AT&T Technical Journal
Vol. 64, No. 10, December 1985
Printed in U.S.A.

Single-Frame Vowel Recognition Using Vector Quantization With Several Distance Measures

By L. R. RABINER and F. K. SOONG*

(Manuscript received June 18, 1985)

One of the most fundamental concepts used in the standard pattern recognition model for speech recognition is that of distance between pairs of frames of speech. Several distance measures have been proposed and studied in the context of an overall speech recognizer. The purpose of this investigation was to isolate the effects of different distance measures in a recognizer from the other types of processing typically used in recognition. The way in which this isolation was achieved was to use a recognizer based on single-frame distance scores, using a vector quantization approach to give the single-frame reference patterns required by the recognizer. The vocabulary for recognition was the set of continuant vowels extracted from carrier words. A speaker-dependent vowel recognition experiment was carried out using seven talkers (four male, three female) and five distance measures. Results indicated that there were differences in performance for the different distance measures when the number of code-book patterns per vowel was one or two; however, when the number of code-book patterns was four or more, these differences in performance became insignificant.

I. INTRODUCTION

In the past several years, interest has focused on defining and studying distance measures for speech recognition that reflect meaningful differences between pairs of speech spectra.[1-7] Although several different distance measures have been proposed,[1-4] and they have been studied in a variety of recognition systems,[5-7] as yet there is little

* Authors are employees of AT&T Bell Laboratories.

consistency in the reported performance of different recognizers using different distance measures. For example, although Shikano and Sugiyama[3] found consistent recognition performance improvements using the weighted likelihood ratio distance measure (as opposed to an unweighted likelihood ratio distance measure) for a Japanese speech recognition system, Nocerino et al.[7] were not able to match these results in English alpha-digit recognition experiments. Similarly, although Davis and Mermelstein[6] achieved the best performance among several distance measures with a mel-based cepstral distance measure, this result has not been confirmed in other recognition tests.

There are several possible explanations for the discrepancies in results obtained in the various investigations of distance measure performance cited above. One explanation is that the basic feature measurements of each of the recognizers were different in all cases, for example, filter bank analysis versus Linear Predictive Coding (LPC), different recording conditions and bandwidths, etc. These differences in recognizer front ends could account for the differences in performance, but if this were the case then the robustness of the distance measure would become a major issue. A second explanation is the difference in vocabulary, talkers, and transmission conditions (e.g., telephone line versus microphone input). Again these differences could be important, but they should not be factors for a robust distance measure. A third explanation is that the experimental results did not just reflect differences in distance measures but were affected by the interaction between the components of the recognizer and the distance measure. Thus, for example, improved performance for a distance measure might be overshadowed by the power of dynamic time warping, which could compensate for a distance measure with poorer performance.

It is the purpose of this paper to investigate the last possibility discussed above—namely to isolate the effects of different distance measures from all the other temporal alignment processing used in recognition. The way in which we accomplish this goal is to design a recognizer that makes its decisions based on single frames of speech. In this manner any real differences in distance measures will manifest themselves as differences in recognition scores.

The implication of using single-frame distance scores for recognition is that the only vocabulary that can be considered is the set of continuant (steady) vowel sounds. We have considered ten such vowel sounds and they are listed in Table I, along with carrier words in which the vowels occur. One side benefit of the experiments to be reported here is that a range of performance scores for single-frame recognition of vowel sounds will be established and can be used to assess future recognition algorithms in much the same way as digit

Table I—List of vowel sounds and typical carrier words

Vowel	Carrier Word
ee	beet
I	bit
e	bet
ae	bat
a	father
uh	butt
ow	bought
oo	boot
er	Bert
U	foot

and alphabet scores have become standardized for isolated word recognition.[8]

Based on the above discussion a series of speaker-dependent recognition tests was performed in the following manner. Each of seven talkers (four male, three female) spoke the carrier words in Table I ten times each, in two separate recording sessions, over a dialed-up telephone line. Each talker also created, in a separate recording session, a single robust pattern for each of the ten carrier words. For diagnostic purposes, an isolated word recognition test was performed on the 100 isolated word tokens for each talker. All words which were misrecognized were manually checked to make sure that no recording errors (by either the talker or the automatic recording system) were made.

The way in which the vowel frames, of each of the ten recordings of each carrier word, were selected was as follows. The energy contour of the word was measured, and the vowel portion was defined as the set of frames whose log energies were contiguous to and within 6 dB of the global energy peak of the word. The first five replications of each carrier word were used as a training set, and a series of LPC Vector Quantization (VQ) code books were designed from the vowel frames for each vowel and for each talker. The second five replications were used as an independent test set for recognition purposes.

Five distance measures were used in the evaluations, namely the likelihood ratio;[1,2] the weighted likelihood ratio;[3] the cepstral distance;[5] a weighted cepstral distance;[9] and a bandpass liftered, weighted cepstral distance.[10]

The overall results of the single-frame recognition tests show that for speaker-trained code books with moderate size—that is, either four or eight vectors per vowel—there were no significant differences in performance for the five distance measures. For code books with one or two vectors per vowel, the two weighted cepstral distances per-

formed best; the likelihood ratio was third; the (unweighted) cepstral distance was fourth; the weighted likelihood ratio was last.

The outline of this paper is as follows. In Section II we discuss the speech analysis system, show how we extracted the vowel frames from each carrier word, review the process of creating VQ code books, and present the five distance measures used in our experiments. In Section III we summarize the experimental conditions and present the results of the word recognition tests, the code-book design, and the single-frame recognition tests. In Section IV we discuss the results and give general conclusions.

II. SINGLE-FRAME, VQ-BASED RECOGNITION SYSTEM

A block diagram of the single-frame, VQ-based recognition system is given in Fig. 1. For each vowel frame, an LPC analysis is performed to give either an LPC vector or an LPC derived cepstral vector. We denote the resulting vector as a. This vector is then passed to a series of ten vector quantizers (VQ's), one for each of the ten vowels, and the minimum VQ distance, ϵ^i, from the VQ for the ith vowel is computed as

$$\epsilon^i = \min_{1 \le m \le M} [d(\mathbf{a}, \mathbf{b}_m^i)], \tag{1}$$

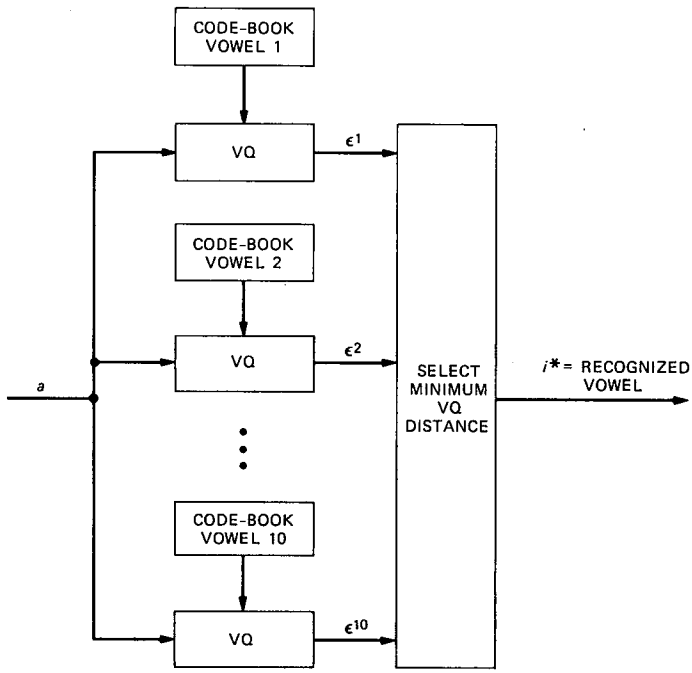

Fig. 1—VQ-based single-frame vowel recognizer.

where we assume that the ith vowel code book consists of the set of M vectors $\mathbf{b}_m^i$, $1 \leq m \leq M$. The local distance measure of eq. (1) can be any of five measures, namely the likelihood ratio; the weighted likelihood ratio; the (unweighted) cepstral distance; the weighted cepstral distance; and the bandpass liftered, weighted cepstral distance. The recognized vowel, i^*, is chosen as the one whose VQ distance ϵ^{i^*} is minimum, that is,

$$i^* = \underset{1 \leq i \leq 10}{\operatorname{argmin}} [\epsilon^i]. \tag{2}$$

In the following sections we briefly review the LPC analysis conditions, the method of extraction of vowel frames from carrier words, the procedure for VQ code-book formation, and the five distance measures used in this study.

2.1 LPC analysis conditions

The speech signal, $s(n)$, was recorded off a dialed-up, local, telephone line. We used a sampling rate of 6.67 kHz. The speech signal is digitized and then preemphasized using a first-order digital network with transfer function $H(z) = 1 - 0.95z^{-1}$. The signal is then blocked into frames of size $N = 300$ samples (45 ms), with consecutive frames spaced by L samples (15 ms). A Hamming window is applied to each speech frame and an eighth-order ($p = 8$) autocorrelation analysis is performed. The zeroth-order autocorrelation is the energy for the frame, and it is used as the basis for word detection[11] and energy normalization. An eighth-order LPC analysis is done on each frame, using the autocorrelation method of LPC,[12] to give the LPC vector for that frame. If a cepstral representation is required, a simple transformation of the LPC vector is performed.[12]

2.2 Extraction of vowel regions from carrier words

The way in which the vowel frames were extracted from the isolated word tokens is illustrated in Fig. 2. Basically we used the log energy contour of the word to find the vowel region—which was arbitrarily defined as the set of frames—in the vicinity of the maximum energy vowel frame, such that the log energy of each frame was within E_{DIF} (dB) of the vowel maximum energy, $E_{\max}$. After some preliminary experimentation, a value of $E_{\text{DIF}} = 6$ dB was used. Thus, for a typical carrier word, as illustrated in Fig. 2a, we first located the frame of maximum energy, t_v, and then, by searching in the local region around t_v, found the beginning, t_B, and ending, t_E, frames of the vowel. Although this procedure worked well, in general, there were some specific cases in which it failed. One such example is illustrated in Fig. 2b, in which the carrier word had a stop release at the end (e.g., boot) whose energy exceeded the maximum vowel energy. The simple strat-

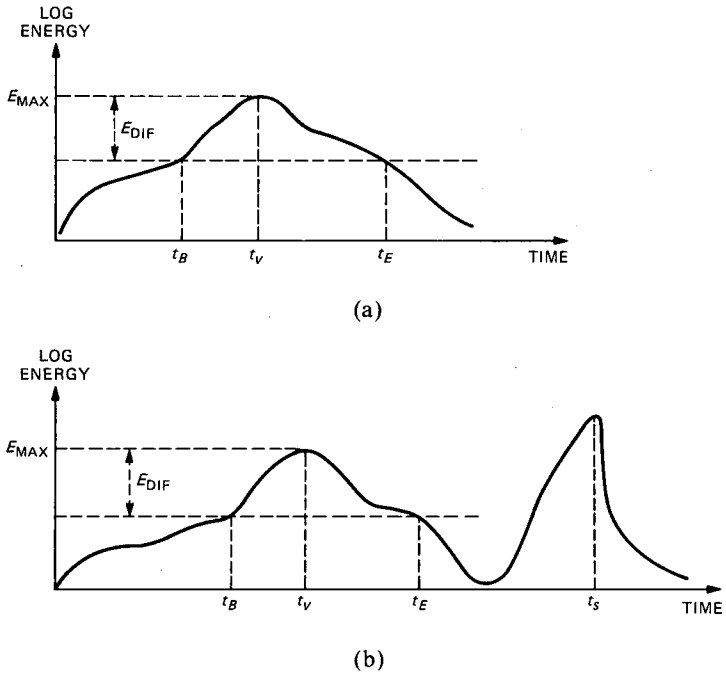

Fig. 2—Illustrations of how vowel frames were extracted from the carrier word.

egy of finding the frame with the maximum energy across the word would fail in this case. Hence a check was made to ensure that all strong local maxima of the energy contour were found, and that the correct vowel maximum was located.

2.3 VQ code-book design

The code-book training set for each vowel (and for each talker) consisted of all the "vowel frames" that occurred in five occurrences of each carrier word. In fact, there were between 35 and 90 training vectors for each vowel. From these training vectors a series of VQ code books were designed with 1, 2, 4, and 8 vectors per vowel, using a standard VQ code-book design algorithm.[13,14] The distance measure used in the code-book design was the same one used in the single-frame recognizer—that is, each of the five distance metrics was used. The centroid of the vectors in each cluster was chosen to represent the whole cluster. In our VQ design algorithm the centroid was chosen to minimize the average distortion of the whole cluster.[13]

2.4 Distance measures used in the recognizer

The five distance measures used in the recognizer included the following:

1. Likelihood ratio distance—$d_{\mathrm{LR}}(\mathbf{a}, \mathbf{b})$
2. Weighted likelihood ratio distance—$d_{\mathrm{WLR}}(\mathbf{a}, \mathbf{b})$
3. (Unweighted) cepstral distance—$d_{\mathrm{CEP}}(\mathbf{a}, \mathbf{b})$
4. Weighted cepstral distance—$d_{\mathrm{WCEP}}(\mathbf{a}, \mathbf{b})$
5. Bandpass liftered, weighted cepstral distance—$d_{\mathrm{BPCEP}}(\mathbf{a}, \mathbf{b})$.

The form for computation of the likelihood ratio is

$$d_{\mathrm{LR}}(\mathbf{a}, \mathbf{b}) = 2 \sum_{i=1}^{p} R_b(i)\hat{R}_{x_a}(i) + R_b(0)\hat{R}_{x_a}(0) - 1, \tag{3}$$

where **a** and **b** are the LPC vectors being compared, and

$$R_b(i) = \sum_{j=0}^{p-i} b(j)b(j+i), \qquad 0 \le i \le p \tag{4}$$

$$R_{x_a}(i) = \sum_{n=0}^{N-1-i} x_a(n)x_a(n+i), \qquad 0 \le i \le p \tag{5}$$

$$\hat{R}_{x_a}(i) = \frac{R_{x_a}(i)}{\alpha}, \tag{6}$$

where α is the residual error of the LPC analysis of the frame with autocorrelation $R_{x_a}(i)$.

The form for computation of the weighted likelihood ratio[3] is

$$d_{\mathrm{WLR}}(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{q} \left[\frac{R_{x_a}(i)}{R_{x_a}(0)} - \frac{R_{x_b}(i)}{R_{x_b}(0)}\right][c_a(i) - c_b(i)], \tag{7}$$

where $R_{x_a}(i)$ and $R_{x_b}(i)$ refer to the signal autocorrelations of the frames corresponding to vectors **a** and **b**, and $c_a(i)$ and $c_b(i)$ are the corresponding LPC-derived cepstral vectors. It should be noted that we use $q > p$ terms, in the summation of eq. (7), to approximate the infinite summation of the true weighted likelihood ratio distance. In particular we have used $q = 2p$ (16), where the "extended" autocorrelations and cepstra were derived from the so-called "maximum entropy" extension of the first $(p + 1)$ terms.[15]

The form for the (unweighted) cepstral distance is

$$d_{\mathrm{CEP}}(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{q} [c_a(i) - c_b(i)]^2, \tag{8}$$

where we have again used the cepstrum extended to $q = 2p$ terms. The form for the weighted cepstral distance[9] is

$$d_{\mathrm{WCEP}}(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{p} w_i[c_a(i) - c_b(i)]^2, \tag{9}$$

where

$$w_i = \sigma_1^2/\sigma_i^2 \tag{10}$$

and σ_i^2 is the sample variance of the ith cepstral coefficient, where the averaging is over the individual vowel sounds, that is,

$$\sigma_i^2 = \frac{\sum_{v=1}^{10} [\sigma_i^2]_v \cdot n_v}{\sum_{v=1}^{10} n_v} \tag{11}$$

with $[\sigma_i^2]_v$ being the variance of $c(i)$ over the n_v frames in the training set for vowel v. Typically the weighting function w_i increases monotonically with the index i.

Finally, the form for the bandpass liftered, weighted cepstral distance[10] is

$$d_{\text{BPCEP}}(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{q} w_i'[c_a(i) - c_b(i)]^2, \tag{12}$$

where q was set to 12, and w_i' had the form of a bandpass lifter, that is, a raised sinewave of the form

$$w_i' = 1 + 6 \sin\left(\frac{\pi i}{12}\right). \tag{13}$$

III. EXPERIMENTAL EVALUATIONS AND RESULTS

A series of recognition tests was run in which each of seven talkers (four male, three female) first created robust training tokens of each carrier word[16] and then, in separate recording sessions, spoke each carrier word ten times each. The first five such recordings were used as a training set for the VQ code books; the second five recordings were used as an independent test set. The robust tokens were used in an isolated word recognition test to check the validity of the recorded carrier words. The results of the isolated word recognition test are given in Table II. It can be seen that for three of the talkers (1, 4, and

Table II—Word recognition errors for carrier words for each talker (100 recognition trials per talker)

	Talker						
Word	1(M)	2(M)	3(M)	4(F)	5(M)	6(F)	7(F)
beet							
bit			6		2		
bet		2	1		3		
bat			2		1		1
father							
butt			1		1		1
bought					1		
boot							
Bert							
foot							
TOTALS	0	2	10	0	8	0	2

6) there were no word errors; for talkers 2 and 7 there were 2 word errors (out of 100 trials each); for talkers 3 and 5 there were 10 and 8 word errors. The overall isolated word recognition accuracy for the seven talkers is 96.9 percent. The word "bit," which accounted for 8 of the 22 recognition errors, was confused with the word "bet" in all such cases.

The results given in Table II indicate that there is a lot of variability in the recognition performance on the isolated words across both talkers and vocabulary words.

3.1 Single-frame vowel recognition results

The results of the single-frame vowel recognition tests are given in Table III and are shown plotted in Fig. 3. The data in Table III are the average vowel error rates in percent averaged over the ten vowels and the seven talkers as a function of VQ code-book size and distance measure for both the training and testing sets, that is, there were about 4000 recognitions per set. Figure 3 shows these same data in graphical form. Several observations can be made from these results, including the following:

1. There are significant degradations in performance, for all distance measures and for all code-book sizes, between the training and testing sets of data. Thus for the VQ code-book size of one we see degradations of 3 to 4 percent, whereas for the VQ code-book size of eight we see degradations of from 9 to 10 percent in averaged vowel error rates.

2. The effects of different distance measures can be seen primarily for code-book sizes of one and two vectors per vowel, in which case the two weighted cepstral distances consistently outperformed the other three metrics, and the weighted likelihood ratio consistently performed the worst of the five measures. For code-book sizes of four and eight vectors per vowel, there were no significant performance differences among the five distance measures.

3. For the independent test set there was an average vowel error

Table III—Average word error rate (%) as a function of VQ code-book size and distance measure for both the training and testing sets

Distance Measure	Results on Training Set: VQ Code-Book Size 1	2	4	8	Results on Testing Set: VQ Code-Book Size 1	2	4	8
d_{LR}	17.6	12.2	7.0	3.4	21.6	16.9	14.1	12.9
d_{WLR}	18.8	13.6	7.2	3.8	22.4	18.9	14.2	12.5
d_{CEP}	18.6	11.8	7.1	3.5	21.6	17.4	13.7	13.4
d_{WCEP}	16.5	11.0	6.7	3.8	20.0	16.5	14.2	13.3
d_{BPCEP}	16.7	10.5	6.4	3.2	19.4	15.5	13.4	12.4

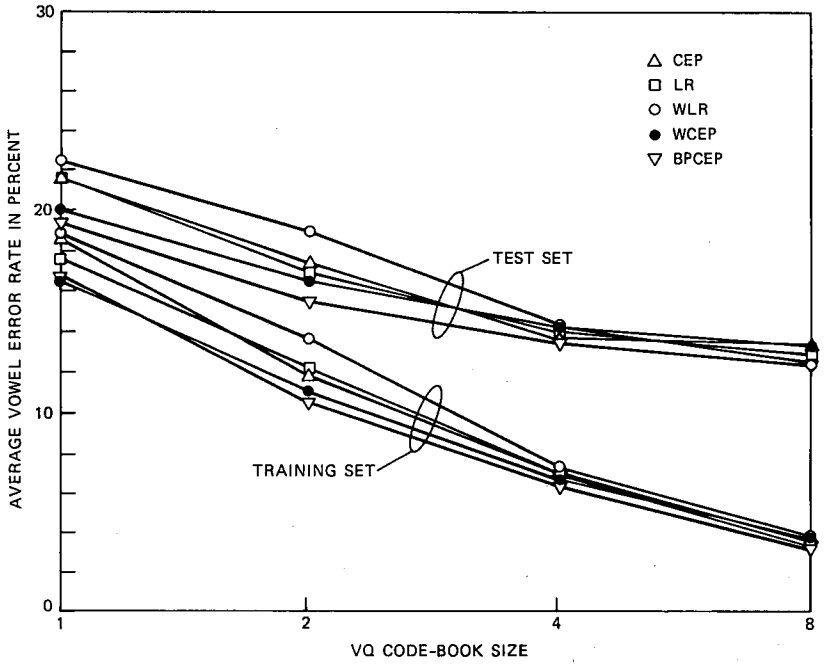

Fig. 3—Average vowel error rate (%) versus code-book size for each of the four distance measures and for the testing and training sets of data.

rate of about 20 to 22 percent for a single code-book vector per vowel, and the error rate dropped to about 13 percent for eight code-book vectors per vowel. Thus we conclude that vowel recognition (among the ten vowels in Table I) cannot be performed reliably using any of the distance measures we have considered, in the framework of a single-frame VQ code book-based recognizer.

IV. DISCUSSION AND CONCLUSIONS

The results presented in Section III can be interpreted as follows. In the case where we have a good representation of the patterns to be recognized, the effects of different distance measures on recognition performance are small. Such was the case when we used four or eight code-book vectors to characterize each vowel in the vocabulary. However, when the representation of the patterns to be recognized becomes more coarse, then the effects of different distance measures start to become important. In these cases a better characterization of speech sound differences, as obtained from a good distance measure, should give better recognition scores. Such was the case when we used one or two code-book vectors to characterize each vowel.

There is another important observation that can be made from the

results presented in Table III. We see a big difference in average vowel error rates between comparable test conditions (distance measure, VQ code-book size) for the training and testing sets, especially when we have four or eight vectors per vowel code book. Thus, in a sense, the effects of different distance measures are small when the code-book vectors begin to characterize well the seemingly insignificant details of the training set, and are larger when the code-book vectors characterize mainly the gross spectral behavior of the vowels. For real-world recognition systems it is most probably the latter case that is the more important one in that the reference patterns would be expected to characterize the gross behavior of spectral variations with time. In general there is not enough training data to reach the point where we have characterized the fine spectral variations of words reliably.

The conclusion we reach from the above discussion is that the results for small code-book sizes, in which there were significant effects of different distance measures, are probably more representative of real recognition systems than the results for large code-book sizes. In these cases—as is evidenced by recent investigations by Tokhura,[9] Juang et al.,[10] and Nocerino et al.,[7]—the weighted cepstral distances and the likelihood ratio would be expected to give better recognition performance than the unweighted cepstral distance or the weighted likelihood ratio measures.

V. SUMMARY

We have presented results on speaker-dependent, single-frame, VQ-based, vowel recognition for five different distance measures and for four different size VQ code books. Our results indicate that for small code-book sizes (one or two vectors per vowel) there is improved recognition performance using a weighted cepstral distance rather than the likelihood ratio, the unweighted cepstral distance, or the weighted likelihood ratio measures. For larger code-book sizes (four or eight vectors per vowel) the performance differences among the five distance measures decrease. For practical recognizers, the weighted cepstral distances appear to have advantages for application to speaker-independent systems and for large vocabulary recognizers. These advantages include increased efficiency of representation, reduced complexity of computation, and improved performance.

REFERENCES

1. F. Itakura and S. Saito, "Analysis-Synthesis Telephony Based on the Maximum Likelihood Method," Proc. Int. Cong. Acoustics, Tokyo, Japan, Paper C5-6, 1968.
2. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoust., Speech, Signal Process., *ASSP-23*, No. 1, (February 1975), pp. 67–72.

3. K. Shikano and M. Sugiyama, "Evaluation of LPC Spectral Matching Measures for Spoken Word Recognition," Trans. IECE, J65-D, No. 5 (May 1982), pp. 535–41.
4. D. H. Klatt, "Prediction of Perceived Phonetic Distance From Critical Band Spectra: A First Step," Proc. ICASSP '82, (May 1982), pp. 1278–81.
5. A. H. Gray Jr. and J. D. Markel, "Distance Measures for Speech Processing," IEEE Trans. Acoust., Speech, Signal Process., *ASSP-24,* No. 5, (October 1976), pp. 380–91.
6. S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Trans. Acoust., Speech, Signal Process., *ASSP-28,* No. 4, (August 1980), pp. 357–66.
7. N. Nocerino et al., "Comparative Study of Several Distortion Measures for Speech Recognition," Speech Commun., *4,* No. 4 (November 1985).
8. G. R. Doddington and T. B. Schalk, "Speech Recognition-Turning Theory to Practice," IEEE Spectrum, *18* (September 1981), pp. 26–32.
9. Y. Tokhura, "Speaker Independent Recognition of Isolated Digits Using a Weighted Cepstral Distance," J. Acoust. Soc. Am., Suppl. 1, *77,* Paper E13 (Spring 1985), p. S11.
10. B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the Use of Bandpass Liftering in Speech Recognition," unpublished work.
11. L. F. Lamel et al., "An Improved Endpoint Detector for Isolated Word Recognition," IEEE Trans. Acoust., Speech, Signal Process., *ASSP-29,* No. 4 (August 1981), pp. 777–85.
12. J. D. Markel and A. H. Gray Jr., *Linear Prediction of Speech,* Springer-Verlag, 1976.
13. Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantization," IEEE Trans. Comm., *COM-28,* No. 1 (January 1980), pp. 84–95.
14. B. H. Juang, D. Wang, and A. H. Gray, Jr., "Distortion Performance of Vector Quantization for LPC Voice Coding," IEEE Trans. Acoust., Speech, Signal Process., *ASSP-30,* No. 2 (April 1982), pp. 294–303.
15. J. P. Burg, "Maximum Entropy Spectral Analysis," PhD. Thesis, Stanford Univ., 1975.
16. L. R. Rabiner and J. G. Wilpon, "A Simplified, Robust Training Procedure for Speaker Trained, Isolated Word Recognition Systems," J. Acoust. Soc. Am., *68,* No. 5 (November 1980), pp. 1271–6.

AUTHORS

Lawrence R. Rabiner, S.B. and S.M., 1964, Ph.D., 1967 (Electrical Engineering), The Massachusetts Institute of Technology; AT&T Bell Laboratories, 1962—. From 1962 through 1964 Mr. Rabiner participated in the cooperative plan in electrical engineering at AT&T Bell Laboratories, in Whippany and Murray Hill, New Jersey. He worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently, Mr. Rabiner is engaged in research on speech communications and digital signal processing techniques. He is coauthor of *Theory and Application of Digital Signal Processing* (Prentice-Hall, 1975), *Digital Processing of Speech Signals* (Prentice Hall, 1978), and *Multirate Digital Signal Processing* (Prentice-Hall, 1983), Member, National Academy of Engineering, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, Fellow, Acoustical Society of America, IEEE.

Frank K. Soong, B.S., 1973, National Taiwan University, M.S., 1977, University of Rhode Island, Ph.D., 1983, Stanford University, all in Electrical Engineering; AT&T Bell Laboratories, 1982—. From 1972 to 1975 Mr. Soong served as a teacher at the Chinese Naval Engineering School at Tsoying, Taiwan. In 1982 he joined the technical staff at AT&T Bell Laboratories, where he engaged in research in speech, coding, and speaker recognition. Member, IEEE.