The Use of Waveform Asymmetry to Identify Voiced Sounds

DAVID J. COMER, Member, IEEE
Department of Electrical Engineering
University of Calgary
Calgary, Alberta, Canada

Abstract

A simple speech measurement is described that allows the identification of voiced and unvoiced sounds. In addition the measurement can be used to distinguish between a limited number of different vowel sounds. The measurement consists of examining the difference in magnitude between the positive and negative peaks of the speech waveform. This asymmetry measurement can become completely volume insensitive if preceded by a frequency-dependent phase shifter. Optimization of the measurement for individual speakers is quite easily accomplished.

Manuscript received December 11, 1967; revised April 26, 1968.

The author was with the Advanced Systems Division, IBM Corporation, from 1959 to 1964.

Introduction

In 1959 the Automatic Speech Recognition group at IBM Corporation began to examine the possibility of identifying voiced sounds and classifying vowel sounds by using waveform asymmetry measurements. It became immediately apparent that asymmetry provides an effective means of differentiating between vowel sounds of a limited vocabulary. Subsequent work led to the development of a 15-word vocabulary, voice-controlled adding machine called Shoebox [1]. This device was demonstrated at the Seattle World Fair in 1962.

One of the main features of the asymmetry measurement is the minimal time required to optimize performance for different speakers. It has long been recognized that the characteristics of a given phoneme vary from speaker to speaker [2]. In developing a speech recognition device to be used by a large group of people, either the measurements must be tolerant of the differences in speaker characteristics or a simple adjustment of measurements to account for these differences must be possible. By means of adjustable asymmetry measurements, Shoebox performed effectively for twenty different operators, both male and female, at the Seattle World Fair.

This paper examines the waveform asymmetry measurement in some detail, noting the advantages and disadvantages in using this measurement as an independent parameter to identify vowel sounds. The results of more recent controlled tests conducted by the Industrial Psychology Department of IBM are reported. These tests indicate that the asymmetry measurement is effective for both male and female speakers.

Waveform Asymmetry

Waveform asymmetry is a measure of the difference in magnitude between positive and negative peaks of a waveform. Consider the signals of Fig. 1. The upper waveform is a sinusoid whose peaks are symmetrical about the base line. The lower waveform is asymmetrical with respect to peak values of the signal. Note that the net area under the waveform over a period equals zero even though the magnitudes of the positive and negative peaks differ.

Fig. 2 shows the block diagram of one possible asymmetry detector. A rectifier passes the positive peaks of the waveform to an envelope demodulator while a second rectifier passes the negative peaks to a second demodulator. If the time constants of the demodulator circuits are correctly chosen, the demodulator outputs will be approximately equal to the peak values of their respective inputs. The output of the demodulator associated with the negative peaks will be negative; thus when summed with the positive output of the other demodulator, the result is a signal proportionl to the peak asymmetry. Fig. 3 shows the waveforms that would appear at various points in the detector for the input shown. The output of the summer can be followed by an RC network to obtain the average value of asymmetry.

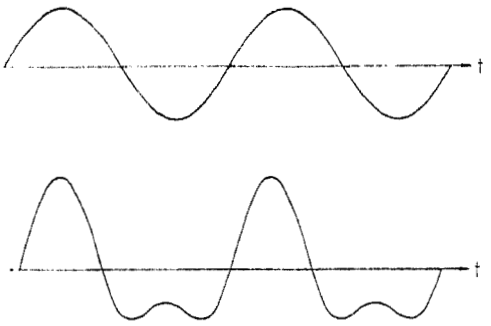


Fig. 1. A symmetrical and an asymmetrical waveform.

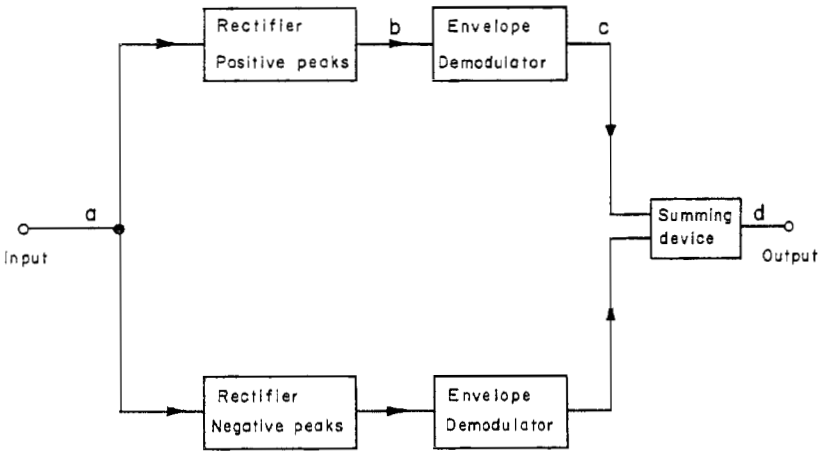
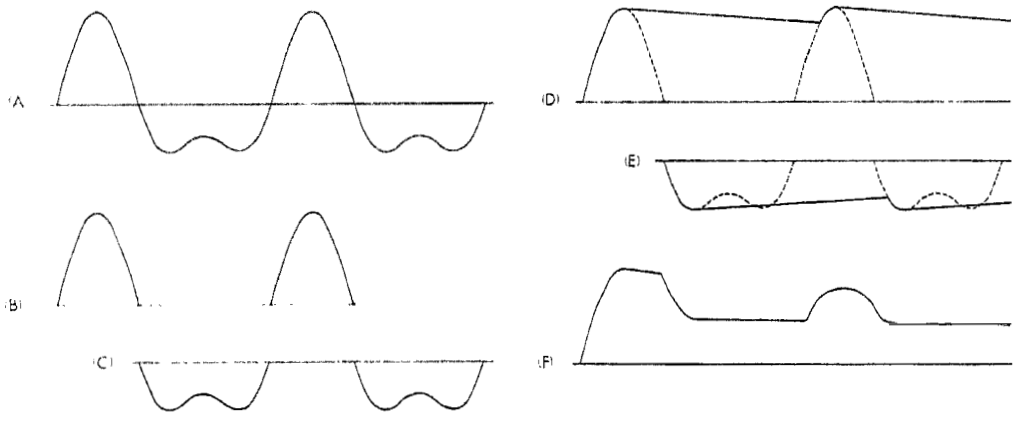


Fig. 2. Block diagram of an asymmetry detector.

Fig. 3. Asymmetry detector waveforms. (A) Input. (B) Rectified positive peaks. (C) Rectified negative peaks. (D) Demodulated positive peaks. (E) Demodulated negative peaks. (F) Summed demodulator outputs.



Asymmetry of Voiced Sounds

Fig. 4 shows the approximate waveforms for three different phonemes. All unvoiced sounds, composed of nonharmonically related components, will be symmetrical about the base line when averaged over relatively long periods of time. Examples of phonemes with symmetrical waveforms are |f|, |th|, and |s|.

The |a| sound shown possesses positive asymmetry since the positive peaks are larger than each succeeding negative peak. The |e| sound has negative asymmetry. All voiced sounds exhibit asymmetry as spoken or can be modified to exhibit asymmetry as explained in the next section. This measurement can quite accurately identify voiced sounds as opposed to nonvoiced sounds and, therefore, can perform the function of segmenting a word into voiced or unvoiced portions. Segmentation of a word is very important in a limited vocabulary device since the sequence of voiced and unvoiced portions of the word aids identification [2]. As explained later, sequence infor-

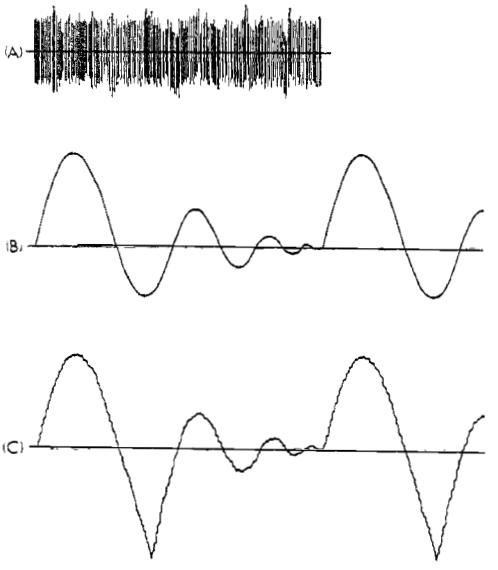


Fig. 4. Waveforms of three phonemes. (A) The |s| sound. (B) The |a| sound. (C) The |e| sound.

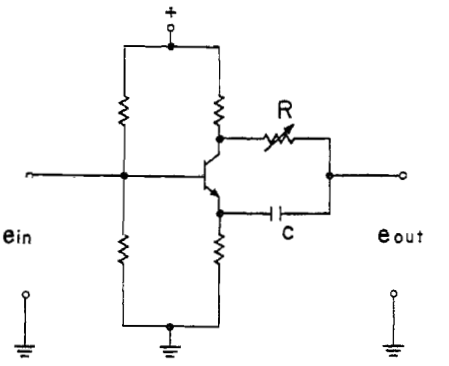


Fig. 5. Phase shift network.

mation is used in Shoebox. This measurement has proven to be very consistent and requires no adjustment from one speaker to another.

Vowel Separation by Asymmetry

In addition to identifying voiced sounds, the asymmetry measurement can be used for vowel separation on a limited basis. For example, it would be a simple matter to separate the |a| and |e| sounds of Fig. 4 using this measurement. If the |a| sound is applied to the detector a positive output results, while the |e| sound causes a negative output. Other pairs of phonemes can be compared and separated on this basis. Unfortunately, while vowel separation is consistently accurate for a single speaker, the accuracy decreases considerably as different speakers generate these same vowel sounds. To accommodate differences in speakers an adjustable phase shift network can precede the asymmetry detector. The all-pass phase shifter of Fig. 5 allows individual optimization of the asymmetry measurement for each speaker. The inclusion of this phase shifter came about after it was observed that the asymmetry of a vowel sound is dependent upon the phase relationships of its harmonic components. To allow an additional degree of freedom, adjustable weighting resistors are used at the output of both the positive and negative envelope demodulators. After adjusting the phase shifter one vowel sound might possess a great deal of positive asymmetry while a second vowel might be only weakly negative (or even slightly positive). Weighting the negative envelope can cause the second sound to have more negative asymmetry at the expense of less positive asymmetry for the first sound.

It should be noted that when the phase shifter and weighting resistors are adjusted such that one vowel sound has positive asymmetry while another has negative asymmetry, the separation of these vowels is independent of volume. That is, the sounds can be spoken very loudly or very softly and the separation can still be accomplished since it is a polarity sensitive measurement. This insensitiveness to volume variations extends the dynamic range of volume over which a speech recognition device can perform effectively.

While there is no doubt that asymmetry can be a useful measurement for speech recognition devices, one might ask just what quantity or quantities are actually being measured. Furthermore, there may be other measurements that supply essentially the same information that can be used in place of the asymmetry measurement. The next section discusses these points after developing the appropriate theory of asymmetry detection.

The Origin of Asymmetry

It is well known that vowel sounds give rise to quasi-periodic waveforms with a duration extending over several repetitive periods [3]. The fundamental frequency is determined by the excitation of the vocal cords while the resonant cavities of the throat and mouth determine the relative magnitude of the harmonic components. Since the vowel sounds can be resolved into harmonically related components, some information regarding asymmetry can be derived from examining waveforms made up of these components.

If a repetitive waveform is made up of a fundamental plus higher order odd harmonics, half-wave symmetry will result. This symmetry is independent of the phase relationship of the various harmonics to the fundamental. It is obvious that a waveform containing only odd harmonics can possess no peak asymmetry.

Consider a waveform composed of a fundamental and a second harmonic component as shown in Fig. 6. If there is no phase difference between the two components, the sum of the components will be a symmetrical waveform as shown. However, if the second harmonic is

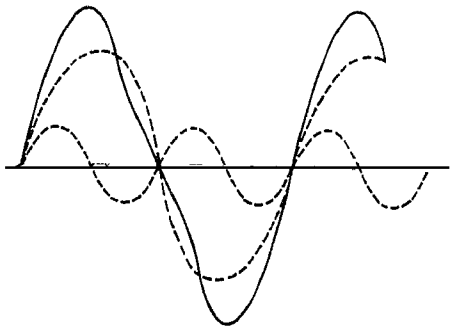


Fig. 6. Symmetrical waveform composed of the fundamental and second harmonic with no phase difference.

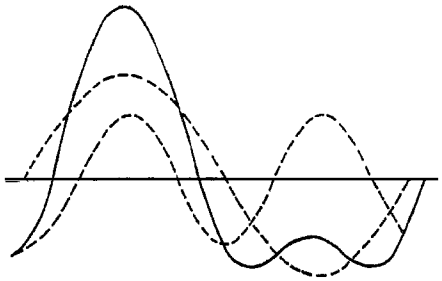


Fig. 7. Asymmetrical waveform composed of the fundamental and second harmonic with a phase difference of 45°.

shifted in phase by −90° (−45° with respect to the fundamental frequency) the resulting waveform is quite asymmetrical as shown in Fig. 7. The asymmetry is then a function of two quantities: the relative magnitude of the second harmonic to the fundamental and the phase relationship of the two components. If the waveform of Fig. 7 were applied to the asymmetry detector, the resulting output would indicate the presence of an even harmonic, but would not be an absolute measure of either the relative magnitude or the phase difference. A change in asymmetry can be effected by modifying either one of these two quantities.

If many harmonics are present in the input waveform, the asymmetry is a very complex function of the magnitudes and phases of all even harmonics relative to all odd harmonics. For most vowel sounds the amplitudes of the higher order harmonics are much smaller than those of the lower frequency components. This means that the waveform asymmetry is determined mainly by the fundamental and the first few harmonics. The first four components are often the major contributors to asymmetry and thus higher order harmonics can be neglected.

Because relative phase contributes to the asymmetry of the waveform, it follows that a change in relative phase will modify the asymmetry. To effect the change in phase, the circuit of Fig. 5 is used. It can be shown [4] that this circuit is an all-pass network described by the transfer function

$$\frac{e_{out}}{e_{in}} = \frac{\tau s - 1}{\tau s + 1} \tag{1}$$

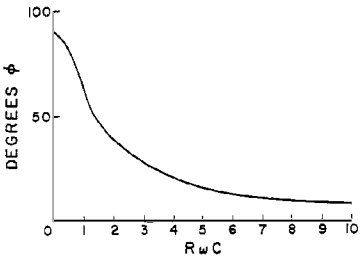
where $s=j\omega$. The phase shift between input and output is

$$\phi_\omega = 2 \arctan \frac{1}{\omega CR} \cdot \tag{2}$$

Equation (2) shows that the phase shift is a strong function of frequency.

It is easy to see how the asymmetry of a vowel waveform is modified when passed through the phase shift circuit. If we assume that the waveform is made up of the fundamental plus a second harmonic that is in phase,

Fig. 8. Relative phase between the fundamental and second harmonic (angle referenced to fundamental frequency).



there will be no waveform asymmetry exhibited. At the output of the phase shifter the phase difference will be

$$\phi = \phi_1 - \frac{\phi_2}{2} \tag{3}$$

where ϕ_1 and ϕ_2 are the phase shifts of the fundamental and first harmonic, respectively, when passed through the phase shifter. The factor of 2 is used to reference all phase shifts to the fundamental frequency. From (2), ϕ can be written as

$$\phi = 2 \arctan \frac{1}{\omega CR} - \arctan \frac{1}{2\omega CR} \cdot$$

If R is changed from a very low value to some high value ϕ varies from 90° to 0° as shown in Fig. 8. At a value of

$$R = 1.6/\omega C$$

the relative phase between the components is 45°. As explained previously this condition results in maximum asymmetry. Thus, any degree of waveform asymmetry can be achieved by varying R.

In the case of a vowel sound there are other important components and they may have relative phase differences before being applied to the phase shifter. Two phase shifters in cascade can be used in order to allow larger phase differences and insure that both positive and negative asymmetry can be obtained.

That asymmetry for a given vowel can be controlled is

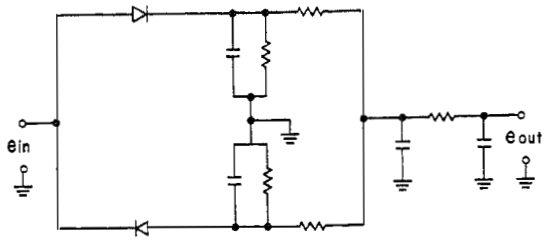


Fig. 9. Schematic of an asymmetry detector.

obvious from the above discussion. This does not guarantee that two given vowel sounds will exhibit a difference in asymmetry that is large enough to be reliably detected. On the other hand, it is reasonable to expect that the relative phases of the harmonic components will be different for different vowel sounds. The cavity of the throat and mouth is shaped differently for each vowel sound causing a different resonant frequency for each vowel sound. If the exciting frequency (determined by the vocal cord excitation) remains relatively constant as the resonant cavity changes, the difference between the resonant frequency and excitation frequency changes. Since the phase shift of a resonant circuit or cavity is strongly dependent on the difference in excitation and resonant frequencies, the components of each vowel sound will be subjected to different amounts of phase shift. Furthermore, the relative magnitudes of the components vary as the cavity is changed. Thus, most vowel sounds should exhibit a different amount of asymmetry. Experimental work verifies this conclusion. Unfortunately, the asymmetry measurement is strongly dependent on volume. If the amplitude of a vowel sound is doubled, the output of the asymmetry detector doubles.

Because of this volume dependence, one of the most useful functions of the asymmetry measurement is to differentiate between two vowel sounds or between two classes of vowel sounds. When used to separate two sounds, the phase shifter and weighting resistors are adjusted to give positive asymmetry for one sound and negative asymmetry for the other sound. Polarity rather than magnitude of the detector output then classifies the sounds. The measurement is completely volume insensitive when used in this manner.

If several pairs of vowels are to be classified, an asymmetry detector must be used for each pair. While this appears at first to be a fairly expensive arrangement, when one considers the simplicity with which the detectors can be realized, it is a reasonable measurement to include in a low-cost speech recognition device. One form of the detector is shown in Fig. 9. The two summing resistors can be replaced by a potentiometer with the wiper arm connected to the output filter. Adjustment of this potentiometer then accomplishes the weighting of envelopes mentioned above.

TABLE I

Segmentation of Vocabulary

Spoken Word	Segmentation Pattern
one	V
two	PV
three	FwV
four	FwV
five	FwVFw
six	FsVFs
seven	FsV
eight	VP
nine	V
oh	V
plus	PVFs
minus	VFs
subtotal	FsVP
total	PVP
false	FwVFs

Shoebox Operation and Test Results

Several asymmetry detectors were used in Shoebox. One detector was used to segment the input words into voiced and unvoiced portions. Table I shows the words of the vocabulary along with the segmentation performed by the device. It should be noted that for unvoiced sounds a further classification is made. The fricative sounds are classified as strong friction (Fs), weak friction (Fw), or plosive (P). In the table, voicing is represented by V.

It can be noted from Table I that several words result in a distinct pattern after segmentation. These words are identified by the use of appropriate logic based on this segmentation pattern. There are certain words that possess a nonunique pattern. The words "one," "nine," and "oh" are composed of voicing only. The words "three" and "four" are both made up of a weak friction sound followed by voicing. These two classes require further subdivision to correctly identify each word.

Three asymmetry detectors are used to remove the uncertainties resulting from the segmentation pattern. One detector is adjusted to give positive asymmetry for the word "four" and negative asymmetry for "three." This measurement along with the segmentation pattern accurately differentiates between each of these words and all other words of the vocabulary. Two asymmetry detectors are used to identify the members of the remaining nonunique class. One detector is so adjusted to give a positive output for "one" and negative outputs for "nine" and "oh." The second detector is adjusted to result in a positive output for "oh" and a negative output for "nine." A suitable logic circuit follows the detectors to identify the results.

When Shoebox was demonstrated at the Seattle World Fair each operator had a particular list of settings for the phase shifter and weighting potentiometer of each asymmetry detector to optimize performance for his or her

TABLE II

Confusion Table for Shoebox

(Each word was spoken a total of 1960 times)

Spoken	Recognized														
	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Oh	Plus	Minus	Sub-total	Total	False
One	1907	24	5	1				2	15	5	1				
Two	20	1851	36	4			1	2	8	17				21	
Three	1	26	1880	14	16		8		4	9				2	
Four	3	4	72	1859	6		8		1	6				1	
Five		9	24	1	1882	1	6	10	1	1			2	2	21
Six					1	1944	2					2	1		10
Seven		2	17	1			1926			1	2	1	10		
Eight	4		5					1913	11	11		15		1	
Nine	21	3	4					4	1924	4					
Oh	77	8		2		1	1	23	27	1820				1	
Plus		5							1	1	1912	33		2	6
Minus		3	2					6	1		8	1936			4
Subtotal		2	15	5	3	3	28						1904		
Total	7	32	14	4	3			9	2	3	3		2	1881	
False				2	8	16				1	5	5			1923

voice. The set-up time for each operator was less than one minute.

The effectiveness of the asymmetry detector in vowel separation led to further speech research in this area by IBM. The Industrial Psychology Department conducted tests on Shoebox to determine the accuracy of identification of the words of the vocabulary for nonselected speakers. Fourteen subjects, seven male and seven female, participated in the tests. A short explanation of the device (approximately 15 minutes in duration) was presented to the subjects. Then each subject was allowed to speak to Shoebox while a technician adjusted the asymmetry detectors for optimum individual performance.

Each subject contributed 420 words per trial with each word occurring 28 times (in random order) per trial. A given subject ran through 5 trials giving a total of 70 trials or 29 400 words. There were a total of 938 errors corresponding to a recognition accuracy of 96.8 percent. The confusion table is presented as Table II.

The class of words consisting of voicing only, that is the "one," "nine," and "oh," have a total error rate of 4 percent. The "three" and "four" have a total error rate of 5 percent. However, these total rates are not an accurate measure of the effectiveness of the asymmetry detectors in separating vowel sounds. The segmenting scheme contributes to the total inaccuracy. The number of errors in identifying "three" and "four" due only to the asymmetry measurement is seen to be 86, giving an error rate of only 2 percent. The number of errors in identifying "one," "nine," and "oh" due only to asymmetry errors is 149, corresponding to an error rate of less than 3 percent. It should be emphasized that the subjects generating Table II are basically untrained spekers of both sexes. A well-trained speaker can generally obtain accuracies exceeding 98 percent with the 15-word vocabulary device.

The use of the asymmetry measurement has been extended to larger vocabulary machines. In this case, the detectors are only capable of classifying vowel sounds into broad subclasses. Further classification requires additional measurements. For small vocabulary speech recognition devices the asymmetry measurement is a very useful measurement.

Conclusions

Several conclusions regarding the usefulness of the asymmetry measurement can be made as follows.

1) The asymmetry measurement is a very gross mea-

sure of the relative magnitudes and phases of the lower harmonic components of a sound. These characteristics could be accurately measured by other means such as bandpass filters, but the amount of hardware required is considerably more.

2) The asymmetry measurement can be used to segment a word into voiced and nonvoiced portions of a word. This method of segmentation is quite accurate and varies little from speaker to speaker.

3) This measurement can be used as a vowel separator and can be so adjusted to be insensitive to volume changes. Again the separation is accomplished with a relatively small amount of hardware.

4) Vowel separation is somewhat inconsistent for different speakers. This problem is partially solved by allowing each speaker to optimize the effectiveness of the measurement. Optimization is accomplished by varying a phase shift control and a weighting potentiometer. These adjustments are extremely simple and direct. Vowel separation schemes involving more hardware are in general much more difficult to optimize.

ACKNOWLEDGMENT

The author would like to thank Dr. R. Hirsch and W. Emmons of the Industrial Psychology Department of the Advanced Systems Division of IBM for conducting the recognition tests reported here.

REFERENCES

[1] W. C. Dersch, "Shoebox—A voice responsive machine," *Datamation*, June 1962.
[2] ——, "The possibility of reducing the redundancy in speech," *Proc. Internat'l Congress on Technology and Blindness*, vol. 2, pp. 171–176, 1962.
[3] B. Gold and C. M. Rader, "Systems for compressing the bandwidth of speech," *IEEE Trans. Audio and Electroacoustics*, vol. AU-15, pp. 131–136, September 1967.
[4] D. J. Comer, "Large deviation phase and frequency modulators," *Electronic Engrg.*, vol. 39, pp. 495–497, August 1967.

David J. Comer (M'64) was born in Tuolumne, Calif., on January 10, 1939. He received the B.S. degree from San Jose State College, San Jose, Calif., the M.S. degree from the University of California, Berkeley, and the Ph.D. degree from Washington State University, Pullman, in 1961, 1962, and 1966, respectively, all in electrical engineering.

From 1959 to 1964 he was employed by the Advanced Systems Development Laboratory of IBM Corporation, San Jose, Calif. In 1964, he taught part-time at San Jose State College. From 1964 to 1966 he was an Assistant Professor of Electrical Engineering at the University of Idaho, Moscow. He is presently an Associate Professor of Electrical Engineering at The University of Calgary, in Calgary, Alberta, Canada. He holds two U. S. patents and is a Consultant in the area of digital circuits. He is the author of the textbook *Introduction to Semiconductor Circuit Design*. His fields of interest include semiconductor devices, circuit design, digital computer design, and machine recognition of speech.

Dr. Comer is a member of Tau Beta Pi, Phi Kappa Phi, the American Society for Engineering Education, and the American Association for the Advancement of Science.