reason barring the simulation of most extant aids for the deaf within the software. The unit cost of the resultant system may therefore be less than a complete set of hard-wired training aids. Further, there is reason to believe that such a system, providing several displays within a simplified common framework, will have a beneficial effect greater than the sum of its parts.

Acknowledgment

The authors wish to thank the other members of the project team, D. Klatt, K. Stevens, D. Dodds, and T. Willemain, for their contributions. They also thank B. Noel for monitoring the experiments, K. Pearsons and S. Fidell for aiding the data analysis, and R. Nickerson for useful comments.

References

[1] P. R. Drouilhet, Jr., and L. M. Goodman, "Pole-shared linear-phase band-pass filter bank," *Proc. IEEE* (Lett.), vol. 54, pp. 701–703, Apr. 1966.

[2] R. M. Lerner, "Band-pass filters with linear phase," *Proc. IEEE*, vol. 52, pp. 249–268, Mar. 1964.

[3] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Amer.*, vol. 24, pp. 175–185, 1952.

[4] K. N. Stevens and G. von Bismarck, "A nineteen-channel filter bank spectrum analyzer for a speech recognition system," NASA Scientific Rep. 2, Bolt Beranek and Newman, Inc., 1967.

A Survey of Digital Speech Processing Techniques

RONALD W. SCHAFER, Member, IEEE
Bell Telephone Lab., Inc.
Murray Hill, N. J. 07974

Abstract

Digital signal processing techniques are becoming increasingly important in speech analysis and synthesis. These techniques can be implemented using a general purpose computer facility (often not in real time), or special purpose hardware realizations can be constructed. This paper discusses some recent work in speech processing including design of digital filter bank spectrum analyzers, homorphic analyzers of speech, predictive coding, and hardware realization of a digital formant synthesizer. The survey concentrates on those speech processing techniques relevant to the development of sensory aids for the deaf.

I. Introduction

Techniques for speech analysis and synthesis have been the subject of research for many years. This research has been stimulated by two basic concerns. One is the need for a more efficient and flexible representation of speech than the acoustic waveform. The other is the desire for a true understanding of the phenomenon of speech. In recent years, digital signal processing techniques have become increasingly important both in speech analysis–synthesis and in basic research, and they will attain an even greater importance in the future.

With the availability of small, high-speed, general-purpose digital computers, the computer has become the principle laboratory facility in speech research [1], [2]. From this increased use of computers it has become clear that speech analysis and synthesis systems that are realized using a digital computer program should not be viewed as a simulation of an analog system. In fact, many techniques that can be easily realized digitally would be difficult if not impossible to realize with analog equipment. Modern integrated circuit technology makes it possible to consider realizing many digital signal processing systems as special-purpose hardware, and the technology is presently available for constructing very complex real-time signal processors at speech sampling rates. Thus, it is now more reasonable to view digital computer programs as simulations of systems that can be realized with *digital* hardware.

The advantages of digital speech analysis systems are many. Several important ones are the following.

1) It is possible to realize very complex and flexible analysis and synthesis systems, and thus obtain very good representations of the speech signal.

2) Digital systems can be designed so that they are highly reliable, stable, compact, lightweight, and low in power consumption.

3) Very efficient (in terms of hardware) realizations can be designed because it is relatively easy to multiplex digital hardware.

These reasons are equally valid whether one is concerned with systems for transmission or storage of speech, learning about the speech process, or with educational or communication aids for people with visual or auditory handicaps. In fact, it seems that digital signal processing should have a great impact on the latter field.

In this paper we will examine a number of digital speech processing systems so as to illustrate the advantages of digital

Manuscript received November 5, 1971.

This paper was presented at the IEEE Electronics Conference, Chicago, Ill., October 1971.

signal processing. As we proceed, we may speculate on possible applications of the techniques in the area of aids to the handicapped. We shall conclude with some comments on experience we have had with hardware realization of a complex speech processing system.

II. Digital Speech Processing Techniques

In this section we discuss a wide range of digital speech processing techniques. We begin by considering a model of speech production based on digital filtering principles.

Digital Model for Speech Production [3], [4]

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 quasi-periodic pulses of air pressure caused by vibration of the vocal cords. Fricative sounds are produced by forming a constriction somewhere in the vocal tract, forcing air through the constriction, thereby creating turbulence that 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 that acts as a time-varying filter to impose its transmission properties on the frequency spectra of the sources. The vocal tract is characterized by its natural frequencies (or formants) that correspond to resonances in the sound transmission characteristics of the vocal tract.

Because the sound sources and vocal tract shape are relatively independent, a reasonable approximation is to model them separately, as shown in Fig. 2. In this digital model, samples of the speech wave are assumed to be the output of a time-varying digital filter that approximates the transmission properties of the vocal tract. Since the vocal tract changes shape rather slowly in continuous speech, and likewise its sound transmission properties, it is reasonable to assume that the digital filter in Fig. 2 has fixed characteristics over a time interval of about 10 ms. Thus the digital filter may be characterized in each such interval by an impulse response or a set of coefficients for a recursive 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 uniform random number generator that produces flat spectrum noise. In both cases, an amplitude control regulates the intensity of the input to the digital filter.

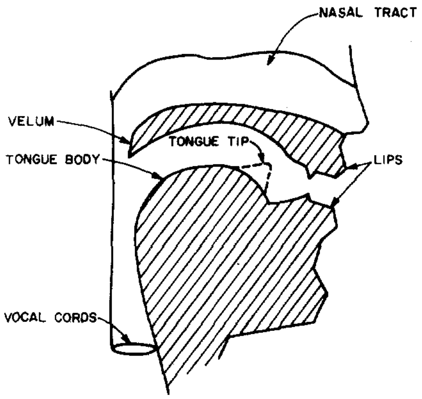

Fig. 1. Schematic drawing of human vocal tract.

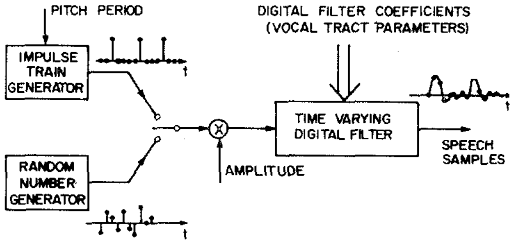

Fig. 2. Model for speech production based on digital signal processing principles.

Taking this model as a representation of the speech process, the problem of speech analysis is simply that of estimating the parameters of the model from the sampled speech waveform. When the parameters of the model are specified as time functions, the model also serves as a basis for speech synthesis. The basic concerns in speech analysis and synthesis are the efficiency and the flexibility of the representation of the speech wave. In the remainder of this section we discuss a number of recently developed analysis–synthesis techniques and comment on the efficiency and flexibility that each scheme affords. We begin with short-time spectrum analysis—a basic technique in several digital signal processing systems.

Short-Time Spectrum Analysis

For discrete-time signals, the short-time Fourier transform takes the form

$$X(\omega, nT) = \sum_{r=-\infty}^{n} x(rT)w(nT - rT)e^{-j\omega rT}$$

$$-\frac{\pi}{T} \leq \omega \leq \frac{\pi}{T} \qquad (1)$$

where T is the sampling period. We note that for a given value of ω, $X(\omega, nT)$ is a sequence defined on the discrete time index nT, while at a given time, $X(\omega, nT)$ is a continuous periodic function of ω with period $2\pi/T$.

The short-time Fourier transform of a sequence x is the Fourier transform of the part of the sequence seen through

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

the window w. The real and imaginary parts of $X(\omega, nT)$ can be obtained by multiplying the sequence $x(nT)$ by cos (ωnT) and sin (ωnT), respectively, and then passing the resulting sequences through digital low-pass filters with impulse responses equal to w. Thus, measurements of $X(\omega, nT)$ at a finite set of frequencies can be obtained with a bank of digital filters.

Alternatively, if w is chosen as a time-limited window of duration N samples, then (1) can be efficiently evaluated at a set of frequencies

$$\omega_k = \frac{2\pi}{NT} k, \qquad k = 0, 1, \cdots, N-1,$$

using a fast Fourier transform (FFT) algorithm.[2]

Both the filter bank and FFT methods of short-time Fourier analysis can be realized either with special-purpose hardware or as programs on a general-purpose digital computer. Spectrum estimation using the FFT offers the advantage of flexibility in the choice of analysis window and, for programmed spectrum analyzers, the FFT is generally faster than a filter bank analyzer for anything but the coarsest frequency resolution. On the other hand, a filter-bank spectrum analyzer may be preferable for special-purpose hardware realizations.

The Phase Vocoder

The phase vocoder [5] is a speech analysis–synthesis system that is based directly upon short-time spectrum analysis. We can write the short-time spectrum as

$$X(\omega, nT) = |X(\omega, nT)| \exp[j\phi(\omega, nT)]. \qquad (2)$$

If $X(\omega, nT)$ is evaluated at a set of $M+1$ frequencies ω_k that cover the frequency range of the signal, it can be shown that a sequence $\tilde{x}$ that approximates the input sequence x can be synthesized using the equation[3]

$$\tilde{x}(nT) = X(0, nT) + \sum_{k=1}^{M} |X(\omega_k, nT)| \cdot \cos[\omega_k nT + \phi(\omega_k, nT)]. \qquad (3)$$

Analysis of the phase vocoder system in terms of the previously proposed model is difficult. It is approximately true that if the number of channels is large, then the phase signals contain mostly information about the excitation, while the amplitude signals depend both on the vocal tract transmission properties and the source spectrum [5]. The spectrum magnitude sequences $|X(\omega_k, nT|$ and the phase derivative sequences $\dot{\phi}(\omega_k, nT)$ can be band-limited to achieve a bandwidth saving of about 2:1 [5]. For digital transmission or storage, a total information rate of about 7200 bits/s has been achieved.

In addition to its improved efficiency for transmission, the phase vocoder offers some degree of flexibility in manipulating the basic speech parameters [5]. The frequency range of the signal can be compressed or expanded without affecting the time scale by simply multiplying the phase signals by a factor A, and then synthesizing using (3) with frequencies $A\omega_k$. For example, if $A = \frac{1}{2}$, a 3-kHz frequency band would be compressed into a band of 1.5 kHz. Alternatively, some of the channels could be transposed to a lower frequency band, leaving others unaffected. Such flexibility could be employed in studying the possibilities of frequency transposing aids to the deaf that are of a more sophisticated nature than those that have previously been used [7].

Another feature of the phase vocoder system is the possibility of compression or expansion of the time scale. In order to achieve a time scale compression, we can first record a signal in which the frequency scale has been compressed as above using a value of A that is less than 1. Then the resulting signal can be played back $1/A$ times faster, restoring the original frequency range, but reducing the time duration by a factor $1/A$. Such a system could be applied to produce natural sounding speeded speech for "talking books" for the blind. The complementary situation of time expanded speech can likewise be achieved by recording a signal in which the frequency range is expanded by a factor A greater than one and then playing it back at a rate A times slower. In both cases, A is not limited to integer values.

Homomorphic Processing of Speech

A digital signal processing technique that exploits the properties of the speech model to a greater extent than the phase vocoder is called homomorphic filtering [8], [9]. The theory of homomorphic systems is based on a generalization of the principle of superposition and has been applied in separating signals combined by convolution and multiplication [9]. Homomorphic filters for convolved signals can be used to separate the excitation and vocal tract components of speech waveforms. This is because the parameters of speech vary slowly with time so that a short section of the speech waveform is approximated by a convolution of the excitation waveform and the vocal tract impulse response.

A homomorphic system for estimating the excitation parameters and the vocal tract transmission properties is shown in Fig. 3(A). If we assume that the signal at A is a discrete convolution of an excitation signal and a vocal tract impulse response, then the discrete Fourier transform (DFT) at B is the product of their respective DFT's.[4] Taking the logarithm of the magnitude of the DFT, we obtain at C the

[2] The FFT is a class of computational algorithms for efficiently evaluating the discrete Fourier transform that is defined as

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

[3] Recent work [6] indicates that significant improvement in performance can be achieved by adding a term $\omega_k n_0 T$ to the argument of the cosine in (3).

[4] Since fine frequency resolution is required, the FFT is used to evaluate the short-time transform.

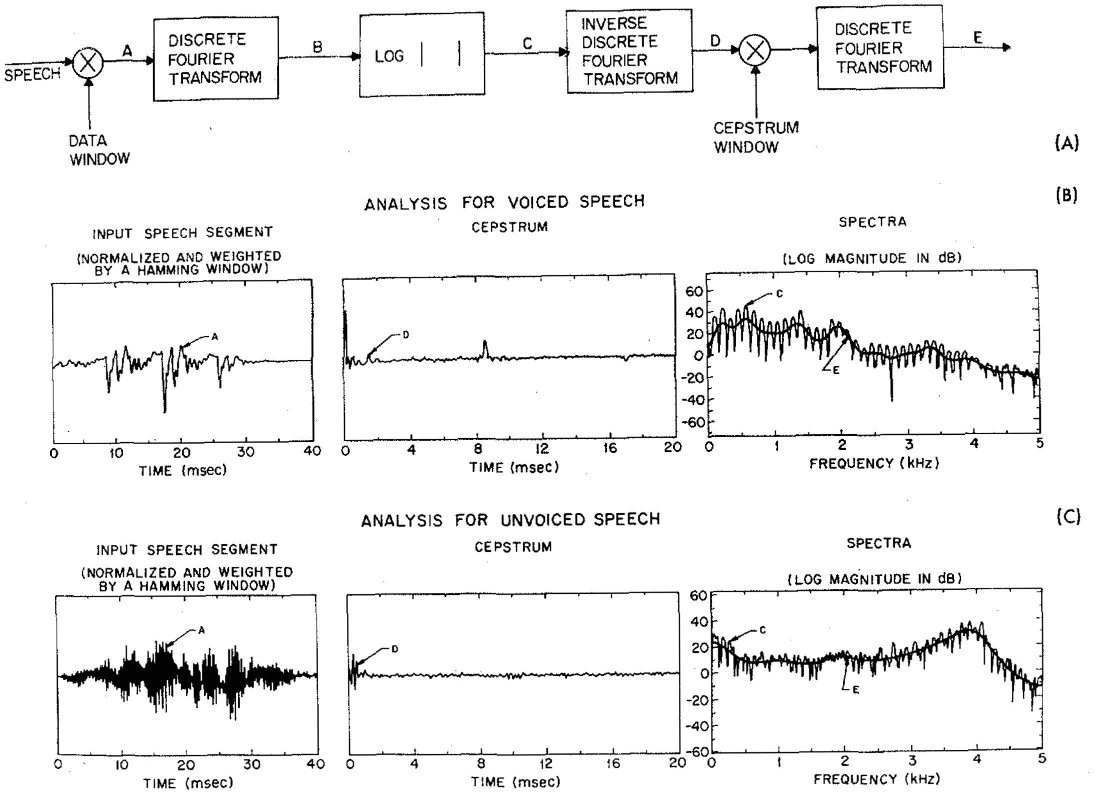

Fig. 3. Homomorphic processing of speech. (A) Basic operations. (B) Analysis for voiced speech segments. (C) Analysis for unvoiced speech segments.

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 DFT, log magnitude, and IDFT is to transform convolution into addition. The value of this transformation can be seen from Fig. 3(B), which depicts the results of such an analysis for voiced speech. The curve labeled A is the input speech section that has been multiplied by a data window that improves the estimate of the short-time spectrum. 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 the period of the input speech segment. The important result is that the cepstrum consists of an additive combination in which the vocal tract and excitation components essentially do not overlap. The situation for unvoiced speech, shown in Fig. 3(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 fundamental period of voiced speech and for determining whether a particular speech segment is voiced or unvoiced [10].

The vocal tract transfer function (spectrum envelope) can be obtained by removing the rapidly varying components of the log magnitude by linear filtering. This can be achieved by multiplying the cepstrum by a "cepstrum window" that only passes the short-time components. The results for voiced and unvoiced speech are labeled E in Figs. 3(B) and (C), respectively.

Formant Analysis and Synthesis: The techniques discussed above have been employed in a system for automatically estimating the vocal tract parameters, pitch period, and amplitude as required in the model of Fig. 2 [3], [11], [12]. Specifically, pitch period is estimated from the cepstrum and the formant frequencies are estimated from the location of resonance peaks in the spectrum envelope. For unvoiced speech, a single complex pole and zero are sufficient to characterize the vocal tract transmission, and these are likewise estimated from the smooth spectrum envelope.

This approach is straightforward except when two resonances become so close together that two distinct peaks are not in evidence in the spectrum envelope. In this case, a new spectrum analysis algorithm called the chirp z-transform [11] allows us to evaluate the transfer function in a manner that sharpens the resonances and produces two distinct peaks. Recently an alternative approach was proposed that uses homomorphic filtering to obtain the spectrum envelope, and then estimates formant frequencies by an iterative spectrum matching procedure [12].

Speech can be synthesized from formant and pitch data obtained in the above manner by using the configuration of Fig. 2, with the digital filter being a cascade of second-order time-varying recursive digital filters. Specifically, the system is characterized by the transfer function

$$V(z) = S(z) \cdot \prod_{k=1}^{M} \left[\frac{G_k}{1 - z^{-1}2e^{-\alpha_k T} \cos(2\pi F_k T) + z^{-2}e^{-2\alpha_k T}} \right] \tag{4}$$

for voiced speech, and for unvoiced speech

$$U(z) = S(z) \frac{G_p(1 - z^{-1}2e^{-\beta T} \cos(2\pi F_z T) + z^{-2}e^{-2\beta T})}{G_z(1 - z^{-1}2e^{-\beta T} \cos(2\pi F_p T) + z^{-2}e^{-2\beta T})} \tag{5}$$

where $S(z)$ is a fixed system that provides appropriate spectrum shaping and is specified by

$$S(z) = \frac{(1 - e^{-aT})(1 + e^{-bT})}{(1 - e^{-aT}z^{-1})(1 + e^{-bT}z^{-1})} \tag{6}$$

and G_p, G_z, and the G_k are constants chosen to produce unity gain at zero frequency. Only the first three formant frequencies F_1, F_2, and F_3 are estimated, and the remaining formant frequencies, bandwidths, and the parameters a and b are held fixed at appropriate values. With efficient coding of these formant, pitch, and amplitude values, speech can be represented by approximately 600 bits/s. Fig. 4 shows an example of automatic analysis and synthesis. The upper set of curves are the estimated parameters that were used to control the speech synthesizer. In the middle is a spectrogram of the original speech utterance, and the lower spectrogram is of the synthetic speech corresponding to the above parameters.

In addition to efficiency, the formant representation of speech offers a great deal of flexibility in manipulating the basic speech parameters. For example, it is possible to independently alter the frequency scale, pitch, and time scale. This flexibility makes formant synthesis attractive for computer voice response applications [3], [13] and could also offer interesting possibilities for aids for the deaf.[5] Certainly the formant analysis technique itself could provide useful data on the characteristics of deaf speech.

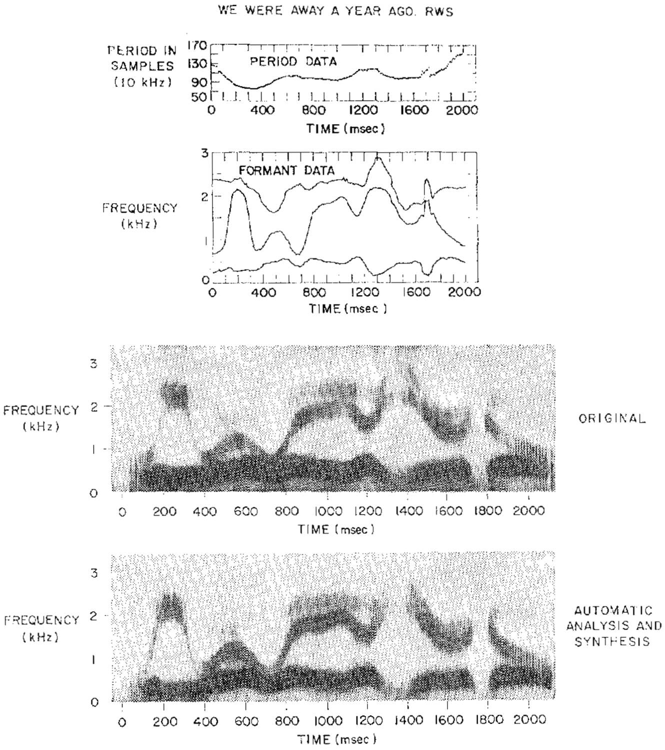

Fig. 4. Formant analysis and synthesis.

The Homomorphic Vocoder: In addition to formant analysis, the system of Fig. 3(A) serves as the basis for an analysis–synthesis system called the homomorphic vocoder [14]. In this system, the low-time samples of the cepstrum (about 30), together with pitch period and amplitude, are used to represent the speech signal. In this case, the last DFT shown in Fig. 3(A) is not part of the analysis. For synthesis, an impulse response is computed from the cepstrum values by performing the inverse of the analysis operations, i.e., DFT, exponentiation, and IDFT. Then the time-varying digital filter of Fig. 2 is realized by explicitly convolving the impulse response with an appropriate excitation.

Using this technique, speech can be represented quite well at 7200 bits/s. Although this representation may be somewhat less flexible than the formant representation, it does offer the possibility of independently manipulating the frequency scale, time scale, and pitch.

Computer-Generated Spectrum Displays: The basic analysis configuration of Fig. 3(A) has also been employed in generating spectrographic displays similar to conventional spectrograms such as those in Fig. 4 [15], [16]. The log magnitude of the short-time spectrum (C in Fig. 3) is computed as in other applications. (In the example of Fig. 3(B), the window is relatively long, corresponding to a narrowband spectrum analysis.) The spectrum can then be shaped in a manner to enhance it for display [15]. Finally, the spec-

[5] For example, the formant frequencies could be lowered to take advantage of low-frequency residual hearing capabilities.

trum values are used to intensity-modulate an oscilloscope or television monitor to produce a display in the manner of the sound spectrograph. Since the log magnitude is available, it is also possible to compute the cepstrum. This can be displayed in the same manner as the spectrum, thereby producing a "cepstrogram" for displaying pitch variations [17].

The primary advantage of computer-generated displays of this type is increased flexibility. For example, the effective analysis bandwidth can be changed by simply changing the data window. Small regions of the spectrogram can be expanded by merely displaying that part on an expanded scale. New ideas for displays such as the cepstrogram can be tried out by simply making program changes. Thus, new ways of displaying the results of very complex acoustic analyses can be explored using a computer picture display.

Predictive Coding

Predictive coding is a digital signal processing technique in which the acoustic wave is analyzed directly [18], [19]. The speech signal is analyzed by predicting the present speech sample as a linear combination of the previous p samples, i.e., the predicted sample is given by

$$s(n) = \sum_{k=1}^{p} a_k s(n-k). \tag{7}$$

The predictor coefficients a_k are obtained by minimizing the mean-squared error between the actual value and the predicted value. This leads to a set of linear equations involving the unknown predictor coefficients and a covariance matrix that can be computed from the original speech signal. An efficient algorithm for solving these equations for the a_k has been given [19].

These coefficients are the parameters of the time-varying digital filter of Fig. 2. In this case, the digital filter is a pth order recursive digital filter with transfer function

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \cdot \tag{8}$$

As in formant analysis, excitation parameters are also estimated from the speech waveform, and using this excitation information in the configuration of Fig. 2, speech can be synthesized with the digital filter specified by (8).

The predictive coding system can serve as a general speech analysis system, since spectrum information can be obtained by simply evaluating $H(z)$ on the unit circle of the z plane, i.e., $z=e^{j\omega T}$ for $-\pi/T \leq \omega \leq \pi/T$. Also, formant frequencies can be obtained by factoring the denominator polynomial in (8). Thus, the flexibility for manipulating basic speech parameters is the same as for formant synthesis. The basic information can be quantized in a number of ways to obtain an efficient representation of the speech signal. Very good quality speech can be synthesized using the predictive coding parameters quantized to 7200 bits/s. Oppenheim and Weinstein [20] have applied predictive coding methods to the impulse responses obtained in the homomorphic vocoder with significant improvements in efficiency and flexibility.

Atal has shown that from the predictor polynomial (denominator of $H(z)$) one can compute the area function of a lossless tube that has the same transmission properties as $H(z)$ [19]. This result may prove to be useful in deriving articulatory displays as aids in teaching the deaf to speak.

III. Hardware Considerations

The previous section has summarized some digital speech analysis–synthesis techniques in which very sophisticated acoustic analyses are performed, yielding very efficient and flexible representations of the speech signal. For the most part, these systems have been realized by program on general-purpose digital computers. Such realizations generally require computation time of from 20 to 100 times longer than real time and, at present, it is generally not possible to obtain real-time operation by purchasing a faster computer. However, it is possible to build special-purpose digital machines to perform a given signal processing function. An example of such a machine is a hardware realization of a digital formant synthesizer [21].

This system is a realization of Fig. 2 where the time-varying digital filter is specified by (4), (5), and (6). The system design is based on the principle that a single arithmetic unit can be multiplexed among all of the required second-order systems. Each of these sections requires two additions, two subtractions, and two multiplications for each sample of the output. Currently available integrated circuits can do about 25 times this number of operations in the time between output samples (100 μs at 10-kHz sampling rate). Thus, by providing shift register storage for the filter coefficients and the delayed filter output, and by providing control logic to insure that the appropriate inputs and coefficients are fed to the arithmetic unit at the correct time, one arithmetic unit can service the entire synthesizer.

A schematic diagram of the synthesizer is shown in Fig. 5. The synthesizer receives control signals specifying the speech parameters from a DDP-516 computer. For voiced speech, a pulse generator produces an input sequence of samples equal to the pitch period. The coefficients of (4) and (6) are obtained from the computer once per pitch period. For unvoiced speech, a random number generator (16-bit maximal length shift register sequence) provides the input, and the coefficients of (5) and (6) are obtained from the computer. The arithmetic unit consists of a three-input adder, multiplier, and a subtractor. Shift register memories hold the delayed filter variables and the coefficients received from the computer. An external clock controls the timing of the synthesizer with a sampling rate of up to 12.8 kHz being possible, and a 12-bit D-A converter provides an

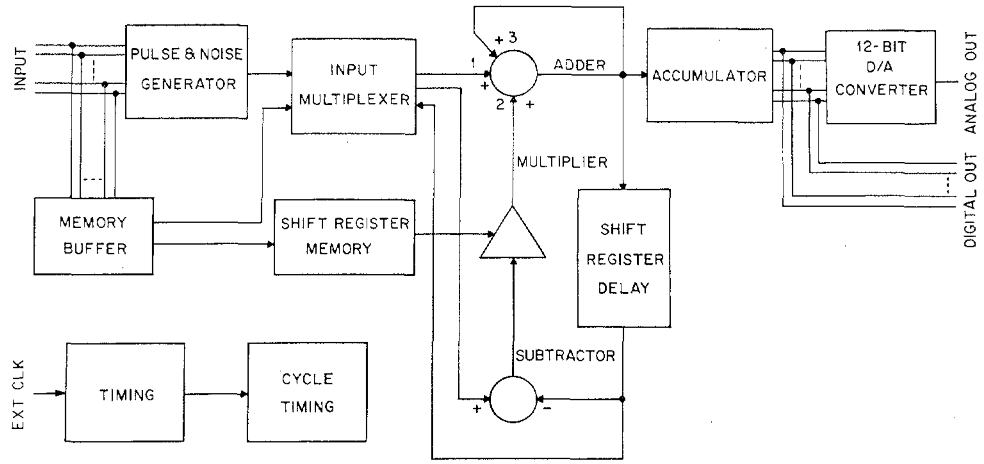

Fig. 5. Schematic diagram of a hardware digital formant synthesizer.

audible output, while 16-bit samples are returned to the computer for display or further processing.

This is one example of a complex digital signal processing system that has been realized in hardware. Digital filters of the same basic design have been built for conventional filtering applications [22] and are now commercially available. One application of such filters would be the realization of a filter bank spectrum analyzer as required, for example, in the phase vocoder.

Speech analysis systems that are based on spectrum analysis, such as homomorphic filtering, are usually realized using FFT programs. In general, real-time operation cannot be achieved using a general-purpose computer. However, a number of special-purpose machines have been built that could be used to perform this time-consuming computation. A tabulation of the characteristics of such machines is given in [23]. This survey shows that one can now obtain a machine that will compute discrete Fourier transforms of 1024 points in time on the order of 10 to 1 ms.

Thus, digital signal processing hardware is currently available for speech processing. However, equipment costs are still high when compared to general-purpose digital computers that, although slower, offer more flexibility. As integrated circuit technology develops, prices should decrease and it may soon be possible to build inexpensive hardware realizations of speech processing systems based on digital signal processing techniques.

IV. Summary

We have discussed several examples of digital signal processing systems that can provide sophisticated analyses of speech signals. Although these systems have been developed primarily for efficient and flexible storage and transmission of speech for computer voice response and telephone transmission applications, another potentially important application is in obtaining an understanding of the fundamental problems of communicating with deaf persons and in developing useful aids to handicapped persons. For the future, integrated circuit technology offers the promise of inexpensive and compact hardware realizations of these complex digital signal processing systems.

References

[1] J. L. Flanagan, "Focal points in speech communication research," presented at the 7th Int. Congr. Acoust. (Budapest), 1971; also, *IEEE Trans. Commun. Technol.*, vol. COM-19, pp. 1006–1015, Dec. 1971.

[2] P. B. Denes, "On-line computers for speech research," *IEEE Trans. Audio Electroacoust.*, vol. AU-18, pp. 418–425, Dec. 1970.

[3] 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.

[4] J. L. Flanagan, *Speech Analysis, Synthesis and Perception.* New York, Berlin: Springer-Verlag, 1965.

[5] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell Syst. Tech. J.*, vol. 45, Nov. 1966.

[6] R. W. Schafer and L. R. Rabiner, "Design of digital filter banks for speech research," *Bell Syst. Tech. J.*, pp. 3097–3115, Dec. 1971.

[7] A. Risberg, "A critical review of work on speech analyzing hearing aids," *IEEE Trans. Audio Electroacoust.*, vol. AU-17, pp. 290–297, Dec. 1969.

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

[9] 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.

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

[11] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Amer.*, vol. 47, part 2, pp. 634–648, Feb. 1970.

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

[13] L. R. Rabiner, R. W. Schafer, and J. L. Flanagan, "Computer synthesis of speech by concatenation of formant-coded words," *Bell Syst. Tech. J.*, vol. 50, pp. 1541–1558, May–June 1971.

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

[15] ——, "Speech spectrograms using the fast Fourier transform," *IEEE Spectrum*, vol. 7, pp. 57–62, Aug. 1970.

[16] P. Mermelstein, "Computer generated spectrogram displays for

on-line speech research," *IEEE Trans. Audio Electroacoust.*, vol. AU-19, pp. 44–47, Mar. 1971.

[17] M. L. Wood, "Computer generated spectrograms and cepstrograms," M.Sc. thesis, Dep. Elec. Eng., Mass. Inst. Tech., June 1971.

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

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

[20] C. J. Weinstein and A. V. Oppenheim, "Predictive coding in a homomorphic vocoder," *IEEE Trans. Audio Electroacoust.*, vol. AU-19, pp. 243–248, Sept. 1971.

[21] L. R. Rabiner, L. B. Jackson, R. W. Schafer, and C. H. Coker, "A hardware realization of a digital formant synthesizer," *IEEE Trans. Commun. Technol.*, vol. COM-19, pp. 1016–1020, Dec. 1971.

[22] L. B. Jackson, J. F. Kaiser, and H. S. McDonald, "An approach to implementation of digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 413–421, Sept. 1968.

[23] G. D. Bergland, "Fast Fourier transform hardware implementations—A survey," *IEEE Trans. Audio Electroacoust.*, vol. AU-17, pp. 109–119, June 1969

Acoustic Analysis of Deaf Speech Using Digital Processing Techniques

HARRY LEVITT, Member, IEEE
Doctoral Program in Speech
City Univ. New York
33 West 42 Street
New York, N. Y. 10036

Abstract

The development of speech-training aids for the deaf requires an understanding of how the acoustic characteristics of deaf speech differ from those of normal speech. The analysis of deaf speech presents problems in that formants may be closely spaced relative to their bandwidths, or unusual variations in voicing may make the separation of source and vocal-tract characteristics more difficult than for normal speech. The fundamental frequency contours of deaf children exhibit unusual characteristics, and a second major problem of interest is the quantitative specification of these contours and how they differ from those of normal children. Two digital-processing techniques which are well suited for these problems are the chirp *z* transform and the short-term orthogonal polynomial analysis. Application of these techniques in the acoustic analysis of the speech of deaf children is discussed.

Introduction

The congenitally deaf child is faced with a doubly severe communication handicap. Normal speech is unintelligible to him and, as a result of the lack of exposure to speech in his early development and an inability to auditorily monitor his own speech production, he has great difficulty in speaking. It is clear from the results of diligent specialized teaching that the second handicap can, in principle, be overcome. Unfortunately, few deaf individuals ever attain a speech quality that is adequate for normal conversation. Hopefully, a much larger proportion of the deaf population could be trained to speak proficiently if suitable instrumentation could be developed for assisting the deaf in learning to speak.

One approach to this problem has been the development of speech-analyzing aids for speech training [1], [2]. Speech-analyzing aids typically operate on the acoustic signal extracting that information which is not normally available to the deaf user, recoding the essential parameters, and transmitting this information back to the user either visually, tactually, or using whatever residual hearing is available. Aids of this type have, thus far, shown a modest degree of success as speech-training aids [3], [4].

In order to develop better speech-analyzing aids for speech training it is necessary to find out more about the acoustic and perceptual characteristics of the speech of deaf children and how these characteristics differ from those of normal speech. Hudgins and Numbers [5], in a perceptual evaluation of articulatory errors that affect intelligibility, found that the most common errors were omission of consonants, voice-voiceless confusions, misarticulation of consonant clusters, vowel substitutions, the reduction of vowels to an indistinct neutral form, and errors of rhythm and timing. Consonants produced at the front of the mouth are substantially less prone to error than those produced at the center or back of the mouth. Also, consonant confusions tend to involve errors of voicing or manner, but seldom of place [6]. For example, the plosive /b/ is often substituted for its unvoiced counterpart /p/ (voicing error) or for the nasal /m/ (manner error), but very rarely for the voiced plosive /d/ (place error).

Most of the articulatory errors in the speech of the deaf involve incorrect movements of the articulators. The linking together of successive phonemes is particularly prone to error. In an acoustic study on voiced and voiceless consonants in the intervocalic position (i.e., between two voiced sounds), Calvert [7] found that deaf speakers generally tended to distort the durational characteristics of phonemes.

Manuscript received November 5, 1971.

This research was supported in part by NIH Grant NS 09252.

This paper was presented at the IEEE Fall Electronics Conference, Chicago, Ill., October 1971.