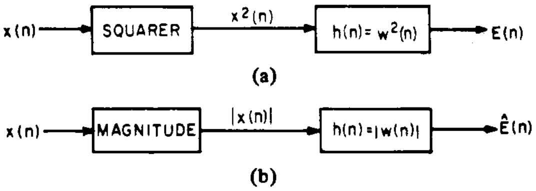


Fig. 7. (a) Implementation of short-time energy calculation using a finite impulse response digital filter. (b) An alternative definition of energy.

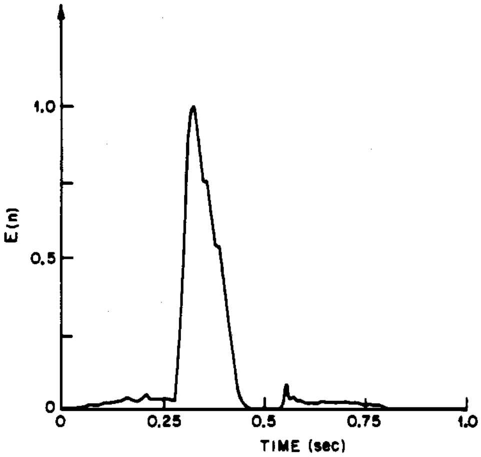


Fig. 8. Normalized energy for the word /six/.

It is to be expected that the function $E(n)$ would display the time varying amplitude properties of the speech signal. However, the definition of (5) requires careful interpretation. First there is the choice of window. The purpose of the window is to attach lower weight to speech samples which occurred further back in time, thus $w(m)$ generally tends to 0 monotonically as m gets larger. When one wants to apply equal weight to the entire interval, a rectangular window is used. The second difficulty involves choice of measurement interval N. If N is too small, i.e., less than a pitch period, $E(n)$ of (6) will fluctuate very rapidly depending on exact details of the waveform. If N is too large, i.e., several pitch periods, $E(n)$ will have very little variation, and will not reflect the changing properties of the speech signal. A suitable practical choice of N is on the order of 100–200 for a 10-kHz sampling rate (i.e., 10–20 ms of speech).

The major significance of $E(n)$ is that it provides a good measure for separating voiced speech segments from unvoiced speech segments. $E(n)$ for unvoiced segments is much smaller than for voiced segments. Also the smaller the value of N, the less smearing there is in locating the exact instant at which unvoiced speech becomes voiced and vice versa. Furthermore, for very high quality speech, the energy can be used to separate unvoiced speech from silence.

One difficulty with energy measurements is that they are very sensitive to large signal levels(because they enter the computation as a square), thereby emphasizing large sample-to-sample variations in $E(n)$. One relatively simple way of alleviating this problem is to use as a measure of energy, the function

$$\hat{E}(n) = \sum_{m=0}^{N-1} |w(m)x(n-m)| \tag{6}$$

where the sum of absolute values is computed instead of the sum of squares. Fig. 7(b) shows an interpretation of (6) as a linear filtering operation on $|x(n)|$. Fig. 8 shows the energy function for the word six for a 10-ms rectangular window. It is easy to see the low energy fricative regions at the beginning and end of six, and the stop gap region during the /*k*/ for which the energy is almost zero. An example of the application of energy measurements is the speech recognition work of Reddy [14].

C. Zero Crossing Measurements

Another very simple time domain analysis method is based on zero crossing measurements. In the context of a digital implementation, a zero crossing can be said to occur between sampling instants n and $n - 1$ if

$$\text{sign}\,[x(n)] \neq \text{sign}\,[x(n-1)]\,. \tag{7}$$

This measurement is trivial to implement and is often used as a gross estimate of the frequency content of a speech signal. Its use is motivated by the observation that if the signal is a sinusoid of frequency f_0, then the average number of zero crossings is

$$n_z = 2f_0 \text{ crossings/s.} \tag{8}$$

However, the interpretation of zero crossing measurements for speech is much less precise, because of the broad frequency spectrum of most speech sounds. Nevertheless, very crude estimates of spectrum properties such as this may often suffice.

For example, it is well known that the energy of voiced speech tends to be concentrated below 3 kHz, whereas the energy of fricatives generally is concentrated above 3 kHz. Thus, zero crossing measurements (along with energy information) are often used in making a decision about whether a particular segment of speech is voiced or unvoiced. If the zero crossing rate is high, the implication is unvoiced; if the zero crossing rate is low, the segment is most likely to be voiced. Zero crossing measurements, coupled with a pitch detection scheme, provide a useful approach to estimation of excitation parameters [34]. Zero crossing measurements have also been useful as representations of speech signals for speech recognition [14].

In implementing zero crossing measurements digitally, there are a number of important considerations. Although the basic algorithm requires only a comparison of signs of two successive samples, special care must be taken in the sampling process. Noise, dc offset, and 60-Hz hum have disastrous effects on zero crossing measurements. Thus for zero crossing measurements a bandpass filter rather than a low-pass filter may be necessary prior to sampling to avoid the said difficulties. Also, the sampling period T determines the time resolution of the zero crossing measurements; thus fine resolution requires a high sampling rate. However, very crude quantization (1 bit in fact) is all that is necessary to preserve the zero crossing information.

D. Short-Time Autocorrelation Analysis

The autocorrelation function of a discrete-time signal $x(n)$ is defined as

$$\varphi(m) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n)\,x(n+m)\,.$$

The autocorrelation function is useful for displaying structure in any waveform, speech being no exception. For example, if a signal is periodic with period P, i.e., $x(n+P) = x(n)$ for all n, then it is easily shown that

$$\varphi(m) = \varphi(m+P). \tag{9}$$

Thus periodicity in the autocorrelation function indicates periodicity in the signal. Also, an autocorrelation function that is sharply peaked around $m = 0$ and falls off rapidly to zero as m increases indicates a lack of predictable structure in the signal.

As we have observed, speech is not a stationary signal. However, the properties of the speech signal remain fixed over relatively long time intervals. As we have already seen, this leads to the notion of short-time analysis techniques that operate on short segments of the speech signal. For example consider a segment of N samples of the signal

$$x_l(n) = x(n+l), \qquad 0 \leqq n \leqq N-1 \tag{10}$$

where l denotes the beginning of the segment. Then the short-time autocorrelation function can be defined as

$$\varphi_l(m) = \frac{1}{N} \sum_{n=0}^{N'-1} x_l(n)x_l(n+m), \qquad 0 \leqq m \leqq M_0 - 1 \tag{11}$$

where M_0 denotes the maximum lag that is of interest. For example, if we wish to observe periodicity in a waveform, then we would require $M_0 > P$. The integer N' is for the moment unspecified.

We can interpret (11) as the autocorrelation of a segment of the speech signal of length N samples beginning at sample l. If $N' = N$, then data from outside the segment $l \leqslant n \leqslant N + l - 1$ is used in the computation. If $N' = N - m$, then only data from that interval is required. In this case, the segment is often weighted by a "window" function that smoothly tapers the ends of the segment to zero. In using the autocorrelation function to detect periodicity in speech, either choice is satisfactory; however, we shall see in Section VII that the distinction is important in analysis methods based on linear prediction. In either case, the direct computation of $\varphi_l(m)$ for $0 \leqq m \leqq M_0 - 1$ requires computational effort proportional to $M_0 \cdot N$. This can be a significant overhead factor.

Short-time analysis methods typically are applied to estimate parameters of the speech model discussed in Section II. The normal assumption is that although a sampling rate ranging from 6 kHz to 20 kHz may be necessary to preserve the essential features of the speech signal in a PCM representation, much lower sampling rates suffice for the slowly varying parameters of the model (50 to 100 Hz is typical). Suppose for example that the sampling rate of the speech signal is 10 kHz and the short-time autocorrelation is to be computed 100 times/s. The estimate of the autocorrelation is generally based upon from 20- to 40-ms segments of the speech signal. (For estimates of periodicity, the window must be long enough to encompass at least two periods of the speech signal.) Thus, for a 10-kHz sampling rate $200 \leqslant N \leqslant 400$, and the autocorrelation estimates must be computed by moving in increments of 100 samples.

In using the short-time autocorrelation function for pitch period estimation, it is desirable that the correlation function be sharply peaked so that a strong peak will stand out at multiples of P, the period. The correlation function of speech is not sharply peaked because there is a great deal of predictable structure in each period of the speech waveform. Sondhi [15] has given several methods of sharpening the peaks in the autocorrelation function. One of these called center clipping is illustrated in Fig. 9. The nonlinear operation of clipping out the middle of the speech waveform is very effective in reducing the sample to sample correlation of the signal. This is illustrated in Fig. 10 which shows a succession of short-time

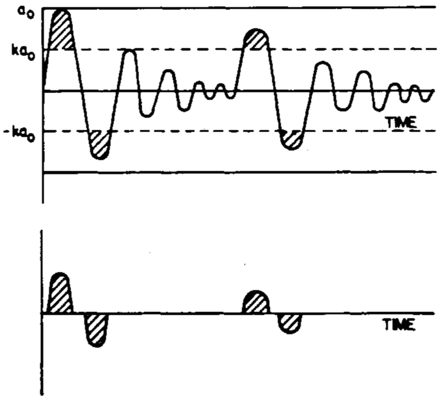


Fig. 9. Illustration of center clipping.

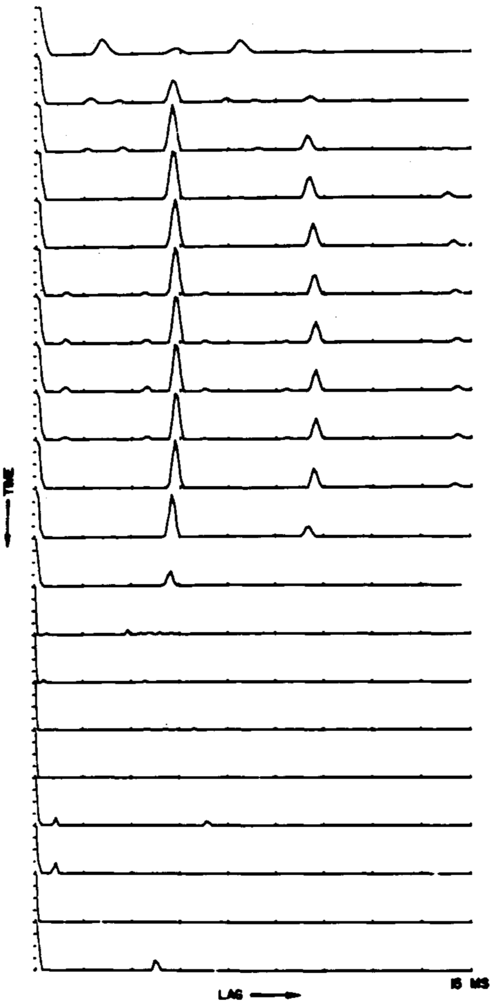


Fig. 10. Sequence of autocorrelation functions for center-clipped speech.

autocorrelation functions each estimated from 30-ms segments of center-clipped speech which are selected at intervals of 15 ms ($66\frac{2}{3}$-Hz sampling rate).

From a set of correlation functions of this type it is possible to estimate the pitch period simply by locating the strong peak that is in evidence during voiced intervals. Sondhi [15] gives a decision algorithm that formalizes this process. This scheme has been found to perform very well in situations where the speech is voiced but the wave shape is almost sinusoidal or when the fundamental frequency is missing [15].

V. Short-Time Spectrum Analysis

Short-time spectrum analysis has traditionally been one of the most important speech processing techniques. As we have previously stated, the fundamental assumption underlying any short-time analysis method is that over a long-time interval, speech is nonstationary but that over a sufficiently short-time interval it can be considered stationary. Thus, the Fourier transform of a short segment of speech should give a good spectral representation of the speech during that time interval. Measurement of the short-time spectrum is the basic operation in the channel vocoder [19], [26] the phase vocoder [18], spectrogram displays [21], [23], and some speech recognition systems [20]. Two methods are commonly used for implementing short-time Fourier analysis. The first uses a bank of bandpass filters. This method was originally used with analog filters and it can be implemented with even greater precision and flexibility with digital filters. The second method uses a fast Fourier transform (FFT) algorithm. This method is fundamentally digital and has no analog counterpart. When implemented on a computer, the FFT method is generally computationally superior to the bank-of-filters model.

A. Filter Banks for Short-Time Spectrum Analysis

Fig. 11 shows a simple way of implementing a short-time spectrum analyzer using a bank of bandpass filters. If the filter passbands are chosen to cover the speech band, then, roughly speaking, the outputs can be thought of as a Fourier representation of the input speech signal. If the filters are carefully designed, the sum of all the filter outputs will be a good approximation to the original speech signal [24]. This is the basis for communication systems such as the channel vocoder and the phase vocoder.

Based on some fundamental ideas of spectrum analysis, the discrete short-time spectrum of $x(n)$ is defined as

$$X_l(\omega) = \sum_{n=-\infty}^{l} x(n)\, h(l-n)\, e^{-j\omega n} \qquad (12a)$$

$$= |X_l(\omega)|\, e^{j\theta_l(\omega)} \qquad (12b)$$

$$= a_l(\omega) - jb_l(\omega). \qquad (12c)$$

Equation (12) can be interpreted in a number of ways. As shown in Fig. 12, one interpretation is that $X_l(\omega)$ is the Fourier transform of a sequence $x(n)$ that is weighted by a "window" $h(l-n)$. Thus the short-time Fourier transform is a function of both frequency ω and the discrete time index l. A second interpretation follows if we assume that $h(n)$ is the impulse response of a low-pass digital filter. Assume that we wish to evaluate the short-time transform at frequency ω. Then $X_n(\omega)$ is seen to be the output of the low-pass filter with input $x(n)\, e^{-j\omega n}$. This is depicted in Fig. 13(a). To avoid complex arithmetic, the system of Fig. 13(a) is generally implemented as shown in Fig. 13(b) where the output parameters are $a_n(\omega)$ and $b_n(\omega)$, the real and imaginary parts of the spectrum. The bandwidth of the low-pass filter determines the frequency resolution. Typically, this bandwidth is on the order of 50 Hz. Thus the spectrum signals can be sampled at a much lower rate (~100 Hz) than the speech signal itself.

Using digital filters, it has been shown [24], [25] that the short-time Fourier transform can be a very good representation of the speech signal in the sense that the output obtained by summing appropriately modulated bandpass channels can be made indistinguishable from the input. This requires a bit

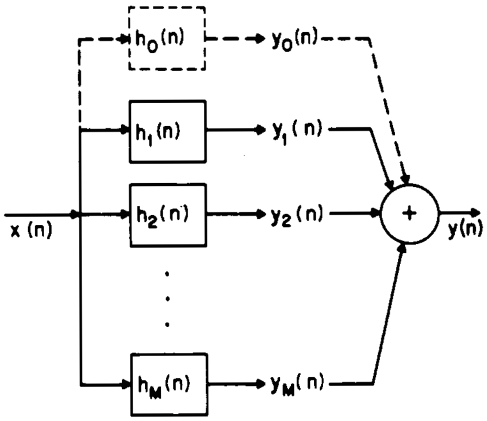


Fig. 11. A bank of bandpass filters.

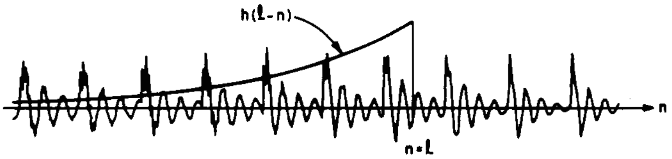


Fig. 12. Illustration of computation of the short-time Fourier transform.

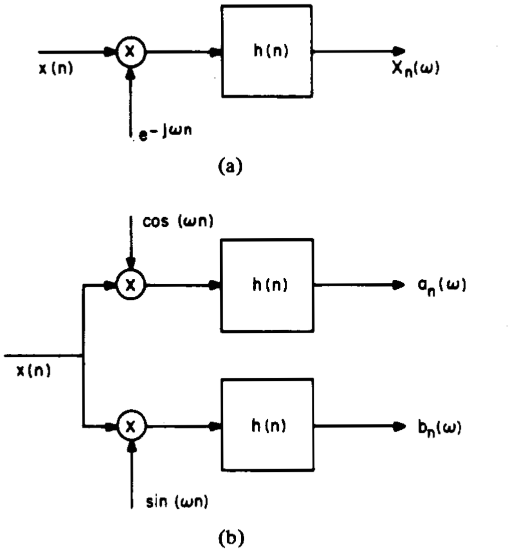


Fig. 13. Short-time Fourier analysis and synthesis for one channel centered at ω.

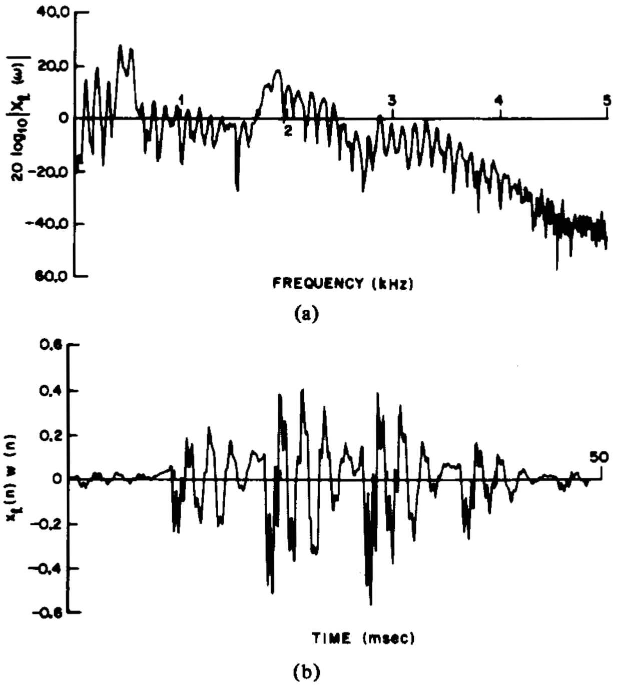


Fig. 14. (a) Log magnitude of the short-time transform. (b) Corresponding windowed speech segment. (N = 500.)

rate on the order of the bit rate required for comparable PCM representation. However, the resulting representation of the speech signal permits greater flexibility in the sense that the spectral parameters $a_n(\omega)$ and $b_n(\omega)$ provide information about the parameters of the speech model in a convenient and useful form. For example the time and frequency dimensions of a speech signal can be independently manipulated through simple manipulations of the spectral parameters [18].

B. Use of the FFT for Short-Time Spectrum Analysis

The FFT is a set of highly efficient algorithms for evaluating the discrete Fourier transform (DFT) expressions

$$F(k) = \sum_{n=0}^{M-1} f(n) \exp\left(-j\frac{2\pi}{M}kn\right), \qquad k = 0, 1, \cdots, M-1 \tag{13}$$

and

$$f(n) = \frac{1}{M}\sum_{k=0}^{M-1} F(k) \exp\left(j\frac{2\pi}{M}kn\right), \qquad n = 0, 1, \cdots, M-1. \tag{14}$$

For using these expressions, it is convenient to define the short-time transform as

$$X_l(\omega) = \sum_{n=0}^{N-1} x_l(n)\, w(n)\, e^{-j\omega n} \tag{15a}$$

where

$$x_l(n) = x(n+l), \quad n = 0, 1, \cdots, N-1, \quad l = 0, L, 2L, \cdots. \tag{15b}$$

As in the case of the short-time autocorrelation function, we interpret (15a) as the Fourier transform of a segment of speech N samples long (weighted by a window $w(n)$), beginning at l. The frequency resolution of the spectrum measurement is inversely proportional to the window length N. This is illustrated in Fig. 14. Fig. 14(a) shows the short-time transform and Fig. 14(b) shows the corresponding windowed segment of speech data. A Hamming window [17] of length 50 ms was used. (N = 500 samples at a 10-kHz sampling rate.) Note that the individual harmonics of the pitch period are resolved in the short-time transform. Figs. 15(a) and (b) show the short-time transform and the windowed speech for N = 50 samples. (The speech segment is the first 50 samples of the segment shown in Fig. 14(b).) In this case the frequency resolution is much less than in Fig. 14. We note that the spectrum of Fig. 14 could be considered comparable to a conventional narrow-band spectrogram measurement while Fig. 15 is comparable to a conventional wide-band spectrogram analysis. In particular, Figs. 14 and 15 show typical spectral cross-sections at a particular time. In the first case, both the pitch information and vocal tract transfer function information is present while in the latter case only the general shape of the vocal tract transfer function is preserved.

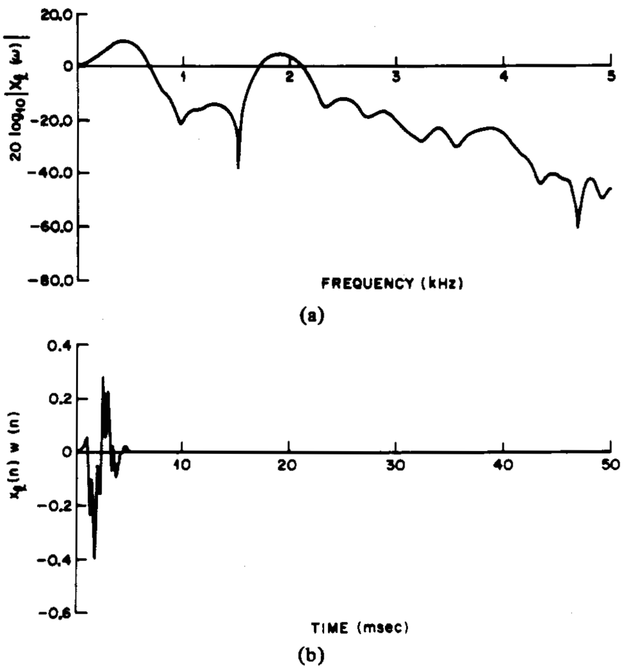


Fig. 15. (a) Log magnitude of the short-time transform. (b) Corresponding windowed speech segment. ($N = 50$.)

An FFT algorithm can be used to compute (15) at equally spaced frequencies $\omega_k = 2\pi k/M$, for $k = 0, 1, \cdots, M - 1$. If $M \geqslant N$, then the sequence $x_l(n)w(n)$ must be augmented with $M - N$ zero valued samples to form a sequence of length M. In this case we can compute

$$X_l\left(\frac{2\pi}{M}k\right) = \sum_{n=0}^{N-1} x_l(n)\,w(n)\,e^{-j2\pi kn/M}, \qquad k = 0, 1, \cdots, M - 1 \quad (16)$$

using an FFT algorithm.

On the other hand if $M < N$, we can take advantage of the periodicity of the complex exponential $\exp(-j2\pi kn/M)$ to express (15a) as

$$X_l\left(\frac{2\pi}{M}k\right) = \sum_{n=0}^{M-1} g(n)\,e^{-j2\pi kn/M}, \qquad k = 0, 1, \cdots, M - 1 \quad (17a)$$

where

$$g(n) = \sum_{r=0}^{[N/M]} x_l(n + r)w(n + r) \qquad (17b)$$

and $[N/M]$ means the largest integer in N/M. This latter feature of FFT spectrum analysis is useful whenever one wishes to only evaluate the transform at intervals of $\omega = 2\pi/M$ but at the same time wishes to obtain the better frequency resolution corresponding to a window of length N. Using the preceding approach, it is also possible to use the FFT to compute the outputs of a uniformly spaced bank of filters as required in a phase vocoder analyzer [24].

An important consequence of the definition of the short-time spectrum in (15) is that $|X_l(\omega)|^2/N$ is the Fourier transform of the short-time autocorrelation function

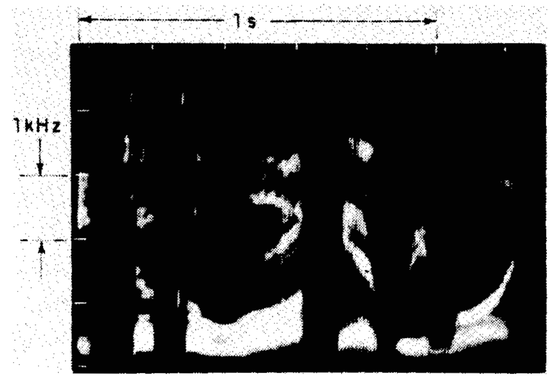


Fig. 16. An example of a spectrogram produced using digital spectrum analysis and computer graphics display. (After Oppenheim [23].)

$$R_l(m) = \frac{1}{N}\sum_{n=0}^{N-1-m} x_l(n)w(n)\,x_l(n + m)w(n + m). \quad (18)$$

That is,

$$R_l(m) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|X_l(\omega)|^2}{N}\,e^{j\omega m}\,d\omega. \qquad (19)$$

Furthermore, it can be shown that if $X_l(2\pi k/M)$ is computed with $M \geqslant 2N$, then $R_l(m)$ is the inverse of $|X_l(2\pi k/M)|^2/N$; i.e.,

$$R_l(m) = \frac{1}{M}\sum_{k=0}^{M-1} \frac{|X_l(2\pi k/M)|^2}{N}\,e^{\,j2\pi kn/M}, \qquad 0 \leqslant m \leqslant N - 1. \quad (20)$$

If we suppose that $R_l(m)$ is required for $0 \leqslant m \leqslant M_0 - 1$, where M_0 is a large number, as in pitch detection, it may be most efficient to first compute the short-time transform using (16), and then compute the autocorrelation function using (20).

C. Short-Time Spectrum Representations of Speech

The short-time spectrum can serve directly as a representation of the speech signal as is the case in many vocoder systems [18], [19], [25], [26] and in some speech recognition systems [20]. In many cases, however, the short-time spectrum is computed as an intermediate step in the estimation of one or more of the time varying parameters of the speech model. In the narrow-band short-time spectrum as in Fig. 14(a), both pitch and vocal tract transfer function information are clearly in evidence, while the wide-band analysis, as in Fig. 15(a), does not preserve the pitch information. Thus there are a variety of methods for estimating fundamental frequency directly from the narrow-band short-time spectrum [22], [27]. Similarly there are a wide variety of methods of estimating parameters such as formant frequencies from the short-time spectrum [16], [26].

One of the most useful tools in speech science is the sound spectrograph. This device produces a plot of energy as a function of time and frequency; i.e., a display of the short-time spectrum. The basis analysis techniques of this section have been used to generate spectrographic displays that are similar to, but in many cases more elaborate and flexible than, con-

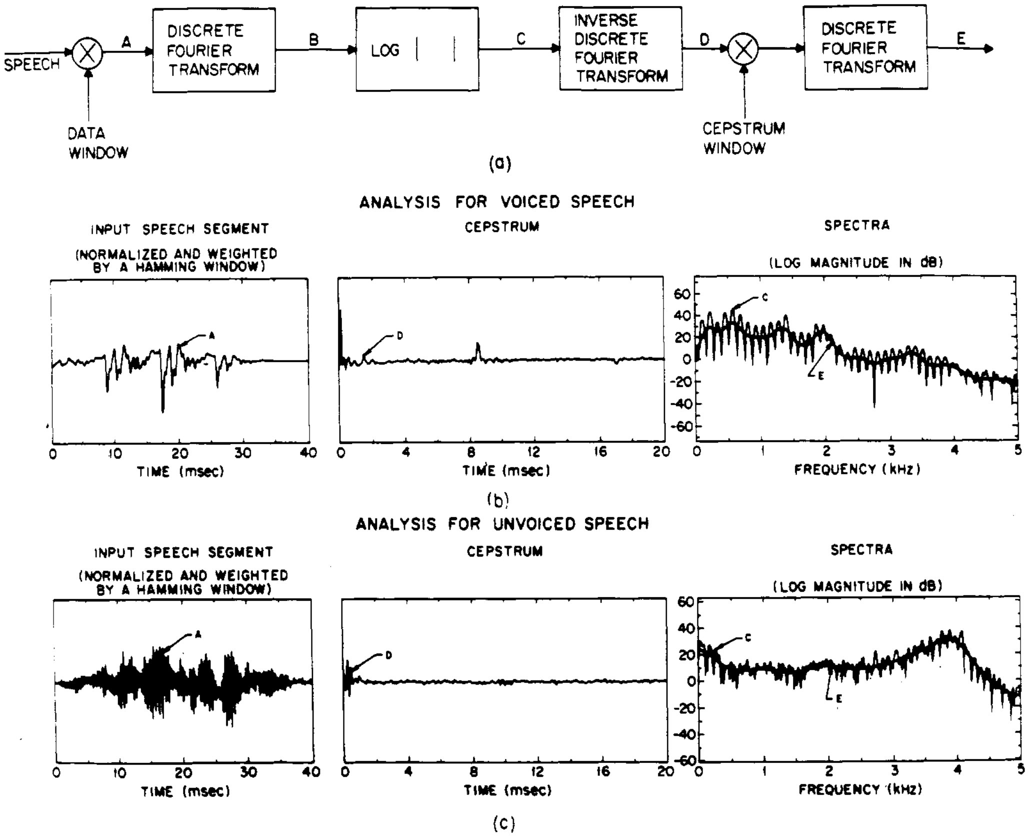


Fig. 17. Homomorphic processing of speech. (a) Basic operations. (b) Analysis for voiced speech. (c) Analysis for unvoiced speech.

ventional spectrograms [21], [23], [28]. As we have pointed out, there is great flexibility for computer spectral analysis in window length and shape or equivalently frequency resolution. Also, the spectrum can be shaped in a manner to enhance it for display, and it is possible to precisely correlate the speech waveform with the spectrographic display.

Such schemes have been implemented in a variety of ways but most of them use (15) to compute a set of short-time spectra at equally spaced time intervals. This set of spectra can be thought of as samples of the two dimensional function $X_l(\omega)$ which can be plotted as a frequency-time-intensity plot on an oscilloscope or television monitor. Using such techniques it has been possible to produce on-line spectrogram displays that are equal in quality to conventional spectrograms and far surpass them in flexibility and innovation. An example of one approach is shown in Fig. 16 [23].

VI. Homomorphic Speech Processing

Homomorphic filtering is a class of nonlinear signal processing techniques that is based on a generalization of the principle of superposition that defines linear systems. Such techniques have been applied in separating signals that have been combined by multiplication and convolution [31]. The application of these techniques to speech processing is again based on the assumption that although speech production is a time varying process, it can be viewed on a short-time basis as the convolution of an excitation function (either random noise or a quasi-periodic pulse train) with the vocal tract impulse response. Thus methods for separating the components of a convolution are of interest.

A. Fundamentals

A homomorphic system for speech analysis is shown in Fig. 17(a). We assume that the signal at A is the discrete convolution of the excitation and the vocal tract impulse response. Then the short-time Fourier transform (i.e., the spectrum of the windowed signal), computed using the FFT method of the previous section, is the product of the Fourier transforms of the excitation and the vocal tract impulse response. Taking the logarithm of the magnitude of the Fourier transform, we obtain at C the sum of the logarithms of the transforms of the excitation and vocal tract impulse response. Since the inverse discrete Fourier transform (IDFT) is a linear operation, the result at D (called the cepstrum of the input at A) is an additive combination of the cepstra of the excitation and vocal tract components. Thus, the effect of the operations, windowing, DFT, log magnitude, and IDFT is to approximately transform convolution into addition. The value of this transformation can be seen from Fig. 17(b), which depicts the results of such an analysis for voiced speech. The curve labeled A is the input speech segment that has been multiplied by a Hamming window. The rapidly varying curve labeled C is the log-magnitude of the short-time transform. It consists of a slowly varying component due to the vocal tract transmission, and a rapidly varying periodic component due to the periodic excitation. The slowly varying part of the log magnitude produces the low-time part of the cepstrum (D), and the rapidly varying periodic component of the log magnitude manifests itself in the strong peak at a time equal to the period of the input speech segment. If we assume that the vocal tract transfer function in the model of Fig. 3 is of the form of an all-pole model,

$$H(z) = \frac{A}{1 - \sum_{k=1}^{p} a_k z^{-k}} = A \prod_{k=1}^{p} \frac{1}{1 - z_k z^{-1}} \tag{21}$$

then the cepstrum of the vocal tract component of the convolution can be shown [30], [47] to be

$$\hat{h}(n) = \begin{cases} 0, & n < 0 \\ \log A, & n = 0 \\ \sum_{k=1}^{p} \frac{z_k^n}{n}, & n > 0. \end{cases} \tag{22}$$

If we assume that the excitation component is a periodic train of impulses, then it can be shown [30] that the cepstrum of the excitation component will also be a train of impulses with the same spacing as the input impulse train. This is clearly reflected in the cepstrum for voiced speech in Fig. 17(b). The important point is that the cepstrum consists of an additive combination in which (due to the $1/n$ falloff) the vocal tract and excitation components essentially do not overlap. The situation for unvoiced speech, shown in Fig. 17(c), is much the same with the exception that the random nature of the excitation component of the input speech segment (A) causes a rapidly varying random component in the log magnitude (C). Thus in the cepstrum (D), the low time components correspond as before to the slowly varying vocal tract transfer function; however, since the rapid variations of the log magnitude are not, in this case, periodic, there is no strong peak as for the voiced speech segment. Thus, the cepstrum serves as an excellent basis for estimating the fundamental period of voiced speech and for determining whether a particular speech segment is voiced or unvoiced [29].

The vocal tract transfer function, often called the spectrum envelope, can be obtained by removing the rapidly varying components of the log magnitude spectrum by linear filtering. One approach to this filtering operation involves computing the IDFT of the log magnitude spectrum (to give the cepstrum), multiplying the cepstrum by an appropriate window that only passes the short-time components, and then computing the DFT of the resulting windowed cepstrum. This method corresponds to the fast convolution method [45]-[49], in this case being applied to filter a function of frequency rather than a function of time. The results for voiced and unvoiced speech segments are labeled E in Figs. 17(b) and (c), respectively.

The smoothed spectrum obtained by the above method is in many respects comparable to a short-time spectrum obtained by direct analysis using a short data window. The major difference, however, is that the cepstrum method is based upon the initial computation of a narrow-band spectrum, which involves a wide time window, while the wide-band spectrum is computed using a very narrow-time window. The smoothing is done upon a narrow-band log-magnitude spectrum rather than upon the short-time Fourier transform itself, as is the case for wide-band analysis. Thus, for speech segments in which the basic parameters such as pitch period and formant frequencies are not changing, we should expect the cepstrum method to produce superior results to direct spectrum analysis. When the speech spectrum is changing rapidly, as in the case of a voiced/unvoiced boundary, the direct method may produce a better representation than the cepstrum method due to its shorter averaging time.

B. Estimation of Formant Frequencies and Pitch Period

The results depicted in Fig. 17 suggest algorithms for estimating basic speech parameters such as pitch period and formant frequencies. Specifically, voiced/unvoiced classification of the excitation is indicated by the presence or absence of a strong peak in the cepstrum [29]. The presence of a strong peak for voiced speech is dependent upon there being many harmonics present in the spectrum. In cases where this is not true, such as voiced stops, zero crossing measurements are helpful in distinguishing voiced from unvoiced speech [34]. If a strong peak is present, its location is a good indicator of the pitch period.

The smoothed spectrum retains peaks at the vocal tract resonances or formant frequencies. One approach to estimating the formants is to search the smooth spectra for peaks and then decide which peaks correspond to formants [34]. Another approach uses iterative methods to adjust the parameters of a model similar to (21) until a good match to the smooth spectrum is obtained [33].

An illustration of the use of homomorphic processing is given in Fig. 18. On the left are shown a sequence of cepstra computed at 20-ms intervals. The strong peak indicates that the speech is voiced during the entire interval. On the right are successive short-time spectra and homomorphically smoothed short-time spectra. The lines connecting the peaks of the smooth spectra show the formant frequencies automatically estimated from the spectrum peaks. The peak-picking approach is relatively simple except when two formants merge as in the third and fourth frames from the top and the last 4 frames from the bottom. In this case it is useful to evaluate the vocal tract transfer function on a contour which passes closer to the poles thereby sharpening the resonances [34].

Speech can be synthesized from formant and pitch data by using the estimated parameters to vary the parameters of the model of Fig. 3. With efficient coding of the parameters, speech is thus represented by about 1000 bits/s [2]. In addition to this high efficiency, the formant representation offers great flexibility in manipulating basic speech parameters. Also, since so much of the speech model is built into the representation, these parameters are very useful for other purposes such as speech recognition and speaker verification.

C. The Cepstrum as a Representation of Speech

The low-time samples of the cepstrum contain mostly information about the vocal tract transfer function $H(z)$ of (21). It can be shown [31], [47], that the following recurrence formula relates the vocal tract impulse response $h(n)$ to the cepstrum $\hat{h}(n)$ of (22):

$$h(n) = \begin{cases} \hat{h}(n)h(0) + \sum_{k=0}^{n-1} \left(\frac{k}{n}\right) \hat{h}(k)h(n-k), & 1 \leqq n \\ e^{\hat{h}(0)}, & n = 0. \end{cases} \tag{23}$$

Also using (23) it is easily shown that the coefficients a_n in (21) are related to the cepstrum by

$$a_n = \hat{h}(n) - \sum_{k=0}^{n-1} \left(\frac{k}{n}\right) \hat{h}(k) a_{n-k}, \qquad 1 \leqq n \leqq p. \tag{24}$$

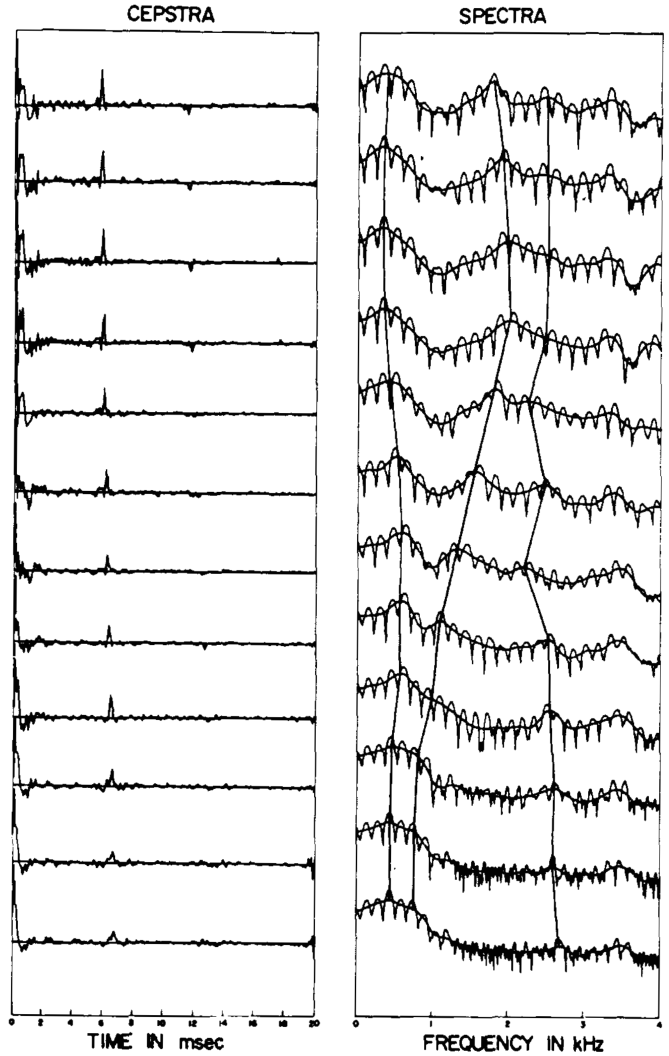


Fig. 18. Cepstra and spectra for a region of voiced speech.

Since the cepstrum contains all of the information of the short-time spectrum, it can be viewed as still another representation of the speech signal. This principle has been applied in a speech analysis synthesis scheme called the homomorphic vocoder [32]. In this system, the low-time cepstrum values and an estimate of pitch period serve as a representation of the speech signal from which an acoustic wave can be reconstructed.

VII. Linear Predictive Analysis

Among the most useful methods of speech analysis are those based upon the principle of linear prediction. These methods are important because of their accuracy and their speed of computation. In this section, we present a formulation of linear predictive analysis and discuss some of the issues which are involved in using it in practical speech applications.

The basic idea behind linear predictive coding (LPC) is that a sample of speech can be approximated as a linear combination of the past p speech samples. By minimizing the square difference between the actual speech samples and the linearly predicted ones, one can determine the predictor coefficients; i.e., the weighting coefficients of the linear combination. The basic philosophy of this scheme is reminiscent of and, in fact, related to the waveform quantization methods discussed in Section III [35]. There it was mentioned that a linear predictor can be applied in a differential quantization scheme to reduce the bit rate of the digital representation of the speech waveform. In this case, as in linear predictive analysis, the predictor coefficients must be adapted (i.e., updated regularly) to match the time-varying properties of the speech signal.

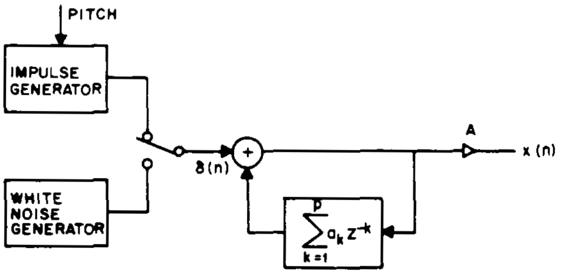


Fig. 19. Digital model for speech production.

A. Fundamental Principles

The use of linear predictive analysis is suggested by the digital model of Section II. Assume that samples of the speech signal are produced by the model of Fig. 3, where over a short time interval the linear system has the transfer function

$$H(z) = \frac{A}{1 - \sum_{k=1}^{p} a_k z^{-k}}. \tag{25}$$

For voiced speech, the system is excited by an impulse train and for unvoiced speech it is excited by random white noise as depicted in Fig. 19. Linear prediction analysis is based on the observation that for such a system the speech samples $x(n)$ are related to the excitation $\delta(n)$ by the following difference equation:

$$x(n) = \sum_{k=1}^{p} a_k x(n-k) + \delta(n). \tag{26}$$

Suppose that we process the speech signal with a linear predictor; i.e.,

$$\tilde{x}(n) = \sum_{k=1}^{p} \alpha_k x(n-k).$$

Then the predictor error is defined as

$$\epsilon(n) = x(n) - \tilde{x}(n) = x(n) - \sum_{k=1}^{p} \alpha_k x(n-k). \tag{27}$$

Note that in this case the prediction is based on the unquantized samples $x(n)$, whereas in Section III, the prediction was based on quantized samples $\hat{x}(n)$. It can be seen by comparing (26) and (27) that if $\alpha_k = a_k$, and if the speech signal really does obey the model of (26), then $\epsilon(n) = \delta(n)$. Therefore, between the excitation impulses of voiced speech, the prediction error should be very small if the predictor coefficients α_k are equal to the parameters a_k of the vocal tract transfer function. Thus the predictor polynomial

$$P(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k}$$

is a good approximation to the denominator of the vocal tract transfer function.[2]

One approach for obtaining the predictor coefficients is based on minimizing the average squared prediction error over a short segment of the speech waveform. That is, we search for the values of α_k that minimize

$$E_l = \sum_{n=0}^{N-1} (x_l(n) - \tilde{x}_l(n))^2$$

$$= \sum_{n=0}^{N-1} \left(x_l(n) - \sum_{k=1}^{p} \alpha_k x_l(n-k) \right)^2 \qquad (28)$$

where $x_l(n)$ is a segment of speech that has been selected in the vicinity of sample l; i.e.,

$$x_l(n) = x(n+l).$$

There are two basic ways of choosing $x_l(n)$ each leading to procedures that are somewhat different in the details of their implementation and the results that are obtained. Leaving $x_l(n)$ unspecified for now, we can find the values of α_k that minimize E_l in (28) by setting $\partial E_l / \partial \alpha_i = 0$, $i = 1, 2, \cdots, p$, thus obtaining the equations

$$\sum_{n=0}^{N-1} x_l(n-i) x_l(n) = \sum_{k=1}^{p} \alpha_k \sum_{n=0}^{N-1} x_l(n-i) x_l(n-k), \qquad 1 \leqslant i \leqslant p. \qquad (29)$$

If we define

$$\varphi_l(i,k) = \sum_{n=0}^{N-1} x_l(n-i) x_l(n-k) \qquad (30)$$

then (29) can be written more compactly as

$$\sum_{k=1}^{p} \alpha_k \varphi_l(i,k) = \varphi_l(i,0), \qquad i = 1, 2, \cdots, p. \qquad (31)$$

This set of p equations in p unknowns can be solved for the unknown predictor coefficients that minimize the average squared prediction error for the segment $x_l(n)$. To do this, the quantities $\varphi_l(i,k)$ must be computed for $1 \leqslant i \leqslant p$ and $1 \leqslant k \leqslant p$. The details of this computation depend upon how $x_l(n)$ is defined.

By a simple substitution of variables, (30) can be written as

$$\varphi_l(i,k) = \sum_{n=-i}^{N-1-i} x_l(n) x_l(n+i-k)$$

$$= \sum_{n=-k}^{N-1-k} x_l(n) x_l(n+k-i). \qquad (32)$$

Clearly, $\varphi_l(i,k) = \varphi_l(k,i)$. We observe from (32) that values of $x_l(n)$ are required outside the interval $0 \leqslant n \leqslant N-1$. If we choose to supply the values outside this interval we note that we then require

$$x_l(n) = x(n+l), \qquad -p \leqslant n \leqslant N-2 \qquad (33)$$

to evaluate $\varphi_l(i,k)$. This method and its attendant details was proposed by Atal [36] and has come to be called the *covariance* method because of the similarity of the matrix $\varphi_l(i,k)$ to a covariance matric.

If we choose not to supply values of the signal outside the interval $0 \leqslant n \leqslant N-1$, then we must resort to using a finite duration window $w(n)$ to reduce the end effects thereby obtaining,

$$x_l(n) = \begin{cases} x(n+l)w(n), & 0 \leqslant n \leqslant N-1 \\ 0, & \text{otherwise.} \end{cases}$$

Using this definition of $x_l(n)$, (32) becomes

$$\varphi_l(i,k) = \sum_{n=0}^{N-1-(i-k)} x_l(n) x_l(n+i-k).$$

$$= \sum_{n=0}^{N-1-(k-i)} x_l(n) x_l(n+k-i)$$

$$\equiv r_l(i-k) = r_l(k-i). \qquad (34)$$

In this case (31) becomes

$$\sum_{k=1}^{p} \alpha_k r_l(|i-k|) = r_l(i), \qquad i = 1, 2, \cdots, p. \qquad (35)$$

From (34) and (18), it is clear that $r_l(n) = NR_l(n)$; i.e., $r_l(n)$ is equal (to within a constant multiplier) to the short-time autocorrelation function, which in turn is related to the short-time Fourier transform $X_l(\omega)$. Thus the method based on (35) is called the *autocorrelation method.* Methods of this type have been proposed by Itakura [38] (the maximum likelihood method) and Markel [41]-[43] (the inverse filter formulation).

The basic difference between the covariance method and the autocorrelation method is the necessity to use a window for the autocorrelation method. For the covariance method the section length is increased by augmenting p samples to enable the first p samples of the section ($x_l(n)$, $0 \leqslant n \leqslant p-1$) to be predicted from speech samples outside the section. Thus an equal number of samples go into the computation of $\varphi(i,j)$ for all indices i and j, and no window is required. For the autocorrelation method one is trying to predict the first p samples from speech samples outside the section. Since these samples are arbitrarily zero, a large error may result. To reduce the error a window is applied which smoothly tapers the signal to zero at the ends of the window.

At this point it is worth noting the mathematical and physical interpretations of using windows in the autocorrelation method. The process of multiplication of a signal by a window is equivalent to a circular convolution of the frequency response of the window with the speech spectrum. Thus a smearing occurs in the speech spectrum. The extent of this smearing depends on the section length N and the actual window used. However, it is clear that with the autocorrelation method, parameters such as formant bandwidths may not be accurately estimated. In many practical applications this is of little or no consequence; however, for vocoder applications it may be significant.

B. Details of Implementation

Both (31) and (35) are a set of p equations in p unknowns that can be expressed in matrix form as

$$\Phi \cdot a = \Psi. \qquad (36)$$

These equations may be solved for the predictor coefficients using any general procedure for solving linear equations. How-

[2] The effects of the glottal pulse shape are included in the predictor polynomial.

ever, if computational efficiency is important, as it usually is, some special properties of the matrix Φ can be exploited to reduce computation. In the case of (31) (the covariance method) Φ is symmetric and positive definite. Utilization of this fact leads to an efficient procedure for solving for the vector a of predictor coefficients that is based on matrix factorization. This method is called the square root method, or the Cholesky decomposition [37].

Similarly, for the autocorrelation method the matrix Φ is symmetric and positive definite and also has the property that the elements along any diagonal are equal. Such a matrix is called a Toeplitz matrix and in this case an even more efficient method for solving the equations can be found [43]. This method is called the Levinson method.

Since computational efficiency is an important consideration in any practical speech analysis scheme, it is worthwhile comparing these two methods of linear prediction in this sense. The square root method for solving the covariance method formulation requires on the order of p^3 operations (multiplications) whereas the Levinson method for solving the autocorrelation formulation requires on the order of p^2 operations. Thus the solution of the equation for the autocorrelation formulation is inherently faster computationally than for the covariance formulation. In particular, for $p = 14$, Makhoul and Wolf [39] note a ratio in computation time of 3.2 to 1 in favor of the autocorrelation method. However, this savings in computation is not significant when viewed in the total framework of the method for two reasons. First the time required to compute the matrix of correlations is significantly greater than the time to solve the matrix equation. For example, for $N = 150$, Makhoul and Wolf [39] note that it takes ten times longer to compute the matrix then to solve the matrix equations using the autocorrelation method. Thus the savings in computation of the Levinson method becomes much less significant. As a second consideration the value of N required for both methods is not the same. For the autocorrelation method (for 10-kHz sampling) a value of N in the range 150 to 300 is generally required. For the covariance method a much smaller value of N can be used if care is taken to begin the section after a pitch pulse. In fact, Atal reports using values of N on the order of 30 with good results [36]. Thus there are many factors which determine computational efficiency.

Another difference between the two methods concerns the roots of the predictor polynomial which are the poles of the digital filter that accounts for the vocal tract transmission properties. For stability of this system, the roots must be inside the unit circle of the z plane. This is not guaranteed by the covariance method [36]; however, given *sufficient computational accuracy* the autocorrelation method guarantees stability [39], [43].

Another consideration in using these two methods is the numerical stability of the matrix inversion. Wilkinson [44] has shown that the square-root method is very stable numerically; no such statement has been made for the Levinson method. Markel [43] has pointed out that when implemented with finite precision arithmetic, the Levinson method requires careful scaling, and it is beneficial if the speech spectrum has been equalized by a simple first-order network.

Until now we have dealt with considerations which can be easily quantified and for which definitive statements can be made. When one becomes seriously interested in using linear predictive methods, several other considerations are involved. These include the necessity for spectrum equalization prior to analysis; the effects of the analog prefilter prior to analog-to-digital (A/D) conversion; the effects of finite word length on the analysis; the desirability of various structures for implementing the system; and finally the ease of building the various alternatives in digital hardware. Markel [43] has provided some excellent insights into several of these issues but most of them are as yet unresolved.

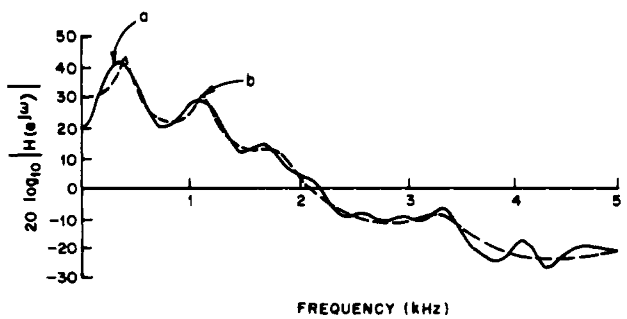


Fig. 20. Comparison of speech spectra. (a) Obtained by cepstrum smoothing. (b) Obtained by linear prediction.

C. Uses of Linear Prediction Analysis

Once the predictor coefficients have been obtained, they can be used in various ways to represent the properties of the speech signal.

1) Spectrum Estimation: If the predictor polynomial is assumed to represent the denominator of the vocal tract transfer function, we can obtain the frequency response of the vocal tract (for a particular segment of the speech signal) as

$$H(e^{j\omega T}) = \frac{A}{1 - \sum_{k=1}^{p} \alpha_k e^{-j\omega kT}}. \tag{37}$$

An example is shown in Fig. 20, where the spectrum obtained using (37) with the predictor coefficients estimated by the autocorrelation method is compared to that obtained by cepstrum smoothing for the same segment of speech. The formant frequencies are clearly in evidence in both plots, however, Fig. 20(b) has fewer extraneous peaks. This is because p was chosen so that at most 6 ($p = 12$) resonance peaks could occur. To determine the appropriate value of p for a given sampling rate, a good rule of thumb is to allow one pair of poles to account for radiation and glottal effects, and one pair of poles for each formant frequency expected in the frequency range $0 \leqslant \omega \leqslant \pi/T$. Thus, for a 10-kHz sampling rate we expect not more than 5 formant frequencies so $p = 12$ should give a good representation of the spectrum. For unvoiced speech it has been shown that a reasonably small prediction error can be obtained with a value of p on the order of 12 [36], [43].

Another point to notice is that the spectrum peaks in Fig. 20(a) are much broader than the peaks in Fig. 20(b). This is an inherent property of the homomorphic method since the Fig. 20(a) was obtained by smoothing the short-time log spectrum.

2) Formant Frequency Estimation: Smooth spectra such as Fig. 20(b) have been used in a peak picking algorithm to estimate formant frequencies in much the same manner as spectra such as Fig. 20(a) were used [41].

If p is chosen as discussed here, it can be assumed that the roots of the predictor polynomial will in general correspond to the formant frequencies. These roots can be obtained by factoring the predictor polynomial. An example is shown in Fig. 21. It is clear by comparing the plot of Fig. 21(b) to the spec-

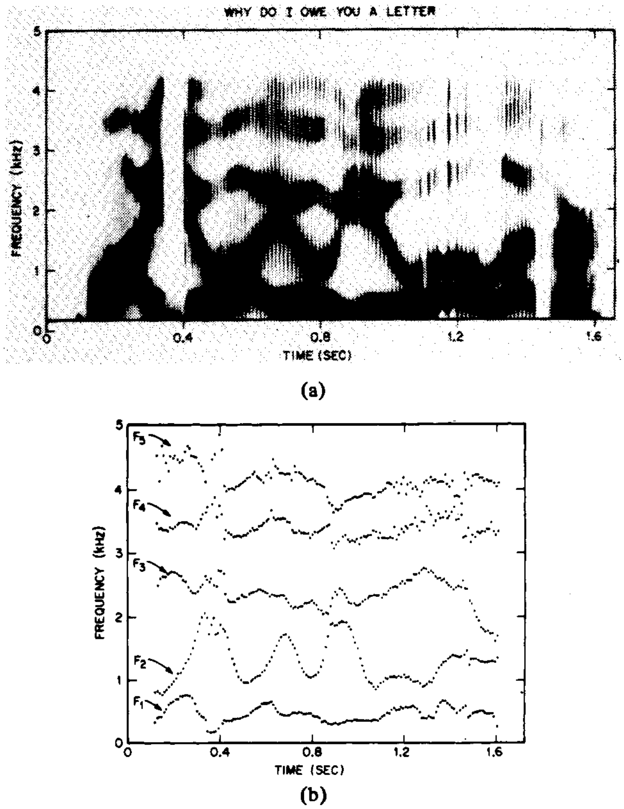


Fig. 21. (a) Spectrogram of predictor polynomial. (b) Roots of predictor polynomial (after Atal [36]).

trogram that the roots of the predictor polynomial are generally very good estimates of the formant frequencies. As with all formant analysis problems the difficulty in the problem lies in giving a particular formant label to a pole. Several reliable algorithms exist for doing this job [34], [41].

3) Pitch Detection: We recall that if we use the predictor coefficients as in our original formulation, then the prediction error

$$\epsilon(n) = x(n) - \sum_{k=1}^{p} \alpha_k x(n-k) \tag{38}$$

should appear very much like the excitation function $\delta(n)$ in Fig. 19. Thus it might be expected that the prediction error signal might be useful as a starting point for determining properties of the excitation; i.e., pitch period and voiced/unvoiced decision. Several procedures of this type have been suggested [40], [42].

4) Relation to the Cepstrum and Autocorrelation Function: In addition to the aforementioned, the basic linear prediction coefficients can be transformed directly into a number of other representations of the speech signal. For example by solving (24) for $\hat{h}(n)$, we obtain the recurrence formula

$$\hat{h}(n) = a_n + \sum_{k=0}^{n-1} \left(\frac{k}{n}\right) \hat{h}(k) a_{n-k} \tag{39}$$

relating the cepstrum of the vocal tract impulse response to the coefficients a_n in (25). Similarly it can be shown [35] that the autocorrelation function of the vocal tract impulse response defined as

$$r(m) = \sum_{n=0}^{\infty} h(n)h(n+m) \tag{40}$$

satisfies the recurrence formula

$$r(m) = \begin{cases} \sum_{k=1}^{p} a_k r(i-k), & m \geqq 1 \\ \sum_{k=1}^{p} a_k r(k) + 1, & m = 1. \end{cases} \tag{41}$$

5) Speech Synthesis: Finally, the predictor coefficients and excitation information can be used in the model of Fig. 19 to reconstruct a speech waveform [36]. In this case it is necessary to estimate the constant A in (25) as well as the parameters of the predictor polynomial. This can be done as part of the computation of the predictor coefficients [43] but in most cases A is simply chosen to match the energy of the synthetic speech to the energy of the original speech [36].

D. Discussion

The underlying structure of linear prediction analysis is that over short sections of speech one can accurately predict the current speech sample from the preceding p samples. Although a wide variety of different formulations of this method have arisen, the inherent similarities between methods are much larger than the supposed differences. To make all the decisions as to which particular method to use, what section duration etc., one must pay strict attention to the ultimate application of the method. Thus for most speech recognition applications, for example, the differences between formulations are *not* significant. For other more stringent applications, such as analysis/synthesis, the differences may indeed be quite significant and may mean the difference between an acceptable and a nonacceptable system.

VIII. Summary

In this paper, we have discussed a wide variety of digital representations of speech signals. These representations have varied in complexity, information rate, and flexibility from simple waveform coding schemes to analysis schemes such as homomorphic filtering and linear prediction analysis which are directed toward the estimation of the parameters of a detailed model of speech production. We have focused our attention almost exclusively on analysis techniques that are of wide applicability. The results of most of these techniques can be applied in a variety of speech processing applications including speech recognition, speech synthesis, and speaker verification.

References

General

[1] G. Fant, *Acoustic Theory of Speech Production.* The Hague, The Netherlands: Mouton, 1970.

[2] J. L. Flanagan, C. H. Coker, L. R. Rabiner, R. W. Schafer, and N. Umeda, "Synthetic voices for computers," *IEEE Spectrum,* vol. 7, pp. 22-45, Oct. 1970.

[3] J. L. Flanagan, *Speech Analysis, Synthesis and Perception,* 2nd ed. New York: Springer-Verlag, 1972.

Waveform Coding

[4] H. S. Black, *Modulation Theory.* Princeton, N.J.: Van Nostrand, 1953.

[5] P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive quantization in differential PCM coding of speech," *Bell Syst. Tech. J.*, pp. 1105-1118, Sept. 1973.

[6] N. S. Jayant, "Adaptive delta modulation with a one-bit memory," *Bell Syst. Tech. J.*, pp. 321-342, Mar. 1970.

[7] —, "Digital coding of speech waveforms," *Proc. IEEE.*, vol. 62, pp. 611-632, May 1974.

[8] R. A. McDonald, "Signal-to-noise and idle channel performance of DPCM systems—particular application to voice signals," *Bell Syst. Tech. J.*, pp. 1123-1151, 1966.

[9] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inform. Theory*, vol. 1T- pp. 7-12, Mar. 1960.

[10] L. H. Rosenthal, R. W. Schafer, and L. R. Rabiner "An algorithm for locating the beginning and end of an utterance using ADPCM coded speech," *Bell Syst. Tech. J.*, vol. 53, pp. 1127-1135, July-Aug. 1974.

[11] L. H. Rosenthal, L. R. Rabiner, R. W. Schafer, P. Cummiskey, and J. L. Flanagan, "A multiline computer voice response system utilizing ADPCM coded speech," *IEEE Trans. Acoust., Speech, and Sig. Processing*, vol. ASSP-22, pp. 339-352, Oct. 1974.

Time-Domain Methods

[12] B. Gold, "Note on buzz-hiss detection," *J. Acoust. Soc. Amer.*, vol. 36, pp. 1659-1661, 1964.

[13] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. Amer.*, vol. 46, no. 2, pp. 442-449, Aug. 1969.

[14] D. R. Reddy, "Computer recognition of connected speech," *J. Acoust. Soc. Amer.*, vol. 42, no. 2, pp. 329-347, Aug. 1967.

[15] M. M. Sondhi, "New methods of pitch detection," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 262-266, June 1968.

Short-Time Spectrum Analysis

[16] 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.*, vol. 33, pp. 1725-1736, Dec. 1961.

[17] R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra*. New York: Dover, 1959.

[18] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell Syst. Tech. J.*, vol. 45, pp. 1493-1509, Nov. 1966.

[19] B. Gold and C. M. Rader, "Systems for compressing the bandwidth of speech," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, pp. 131-135, Sept. 1967; and "The channel vocoder," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, pp. 148-160, Dec. 1967.

[20] T. Martin, "Acoustic recognition of a limited vocabulary in continuous speech," Ph.D. dissertation, Univ. Pennsylvania, Philadelphia, 1970. (Available from Univ. Microfilms, Ann Arbor, Mich.)

[21] P. Mermelstein, "Computer generated spectrogram displays for on-line speech research," *IEEE Trans. Audio Electroacoust.*, vol. AU-19, pp. 44-47, Mar. 1971.

[22] A. M. Noll, "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate," in *Computer Processing in Communications Proceedings*, J. Fox, Ed. New York: Polytechnic Press, 1969.

[23] A. V. Oppenheim, "Speech spectrograms using the fast Fourier transform," *IEEE Spectrum*, vol. 7, pp. 57-62, Aug. 1970.

[24] R. W. Schafer and L. R. Rabiner, "Design of digital filter banks for speech analysis," *Bell Syst. Tech. J.*, vol. 50, no. 10, pp. 3097-3115, Dec. 1971.

[25] —, "Design and simulation of a speech analysis-synthesis system based on short-time Fourier analysis," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 165-174, June 1973.

[26] M. R. Schroeder, "Vocoders: Analysis and synthesis of speech," *Proc. IEEE*, vol. 54, pp. 720-734, May 1966.

[27] M. R. Schroeder, "Period histogram and product spectrum: New methods for fundamental-frequency measurement," *J. Acoust. Soc. Amer.*, vol. 43, no. 4, pp. 829-834, Apr. 1968.

[28] H. R. Silverman and N. R. Dixon, "A parametrically controlled spectral analysis system for speech," *IEEE Trans. Acoustics, Speech, and Sig. Processing*, vol. ASSP-22, pp. 362-381, Oct. 1974.

Homomorphic Speech Analysis

[29] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293-309, Feb. 1967.

[30] A. V. Oppenheim and R. W. Schafer, "Homomorphic analysis of speech," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 221-226, June 1968.

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

[32] A. V. Oppenheim, "A speech analysis-synthesis system based on homomorphic filtering," *J. Acoust. Soc. Amer.*, vol. 45, pp. 458-465, Feb. 1969.

[33] J. Olive, "Automatic formant tracking in a Newton-Raphson technique," *J. Acoust. Soc. Amer.*, vol. 50, pt. 2, pp. 661-670, Aug. 1971.

[34] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Amer.*, vol. 47, no. 2, pp. 634-648, Feb. 1970.

Linear Prediction Analysis

[35] B. S. Atal and M. R. Schroeder, "Adaptive predictive coding of speech signals," *Bell Syst. Tech. J.*, vol. 49, 1970.

[36] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pt. 2, pp. 637-655, Aug. 1971.

[37] D. K. Faddeev and V. N. Faddeeva, *Computational Methods of Linear Algebra*. San Francisco, Calif.: Freeman, 1963.

[38] F. Itakura and S. Saito, "An analysis-synthesis telephony system based on maximum likelihood method," *Electronics Commun. Japan*, vol. 53A, pp. 36-43, 1970.

[39] J. I. Makhoul and J. J. Wolf, "Linear prediction and the spectral analysis of speech," Bolt, Beranek, and Newman Inc., Boston, Mass., BBN Rep. 2304, Aug. 31, 1972.

[40] J. N. Maksym, "Real-time pitch extraction by adaptive prediction of the speech waveform," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 149-153, June 1973.

[41] J. D. Markel, "Digital inverse filtering—A new tool for formant trajectory estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 129-137, June 1972.

[42] J. D. Markel, "The sift algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367-377, Dec. 1972.

[43] J. D. Markel, A. H. Gray, Jr., and H. Wakita, "Linear prediction of speech-theory and practice," Speech Communications Res. Lab., Santa Barbara, Calif., SCRL Monograph 10, Sept. 1973.

[44] J. H. Wilkinson, *Rounding Errors in Algebraic Processes*. Englewood Cliffs, N.J.: Prentice-Hall, 1963.

Digital Signal Processing

[45] B. Gold and C. M. Rader, *Digital Processing of Signals*. New York: McGraw-Hill, 1969.

[46] H. D. Helms, "Fast Fourier transform method of computing difference equations and simulating filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, no. 2, pp. 85-90, June 1967.

[47] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

[48] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

[49] T. G. Stockham, Jr., "High speed convolution and correlation," *AFIPS Proc.*, pp. 229-233, 1966.

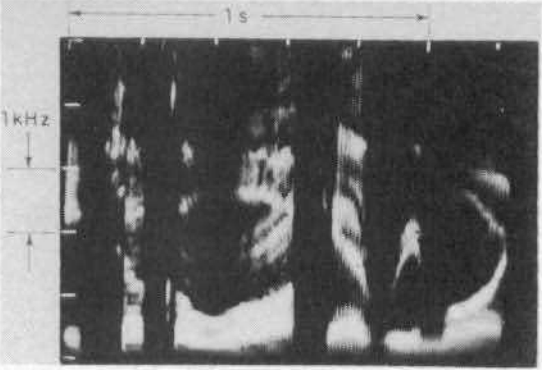
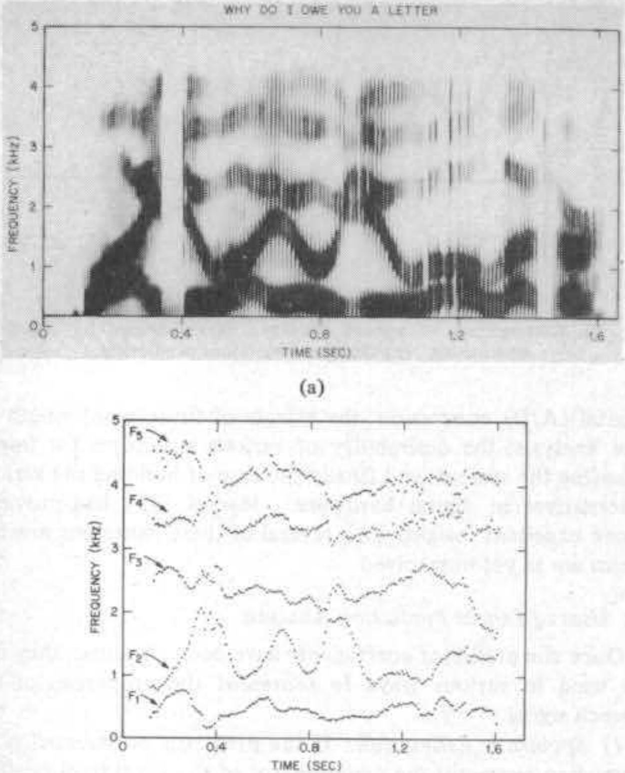


Fig. 16. An example of a spectrogram produced using digital spectrum analysis and computer graphics display. (After Oppenheim [23].)



(a)

(b)

Fig. 21. (a) Spectrogram of predictor polynomial. (b) Roots of predictor polynomial (after Atal [36]).