A Hardware Realization of a Digital Formant Speech Synthesizer

LAWRENCE R. RABINER, MEMBER, IEEE, LELAND B. JACKSON, MEMBER, IEEE, RONALD W. SCHAFER, MEMBER, IEEE, AND CECIL H. COKER, SENIOR MEMBER, IEEE

Abstract—**Terminal analog or formant speech synthesizers have found many applications in speech research. These include investigation of computer voice response, speech synthesis-by-rule, and speech perception studies, among others. Many types of formant synthesizers have been designed and realized either in analog circuitry or as a computer program. In this paper we describe a digital hardware realization of a formant synthesizer which utilizes the technique of digital multiplexing of a single arithmetic unit among several digital filter sections. The advantages of this hardware over conventional analog hardware include: precise control over center frequencies and bandwidths of the resonators in the synthesizer, stability and reliability of the hardware, light weight, small size, and low power consumption. The synthesizer is capable of producing speech in real time at sampling rates up to 12.8 kHz, using 24 bits to process the digital signals internal to the synthesizer. A 12-bit digital-to-analog convertor supplies an immediate analog output for monitoring the speech and a provision is included for returning 16 bits of the output signal to the computer for future processing such as waveform display or spectrum analysis.**

Introduction

THE IMPORTANCE of terminal analog or formant speech synthesizers to speech communication research has been apparent for many years. Among the applications of formant synthesizers are computer voice response [1], [2], speech synthesis-by-rule [3]–[5], and speech perception experiments [6], [7].

There have been many attempts at building hardware to synthesize speech from formant control parameters. The simplest piece of hardware was essentially a vowel generator [8] capable of producing steady-state sounds only. In order to produce continuous speech, provision must be made to allow time variations of the formant control parameters [9]. The solution to this problem was to let a digital computer control the synthesizer [10], i.e., to provide the time-varying control parameters necessary to produce continuous speech. This innovation immediately led to the realization of many computer controlled synthesizers realized from analog circuitry [11]–[13]. Although they provided reasonable solutions to the need for a formant synthesizer, these synthesizers suffered from the typical problems of any complicated piece of analog equipment, i.e., their characteristics were highly temperature and humidity dependent, and they required constant maintenance to remain in good working order. The temperature sensitivity of the hardware was manifested both as drift in the frequency characteristics of the resonators, as well as a tendency for the hardware to become unstable. The analog circuitry also tended to become unbalanced leading to spurious undesirable noises.

To alleviate the problems associated with analog circuitry, computer programs were written to simulate a formant synthesizer [14], [15]. The advantages of an all digital speech synthesizer included precise control of all resonator center frequencies and bandwidths, repeatability of the data, and ease of modification of the synthesizer configuration. Also, complicated higher pole correction networks are not required with a digital synthesizer [16]. Another advantage was that highly sophisticated synthesizers could be designed and simulated [17]. The major disadvantage of the programmed digital synthesizer was that a fast computer (1-μs cycle time) could not perform all the necessary arithmetic operations to synthesize the samples of the speech waveform in "real time" at a 10-kHz sampling rate, for example. Depending on the specific computer used, and the details of the synthesizer, speech synthesis simulations generally required from 2 to 20 times real time.

With the advent of integrated circuit technology, it is feasible to consider building digital hardware for speech synthesis which can synthesize speech in real time. In this paper we describe such a synthesizer which we have designed and built. Although we designed the hardware synthesizer for use with the Bell Labs DDP-516 computer facility for Acoustics Research, the synthesizer is generally useful with a variety of digital computers.

Synthesizer Description

Fig. 1 shows a simplified block diagram of a digital formant synthesizer that has been employed in computer voice response studies [1], [2], [9]. There are two excitation sources: an externally controllable impulse generator whose output consists of a unit pulse once every pitch period (P samples); and a pseudorandom uniform number generator whose output approximates a white noise generator. The impulse generator is generally used to produce voiced sounds, i.e., where the vocal cords are vibrating; and the noise generator is used to produce both unvoiced sounds and whispered speech.

There are two basic signal processing paths in the synthesizer. The upper path consists of an intensity modu-

Manuscript received June 10, 1971; revised July 6, 1971. This paper was presented at the 7th International Congress on Acoustics, Budapest, Hungary, August 1971.

L. R. Rabiner, R. W. Schafer, and C. H. Coker are with the Bell Telephone Laboratories, Murray Hill, N. J. 07974.

L. B. Jackson is with the Rockland Systems Corporation, Blauvelt, N. Y.

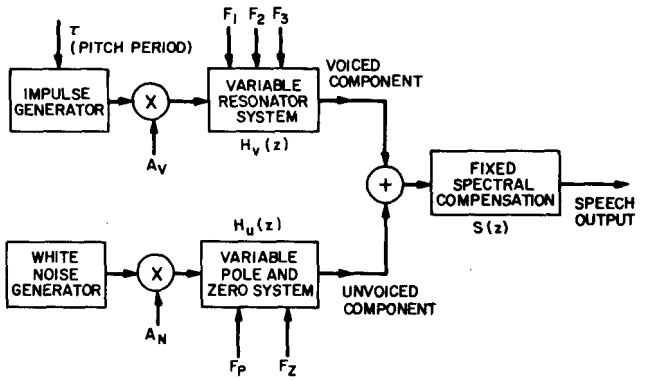


Fig. 1. Synthesizer used in computer voice response studies.

lator (A_v) and a time-varying digital filter consisting of a cascade of L variable resonators (poles). The transfer function of this filter (under steady-state conditions) is

$$H_v(z) = \prod_{k=1}^{L} \left(\frac{1 - \exp(-a_kT)2\cos(b_kT) + \exp(-2a_kT)}{1 - \exp(-a_kT)2\cos(b_kT)z^{-1} + \exp(-2a_kT)z^{-2}} \right) \quad (1)$$

where a_k is the radian bandwidth of the kth pole, b_k is the radian center frequency of the kth pole, and T is the sampling period. Although all the pole center frequencies and bandwidths can be controlled, generally only the lowest three center frequencies are varied as shown by the control signal inputs ($F1$, $F2$, $F3$) to the variable resonator system in Fig. 1. The variable resonator system accounts for the effects of the time-varying shape of the vocal tract on the speech spectrum.

The effects of radiation of sound from the mouth (or nose) into air, and glottal excitation pulse shape must be accounted for. This is the function of the fixed spectral compensation network whose transfer function is of the form

$$S(z) = \frac{(1 - \exp(-aT))(1 + \exp(-bT))}{(1 - \exp(-aT)z^{-1})(1 - \exp(-bT)z^{-1})}. \quad (2)$$

This network consists of two real axis poles (one in the right-half z plane, one the left-half z plane) which approximates the desired transfer function.

The lower path in Fig. 1 consists of a modulator (A_N) which controls the variance of the noise generator output, and another time-varying digital filter consisting of a cascade of a pole and zero. Its transfer function is of the form:

$$H_u(z) = \frac{(1 - 2\exp(-aT)\cos(bT) + \exp(-2aT))(1 - 2\exp(-cT)\cos(dT)z^{-1} + \exp(-2cT)z^{-2})}{(1 - 2\exp(-aT)\cos(bT)z^{-1} + \exp(-2aT)z^{-2})(1 - 2\exp(-cT)\cos(dT) + \exp(-2cT))} \quad (3)$$

where a, b, c, and d are the radian bandwidths and center frequencies of the time-varying pole and zero. Generally, the bandwidths of the pole and zero are fixed, and only the center frequencies vary as shown by the control signal inputs F_p and F_z to the variable pole and zero system in Fig. 1. The output of this system is passed to the fixed spectral compensation system to provide the final unvoiced speech output.

It should be noted that each of the transfer functions (1)–(3) of the synthesizer has the property that at zero frequency the transfer function is unity independent of the center frequencies and bandwidths of any pole or zero. This property is essential to account for the unity transmission of the vocal tract at zero frequency, and is achieved by using resonators which are individually normalized to have this property.

As stated earlier, the synthesizer of Fig. 1 was designed for use in computer voice response studies where the control parameters were automatically estimated from natural speech. Because of the difficulties in estimating control parameters other than pitch period, intensity, three variable formants, and fricative pole, and zero, the digital synthesizer of Fig. 1 is incomplete in several details which are desirable in a general purpose synthesizer. For example, there is no provision for a network to produce the nasal consonants n and m, or a network to produce the voiced fricatives z (as in *z*oo), zh (as in a*z*ure), v (as in *v*ery), and th (as in *th*ere). To synthesize nasal consonants a network consisting of a time-varying pole and zero must be placed in cascade with the variable resonator system of Fig. 1. To adequately synthesize voiced fricatives, a network which modulates the noise generator output by the voiced path output is necessary. Also, for additional flexibility in the synthesizer provision should be made to allow the noise generator output to excite the voiced processing path in order to produce whispered speech.

Fig. 2 shows a block diagram of the synthesizer we have built in hardware. This synthesizer derives its time-varying control parameters pitch synchronously (i.e., it changes all parameters at the beginning of each pitch period) from the DDP-516 computer.[1] Each of the control parameters is specified by one 16-bit word from the computer. This synthesizer is similar in concept to the one in Fig. 1, but differs slightly in its details. Specifically the upper signal processing path consists of six two-pole digital filters [$L = 6$ in (1)] and one two-zero filter, where the bandwidth and center frequency of each filter is controllable. The sixth two-pole filter and the two-zero filter account for a nasal pole and zero, and cancel each other during nonnasal sounds. (Exact cancellation of a pole by a zero is easily accomplished in a digital system.) Four of the two-pole filters (or possibly five during nonnasal sounds) are used to represent the time-varying vocal tract transfer function $H_v(z)$ and the last two-pole filter provides the desired spectral compensation $S(z)$.

[1] When the speech is unvoiced, the user sets the pitch period to whatever value he desires. Thus the coefficients of the digital filters are still varied once per "pitch period."

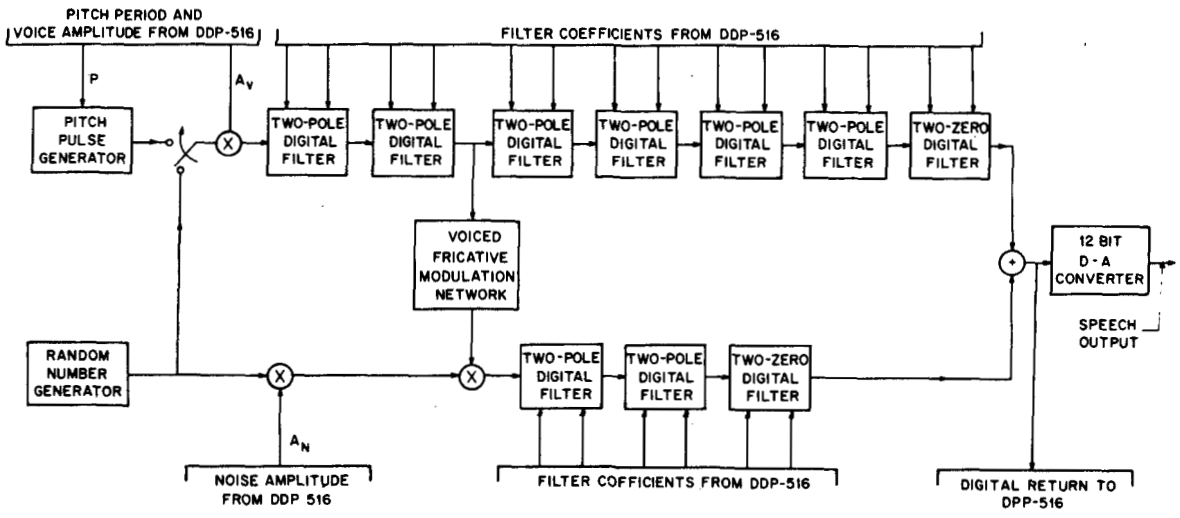


Fig. 2. Synthesizer built in digital hardware.

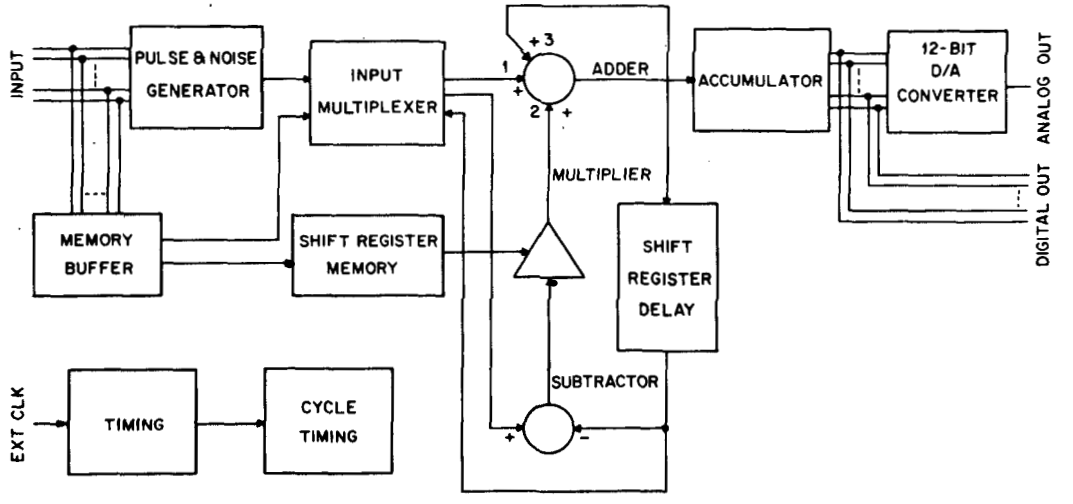


Fig. 3. Schematic logic diagram of synthesizer.

The unvoiced signal processing path consists of two two-pole filters and one two-zero filter. Again the bandwidths and center frequencies of each of the filters can be varied externally. One two-pole and one two-zero filter are used to represent $H_u(z)$, and the remaining two-pole filter is used to provide the necessary spectral compensation $S(z)$. In this synthesizer, for added flexibility, the voiced and unvoiced spectral compensation networks may be different since they are included separately in each path of the synthesizer.

To provide the necessary modulation to produce voiced fricatives (z, zh, v, th) a voiced fricative modulation network takes the voiced output from the second two-pole filter, and modulates the noise output. The details of this highly nonlinear network are described elsewhere [17], but what it basically does is to zero out the noise except for a small part of the pitch period (typically 10–30 percent). Thus rather than having a continuous stream of noise samples excite the unvoiced path, only a small burst of noise is used. Thus the unvoiced output is essentially pitch modulated noise.

The specific pseudorandom number generator used is a 16-bit maximal length shift register sequence [18]. This algorithm generates a random bit from mod-2 sums of the previous 16 bits, shifts out the bit generated 16 clock pulses earlier, and shifts in the new bit. The algorithm used to generate the current bit is

$$X_n = X_{n-1} \oplus X_{n-12} \oplus X_{n-14} \oplus X_{n-15}, \qquad n = 0, 1, 2, \cdots \tag{4}$$

where each X is either 1 or 0, and 1 physically corresponds to a positive excitation pulse, and a 0 to a negative excitation pulse. Thus the noise generator output consists of a random succession of positive and negative pulses. The spectrum of the noise generator output is flat.

The outputs of the two signal processing paths are added, and the 16 most significant bits are returned to

Fig. 4. Front view of hardware synthesizer.

the computer for storage on the disc. This digital return makes possible waveform examination, or spectrum analysis of the synthetic speech. Different versions of the same utterance may be readily compared by listening to the digital waveforms stored consecutively on the disc. Simultaneous with the digital return to the computer, a 12-bit digital-to-analog converter provides immediate acoustic output for monitoring purposes, or for direct recording on analog tape.

All the filter coefficients are 16-bit two's complement integers. Internally in the synthesizer the digital filtering is performed with 24-bit accuracy insuring that the effects of quantization and roundoff remain negligible [16].

Principles of Digital Operation

The basic principle behind the digital hardware is the multiplexing of a single arithmetic unit among all the two-pole filters, and the two-zero filters [19]. The arithmetic operations required to realize a two-pole filter, for example, are two additions, two subtractions, and two multiplications for each output sample. High-speed integrated circuits are capable of doing about 25 times this number of arithmetic operations in the time between output samples (100 μs at a 10-kHz sampling rate). Thus the notion of sharing a single arithmetic unit among many filters attains practical significance in the synthesizer. By providing storage for the filter coefficients, and the delayed outputs of the filters, and by dynamically controlling which inputs go into the arithmetic unit, and where the outputs go, a single arithmetic unit can service the entire synthesizer.

A schematic block diagram of the digital logic used to realize the synthesizer is shown in Fig. 3. The arithmetic unit consists of a three-input adder, a shift register delay (which holds the delayed filter variables), a subtractor, and a multiplier. The length of the shift register delay is 480 bits (20 delayed variables times 24 bits per variable). Another shift register memory of 320 bits (20 filter coefficients times 16 bits per coefficient) holds the multipliers for each of the filter sections. This arithmetic unit can perform a multiplication, an addition, and a subtraction simultaneously in about 3.9 μs, therefore each filter section requires about 7.8 μs per iteration. In this

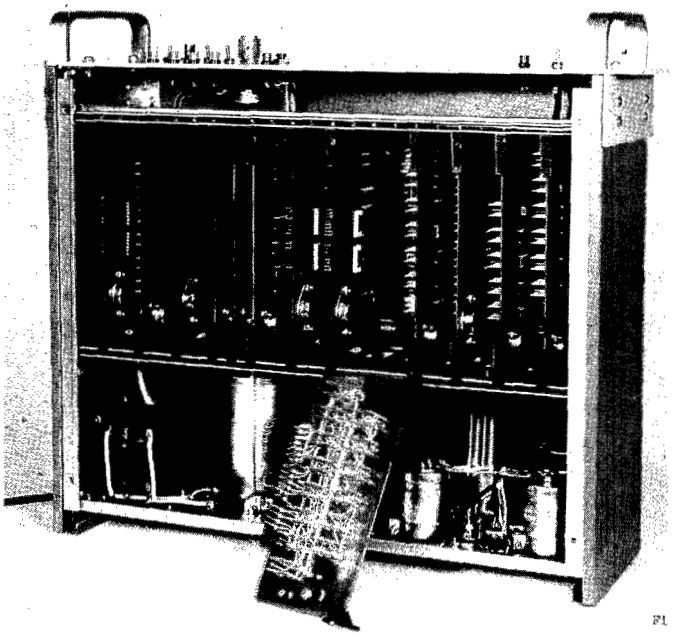


Fig. 5. Top view of hardware synthesizer.

manner the 10 filter sections of the synthesizer require about 78 μs. Thus the synthesizer can operate at sampling frequencies up to 12.8 kHz.

The remainder of the logic diagram is straightforward. The synthesizer control signals come from the computer output line to the input of the synthesizer. A memory buffer transfers the gain coefficients and pitch period to the pulse and noise generator, and the input multiplexer. The memory buffer shifts the filter coefficients to the shift register memory. The pulse and noise generator provides excitation to the arithmetic unit via the input multiplexer. An accumulator sums the voiced and unvoiced outputs and sends the 16 most significant bits back to the computer, simultaneously converting the 12 most significant bits to analog form. The switching and timing logic is determined from timing and cycle timing logic which use an externally supplied clock to determine the basic synthesizer sampling rate. The sampling rate is thus easily changed without any internal modifications to the synthesizer.

Physical Description of Synthesizer

Fig. 4 shows a photograph of the synthesizer. Among the benefits of integrated circuit technology are the small size (5 in by 17 in by 14 in) and light weight (22 lb) of the synthesizer.

For convenience, an overflow light is provided to signal digital overflow in one of the filters, and through a series of bit selector switches, the user may listen to any 12 consecutive bits of the output. This feature may be valuable for noise studies when one is interested in the noisy bits of the synthesizer, rather than the signal bits. Fig. 5 shows a top view of the hardware. Medium speed tran-

sistor–transistor logic (TTL) circuits are used throughout the synthesizer. Thirteen wire wrapped cards comprise the logic—six of which are the arithmetic unit.

Conclusion

Digital hardware is now available for synthesizing speech in real time from formant control data. The hardware is capable of producing high quality speech at sampling rates up to 12.8 kHz, with 24-bit processing accuracy. The synthesizer has been found to be extremely reliable in almost a year of operation.

References

[1] 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.

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

[3] J. Holmes, I. Mattingly, and J. Shearme, "Speech synthesis by rule," *Lang. Speech*, pp. 127–143, 1967.

[4] L. R. Rabiner, "Speech synthesis by rule: An acoustic domain approach," *Bell Syst. Tech. J.*, pp. 17–37, Mar. 1970.

[5] C. H. Coker and N. Umeda, "Text to speech conversion," in *1970 IEEE Int. Conv. Dig.*, pp. 216–217.

[6] K. N. Stevens and A. S. House, "Speech perception," in *Foundations of Modern Auditory Theory*, J. Tobias and E. Schubert, Eds. New York: Academic Press, 1970.

[7] A. E. Rosenberg, R. W. Schafer, and L. R. Rabiner, "Effects of smoothing and quantization of the parameters of formant-coded speech," *J. Acoust. Soc. Amer.*, to be published.

[8] G. Fant, "Speech communication research," *IVA* (Sweden), pp. 331–337, 1953.

[9] J. L. Flanagan, "Focal points in speech communication," this issue, pp. 1006–1015.

[10] J. B. Dennis, "Speech synthesis," Res. Lab. Electron., M.I.T., Cambridge, Mass., Rep. QPR 67, pp. 157–162, Oct. 15, 1962.

[11] G. Fant, J. Martony, V. Rengman, and A. Risberg, "OVE III synthesis strategy," presented at the Speech Communications Seminar, Stockholm, Sweden, 1962, Paper F5.

[12] R. S. Tomlinson, "SPASS—an improved terminal-analog speech synthesizer," *J. Acoust. Soc. Amer.* (Abstract), vol. 38, p. 940(A), 1965.

[13] C. H. Coker and P. Cummiskey, "On-line computer control of the formant synthesizer," *J. Acoust. Soc. Amer.*, vol. 38, 940(A), 1965.

[14] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*. New York: Academic Press, 1965, pp. 203–209.

[15] J. L. Flanagan, C. H. Coker, and C. M. Bird, "Digital computer simulation of formant-vocoder speech synthesizer," presented at the Audio Eng. Soc. Meeting, 1963.

[16] B. Gold and L. R. Rabiner, "Analysis of digital and analog formant synthesizers," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 81–94, Mar. 1968.

[17] L. R. Rabiner, "Digital formant synthesizer for speech-synthesis studies," *J. Acoust. Soc. Amer.*, vol. 43, pp. 822–828, 1968.

[18] S. Golomb, *Shift Register Sequences*. San Francisco: Holden-Day, 1967.

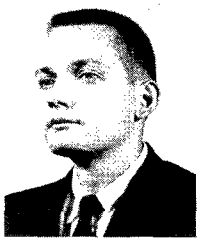
[19] L. B. Jackson, J. F. Kaiser, and H. S. McDonald, "An approach to the implementation of digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 413–421, Sept. 1968.

Lawrence R. Rabiner (S'62–M'67), for a photograph and biography please see page 195 of the April issue of this TRANSACTIONS.

Leland B. Jackson (S'62–M'65) was born in Atlanta, Ga., on July 23, 1940. He received the S.B. and S.M. degrees from the Massachusetts Institute of Technology, Cambridge, in 1963, and the Sc.D. degree from the Stevens Institute of Technology, Hoboken, N. J., in 1970, all in electrical engineering.

During 1961–1962 he was associated with the Bell Telephone Laboratories, Inc., under the M.I.T. cooperative program in electrical engineering. From 1964 to 1966 he was employed by the Sylvania Electronic Systems, Inc., Mountain View, Calif., where he studied and developed signal processing techniques for ionospheric research. From 1966 to 1970 he was a Member of the Technical Staff of the Bell Telephone Laboratories, Inc., Murray Hill, N. J., where he was primarily concerned with the analysis and synthesis of digital filters. In 1970 he joined the Rockland Laboratories, Inc., Blauvelt, N. Y., as Vice President for Engineering.

Dr. Jackson is a member of Tau Beta Pi, Eta Kappa Nu, and Sigma Xi.



Ronald W. Schafer (S'62–M'67) received the B.S.E.E. and M.S.E.E. degrees from the University of Nebraska, Lincoln, in 1961 and 1962, respectively, and the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1968.

During 1962–1963 he was an Instructor in the Department of Electrical Engineering, University of Nebraska, and from 1964 until 1968 he was an Instructor in the Department of Electrical Engineering, M.I.T. During 1965 he was associated with the M.I.T. Electronic Systems Laboratory and from 1966 to 1968 he was a Member of the Research Laboratory of Electronics, M.I.T. In 1968 he joined the Acoustics Research Department, Bell Telephone Laboratories, Murray Hill, N. J., where he is engaged in research on speech analysis and synthesis and digital signal processing techniques.

Dr. Schafer is a member of Eta Kappa Nu, Sigma Xi, and the Acoustical Society of America. He is a member of the IEEE G-AE Committees on Speech Communication and Digital Signal Processing. He received a departmental teaching award at M.I.T. and with A. V. Oppenheim and T. G. Stockham, he was awarded the 1969 G-AE Senior Award.

Cecil H. Coker (S'55–SM'68) received the B.S. and M.S. degrees in electrical engineering from the Mississippi State University, State College, in 1954 and 1956, respectively, and the Ph.D. degree in electrical engineering from the University of Wisconsin, Madison, in 1960.

During 1960 he was an Instructor in the Department of Electrical Engineering, University of Wisconsin. He joined the Bell Telephone Laboratories, Inc., Murray Hill, N. J., in 1961, where he worked for several years on formant analysis and synthesis of speech. Subsequent work included supervision of the development of two laboratory computer facilities. His work on modeling of the articulatory process was begun in 1966. He is curently a Supervisor in the Acoustics Research Department at Bell Laboratories. He holds several patents and has written a number of papers on speech analysis and synthesis.