REPORT:

Frank K. Soong, Aaron E. Rosenberg, *and* ***Biing-Hwang Juang*** *are all members of the technical staff in the Speech Research Department of AT&T Bell Laboratories in Murray Hill, New Jersey.* ***Lawrence R. Rabiner*** *is head of the Speech Research Department. Mr. Soong joined AT&T in 1982 and does research work on speaker recognition as well as speech coding and spectral estimation. He holds a B.S.E.E. from National Taiwan University, an M.S.E.E. from the University of Rhode Island and a Ph.D. in electrical engineering from Stanford University. Mr. Rosenberg joined AT&T in 1964 and is working on signal processing and pattern recognition techniques for automatic speech and speaker recognition. He holds B.S. and M.S. degrees in electrical engineering from the Massachusetts*
(continued on page 26)

A VECTOR QUANTIZATION APPROACH TO SPEAKER RECOGNITION

Introduction

Automatic speaker recognition has long been an interesting and challenging problem to speech researchers.[1-10] The problem, depending on the nature of the final task, can be classified into two different categories: speaker verification and speaker identification. In a *speaker verification* task, the recognizer is asked to verify an identity claim made by an unknown speaker and a decision to reject or accept the identity claim is made. In a *speaker identification* task, the recognizer is asked to decide which out of a population of N speakers is best classified as the unknown speaker. The decision may include a choice of "no classification" (i.e., a choice that the specific speaker is not in a given closed set of speakers). The input speech material used for speaker recognition can be either text dependent (constrained text) or text independent (free text).

In the text-dependent mode, the speaker speaks a prescribed text. The utterance is then compared with prestored patterns of the same text. A fairly short utterance (usually one sentence long) is adequate for text-dependent speaker recognition. In the text-independent mode, the speaker can, in theory, speak any speech material with no constraint. The unconstrained speech input is compared with models that characterize the speaker's features. In general, because of the higher acoustic-phonetic variability of text-independent input, more training material is required to characterize (model) a speaker reliably than in the text-dependent mode.

Recognition procedures generally are quite different for text-dependent and text-independent systems. In the text-dependent mode, samples of the same speech events in reference and test utterances are compared, usually by establishing a nonlinear time alignment between the utterances. There is no possibility of time alignment in true text-independent systems. Speaker characterization is carried out either by statistical averaging over selected acoustic features, or by locating comparable speech events in test and reference utterances.

In this paper we propose a new approach to the speaker recognition problem. The approach is partially motivated by the success of two word-based *vector quantization* (VQ) speech recognition systems. In one, Pan, Soong, and Rabiner [11] used word-based VQ codebooks in an isolated word recognition preprocessor to substantially alleviate the computational complexity of a dynamic programming-based speech recognition system. In the other, Shore and Burton [12] used word-based VQ codebooks and reported good performance in speaker-trained isolated-word recognition experiments. However, instead of using word-based VQ codebooks to characterize the phonetic contents of isolated words, we propose using *speaker-based* VQ codebooks to characterize the variability of short-time acoustic features of speakers. (Related studies are reported in References 13-16.)

Speaker-based VQ Codebook Approach

A set of short-time raw feature vec-

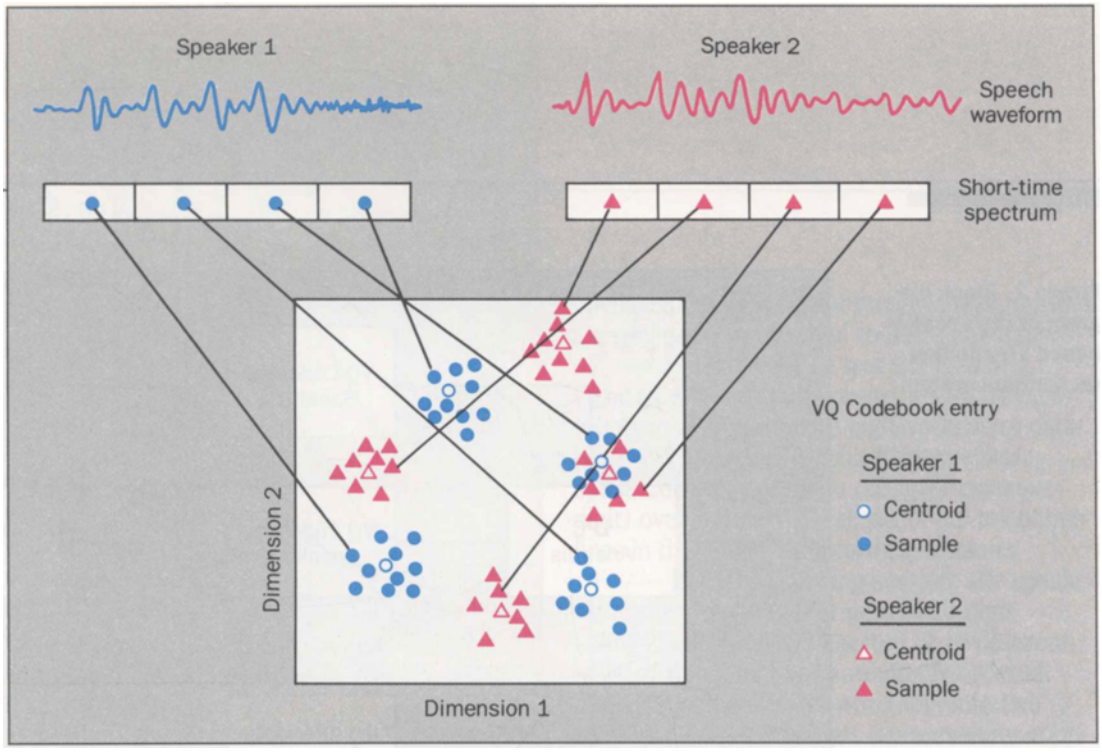

Figure 1. Conceptual diagram illustrating codebook formation. One speaker can be discriminated from another based on the location of centroids.

tors of a speaker can be used directly to represent the essential characteristics of that speaker. These can be acoustical, such as the short-time spectrum, or physiological, such as vocal tract shape measurements, if the training set includes sufficient variations. However, such a direct representation is not practical when the number of training vectors is large. The memory requirements for storage and computational complexity in the recognition phase eventually become prohibitively high. Therefore, an efficient way of compressing the training data has to be found. To compress the original data to a small set of representative points, we used a VQ codebook consisting of a small number of entries as an efficient means of characterizing speaker-specific features.

The speaker-based VQ codebook was generated as follows: Given a set of I training feature vectors, $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_I\}$ characterizing the variability of a speaker, we want to find a partitioning of the feature vector space, $\{S_1, S_2, \ldots, S_M\}$, for that particular speaker where S, the whole feature space, is represented as $S = S_1 \cup S_2 \cup \ldots \cup S_M$. Each partition, S_i, forms a nonoverlapping region and every vector inside S_i is represented by the corresponding centroid vector, $\mathbf{b}_i$, of S_i. The partitioning is done in such a way that the average distortion

$$D = \frac{1}{I} \sum_{i=1}^{I} \min_{1 \le j \le M} d(\mathbf{a}_i, \mathbf{b}_j) \qquad (1)$$

is minimized over the whole training set. The distortion (distance) between the vectors $\mathbf{a}_i$ and $\mathbf{b}_j$ is denoted as $d(\mathbf{a}_i, \mathbf{b}_j)$.

This VQ codebook representation of a speaker's feature space is useful in two ways. The speaker's features can be represented most efficiently for a given data rate. Successful VQ applications to low-bit-rate speech coding have confirmed this. The VQ representation also is useful in discriminating one speaker's voice from the other speaker's voice.

To illustrate this point more clearly, we use the conceptual diagram shown in Figure 1. The figure uses a two-dimensional Euclidean space to represent the short-time spectral fea-

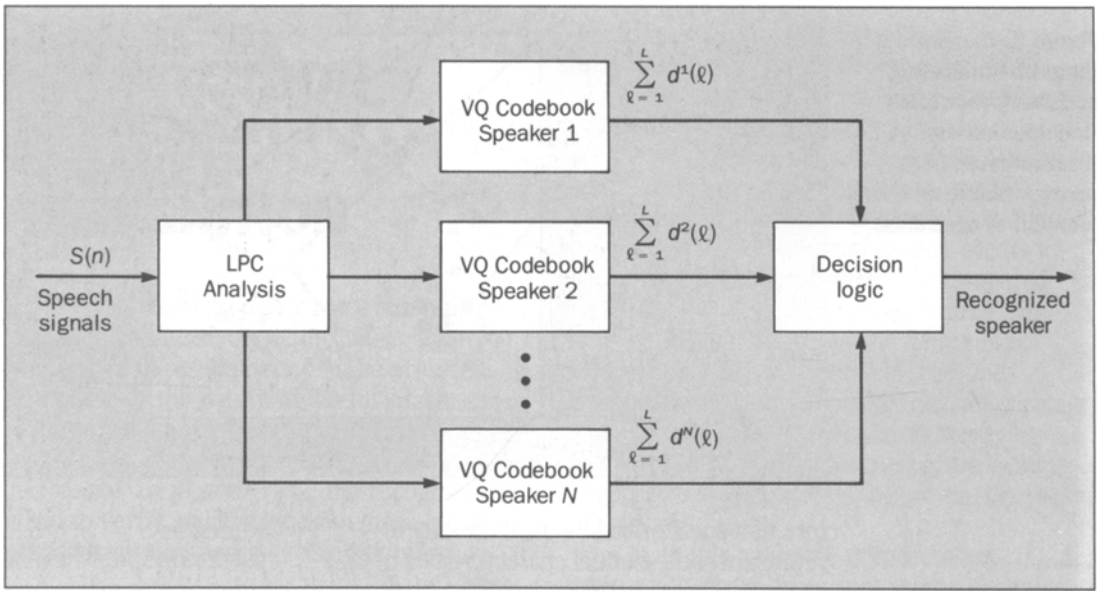

Figure 2. Block diagram of the speaker-based VQ speaker recognition system.

tures. First, the input speech signals of two speakers are analyzed and used to train the VQ codebooks. After the partitioning, we generate two codebooks, each with 4 entries. The VQ codebook entries, denoted as O and Δ, are the centroids of the corresponding partitioning of each speaker's feature space. There are some overlaps between two clusters belonging to the two different speakers. However, with the multi-partitioning of each individual speaker's feature space, the remainder of the centroids are still well separated. Even with some possible overlaps, the two VQ codebook representations are capable of discriminating one speaker from the other.

In this study we use short-time linear prediction coefficient (LPC) vectors as feature vectors. The corresponding distortion measure that measures the similarity between any two feature vectors is the LPC likelihood-ratio-distortion measure. The likelihood ratio distortion between two LPC vectors **a** and **b** is defined as

$$d_{LR}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{b}^T \mathbf{R}_a \mathbf{b}}{\mathbf{a}^T \mathbf{R}_a \mathbf{a}} - 1 \qquad (2)$$

where $\mathbf{R}_a$ is the autocorrelation matrix of speech input data associated with the vector **a**. Using this distortion measure, and the VQ codebook training algorithm proposed by Linde, Buzo, and Gray,[13] we generated speaker-based VQ codebooks of different sizes.

The *speaker identification* system based on this VQ codebook approach is shown in Figure 2. The input speech signal is sampled, endpointed, and LPC analyzed giving the sequence of vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_L$. The resultant LPC vectors are vector quantized

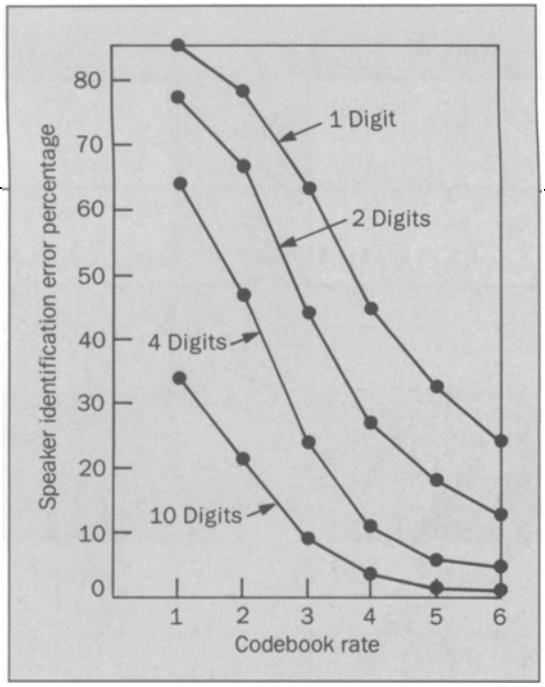

Figure 3. Speaker identification error percentage versus codebook rate for test sequences of 1, 2, 4, and 10 digits.

(encoded) using N codebooks corresponding to N different speakers. The quantization errors (or distortions) with respect to each codebook are individually accumulated across the whole test token. The average distortion with respect to the ith codebook (speaker) is

$$D^i = \frac{1}{L} \sum_{\ell=1}^{L} \min_{1 \le j \le M} d(\mathbf{a}_\ell, \mathbf{b}_j^i) \qquad (3)$$

The N resultant average distortions are compared to find the minimum. The final speaker recognition decision is given by

$$i^* = \underset{1 \le i \le N}{\operatorname{argmin}} D^i \qquad (4)$$

A *speaker-verification* system has a similar structure; however, only the codebook of the claimed identity is used and the resultant average distortion is compared with a preset threshold to reject or to accept the identity claim made by the unknown speaker.

Database, LPC Analysis, and Experiments

Because collecting a truly linguistically unconstrained database of many speakers is not a trivial task, we decided to use a much more constrained database to test the idea first. Over a period of two months, a 100-speaker (50 male and 50 female) digit-vocabulary database was collected. Each of 100 speakers spoke 200 isolated digits (20 utterances per digit) over ordinary, local, dialed-up telephone lines in five different recording sessions.

In each recording session, the speaker was asked to speak four sets of 10-digit strings. Each string consisted of 10 different, isolated digits that were randomly ordered. The 200 isolated digits were split into two parts. The first 100 digits were used for training (codebook generation) while the second 100 digits were used for testing (recognition). The recognition experiments were performed on a digit-independent basis. That is, the speaker was allowed to say any random, isolated digit string.

The analog speech input samples were first bandlimited from 200 Hz to 3200 Hz and then sampled at a 6.67-kHz sampling rate. The speech samples were pre-emphasized by a first-order filter whose transfer function was $H(z) = 1 - 0.95z^{-1}$. A 45-millisecond (ms) Hamming window was used to window the pre-emphasized speech data and an 8th-order autocorrelation analysis was performed. The resultant 9 autocorrelation coefficients were then used to find the LPC feature vector that represented the short-time LPC spectral information of the corresponding frame. The LPC analysis was performed every 15 ms (i.e., 30-ms overlapping between adjacent frames was used).

The following experiments were performed to evaluate the effects of different speaker recognition parameters on the recogni-

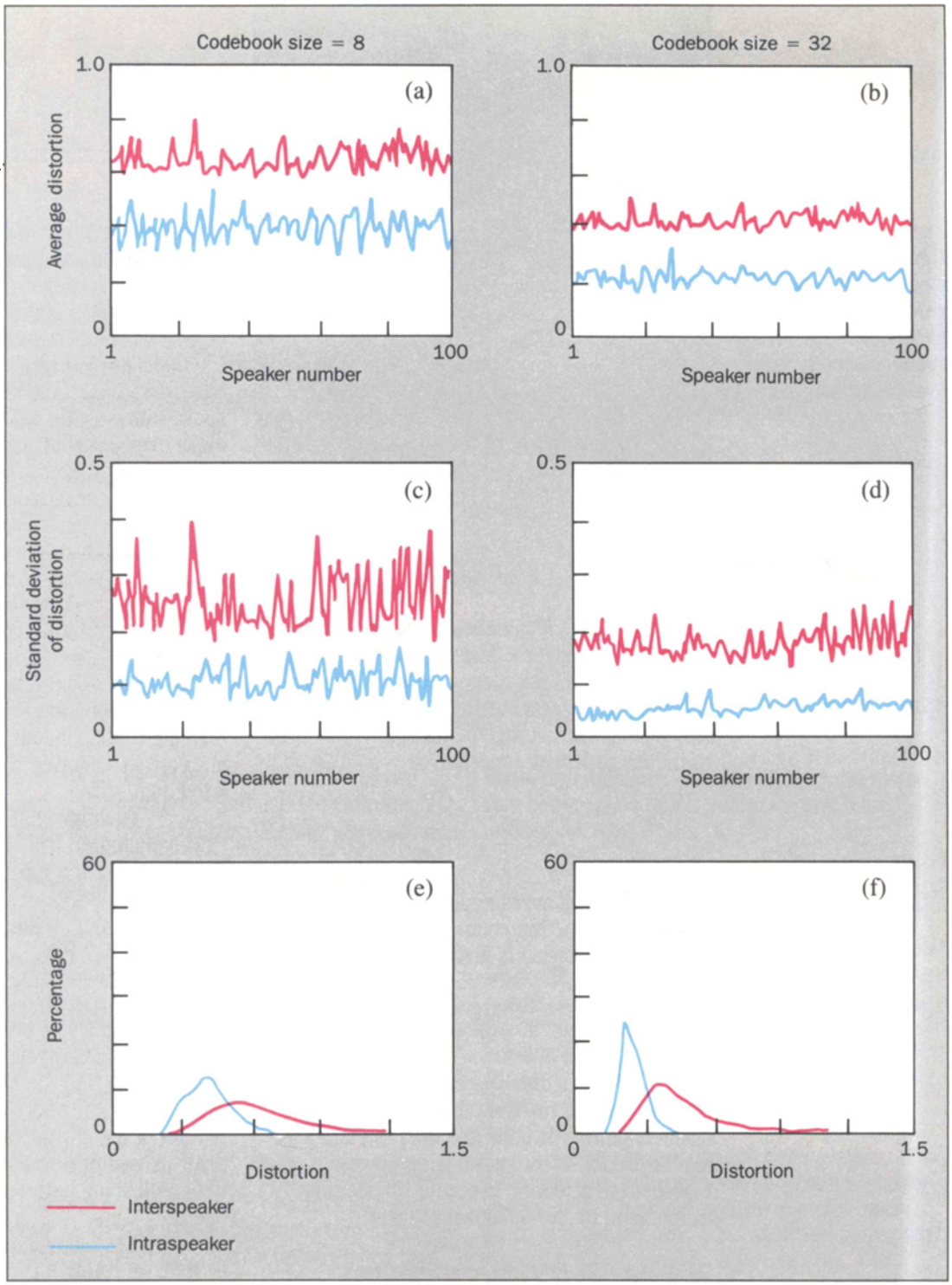

Figure 4. The effects of different codebook sizes on the mean and the standard deviation of the VQ distortions.

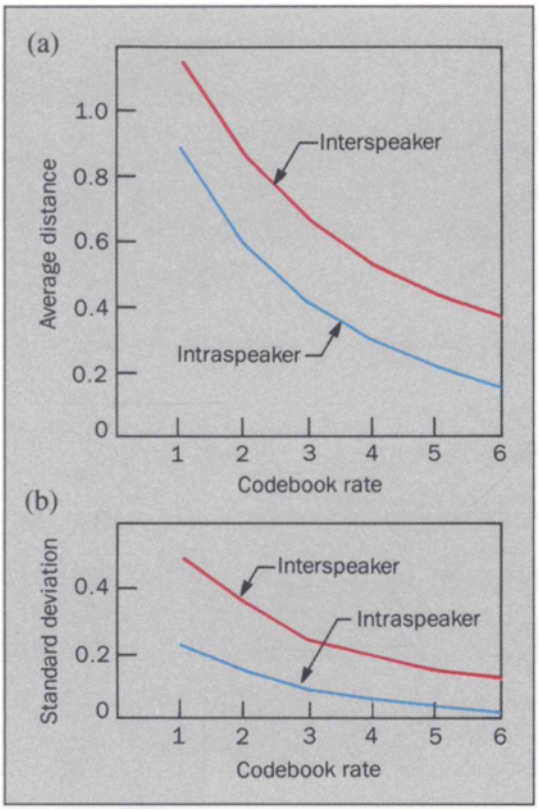

Figure 5. (a) Average distortion versus codebook rate. (b) Standard deviation of the distortion versus codebook rate.

tion performance. In particular we varied:

- *The size of the VQ codebook.* Six codebooks with 2, 4, 8, 16, 32, and 64 codebook vectors were used.
- *The number of digits in a test utterance.* Test utterance strings of one, two, four, and ten random (but different) isolated digits were used in the 100 speaker identification experiments.
- *The digits used in the test utterances.* Sets of 10 repeated digits were used as test utterances. The results were compared with the results obtained by using 10 random but different digits as test utterances.
- *The voiced/unvoiced frames used in the recognizer.* Voiced frames were extracted from the test utterance by using a simple but effective voiced/unvoiced detector based on frame energy and prediction gain. The results were compared with the results obtained when all speech frames were used.
- *The time span between the training and testing material.* The 100 test utterances were divided into 3 different groups according to the recording session sequence. This experiment was designed to study whether the intraspeaker variations and possible channel variations of different recording sessions affected the recognizer performance.
- *Speaker verification experiments.* The speaker identification experiments described above are useful in studying the effects of different parameters and different speech material used. However for most commercial applications, speaker verification systems seem to be more appropriate. We therefore investigate the performance of this VQ-based speaker recognition system in a verification mode.

Results

Effects of Codebook Size and Number of Digits in the Test Token. The speaker identification error percentage is plotted as a function of codebook size in Figure 3. The codebook size is represented by the number of vector entries, M, or by the corresponding codebook rate, $R = \log_2 M$. Four curves corresponding to four

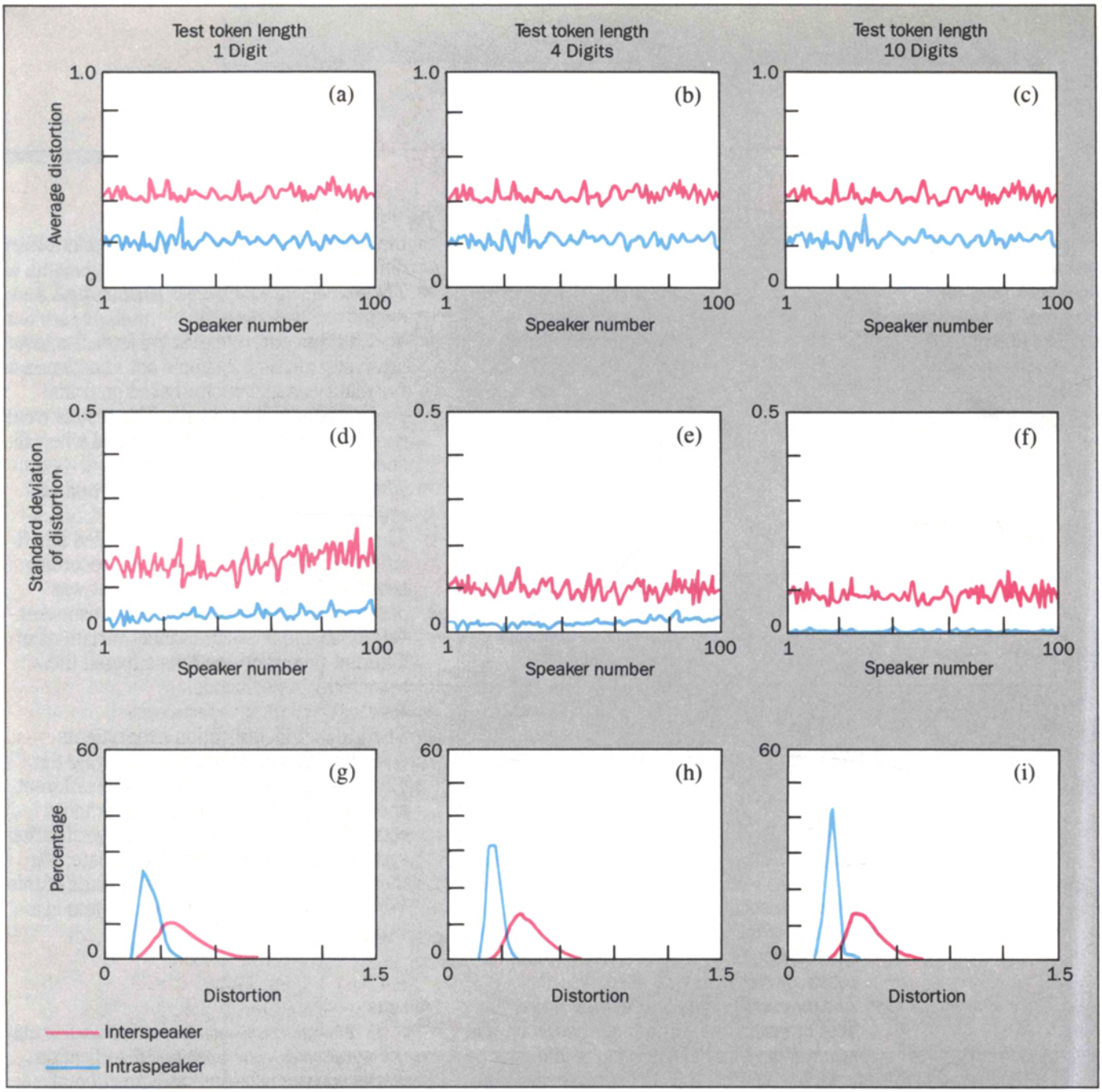

Figure 6. The effects of different test utterance lengths on the mean and standard deviation of the VQ distortion.

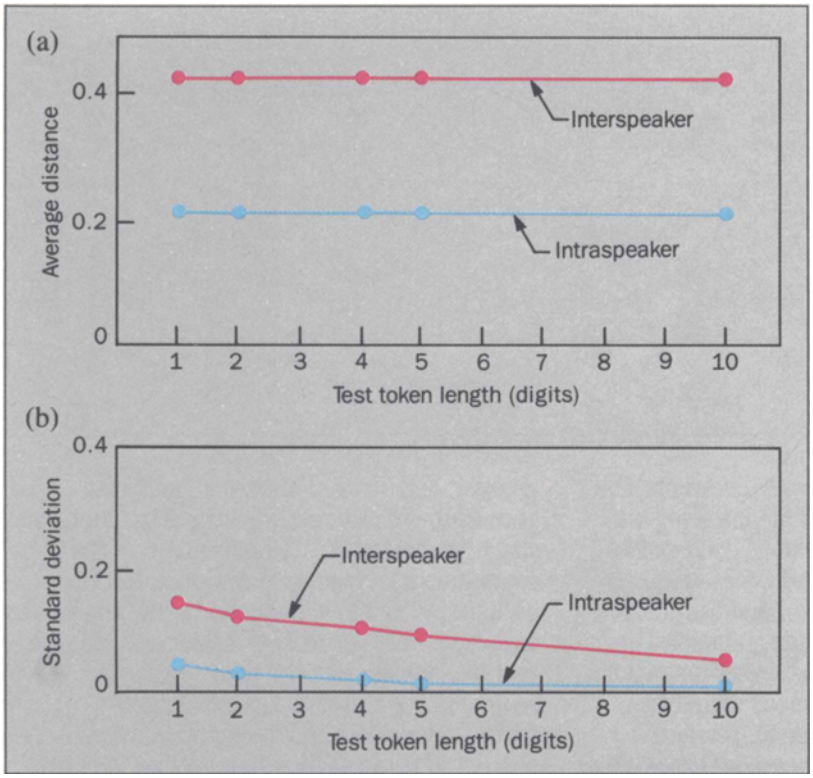

Figure 7. (a) Average distortion versus test utterance length. (b) The standard deviation of the distortion versus test utterance length.

different test utterance lengths (i.e., 1 digit, 2 digits, 4 digts, and 10 digits) are shown in the figure. The identification error rate decreases when either the codebook size or the test token length increases. For a test token of 10 isolated, different digits, the error rate dropped from 34 percent to 1.5 percent when the codebook size was increased from 2 to 64 vectors. With a codebook of 64 vectors, the error rate dropped from 24 percent to 1.5 percent when the test token length was increased from 1 to 10 digits.

Figure 4 further illustrates the effects of codebook size on the recognition error rate. The figure shows the averages [4(a) and 4(b)], standard deviations [4(c) and 4(d)], and histograms [4(e) and 4(f)] of interspeaker and intraspeaker distortion obtained from the speaker identification experiment of 100 speakers. Results using two different codebooks [i.e., 8 vectors (rate 3) and 32 vectors (rate 5)] are given in this figure. The average distortions of both codebooks are fairly constant across all 100 speakers and only a small amount of random fluctuation is observed. The separations between the inter- and the intraspeaker average distortions do not change appreciably when the codebook size varies.

On the other hand, the standard deviations of both the inter- and intraspeaker distortions reduce more noticeably when the codebook size increases from 8 to 32. As a result, the overlapping area between the two histograms in Figures 4(e) and 4(f) is decreased and the recognition error rate is reduced. The separations between the inter- and intraspeaker average distortions stay relatively constant for different codebook sizes. Actually, they decrease slightly with the codebook sizes as illustrated in Figure 5(a). The reductions of the corresponding standard deviations are shown in Figure 5(b).

In Figure 6 the averages, standard deviations, and histograms of inter- and intraspeaker distortions of three different test utterance lengths (1 digit, 4 digits, and 10 digits) are shown. A codebook of 32 vectors was used for all cases. The average distortions per frame are independent of test utterance length and hence are identical for all three different lengths.

However, the standard deviations of

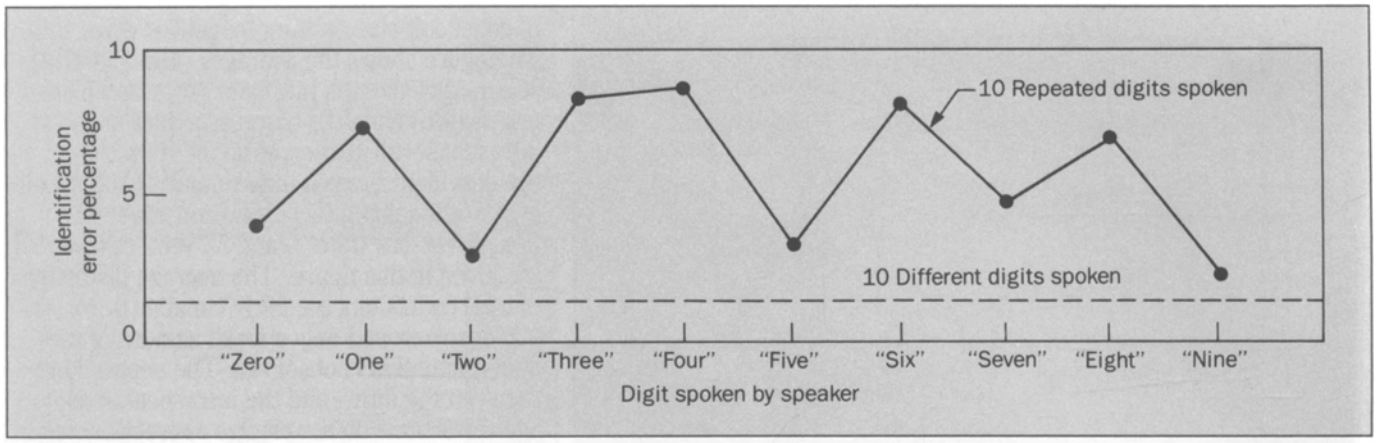

Figure 8. Identification error percentage versus digit uttered by the speaker for repeated digits and different digits.

the distortions are significantly different. The standard deviations of both the intraspeaker and the interspeaker distortions decrease as the number of digits in a test utterance increases [Figures 6(d), 6(e), and 6(f)]. The averages and the corresponding standard deviations of the inter- and intraspeaker distortions versus different test utterance lengths are depicted in Figures 7(a) and (b), respectively.

Effects of Repeated Digits and Different Digits. The speaker identification results obtained using 10 repeated digits are shown in Figure 8. The digit "9" achieved the best results; the digits "3," "4," and "6," provided the worst scores. We believe that the digit "9" produced favorable results because of its relatively long duration and its strong nasal-vowel coarticulation. Nasal-vowel coarticulation has long been known to be effective for speaker recognition purposes.[18] Nasal-vowel coarticulation is unlikely to be modified consciously by a speaker. Therefore, it serves as a reliable speaker identification feature with inherently low intraspeaker variability.

Because it has a stop gap with silence frames, the digit "6" does not achieve good recognition scores. This is not surprising because the short-time spectra in the stop gap carry no speaker-related information. The short-time spectrum of the silence (or the background noise) is matched by any codebook with a possibly very large variance. Thus, the average distortion of the true speaker corrupted by the aforementioned random perturbation makes misrecognition more likely.

It is interesting to note that the performance with 10 different digits (shown as the dashed line in Figure 8) is better than the performance with any 10 repeated digits. This result indicates that statistically less correlated information could, and did, improve the speaker recognition performance.

Effects of Using Only Voiced Speech. Using only voiced speech in the current speaker recognizer was suggested because of the common belief that voiced frames are more effective in characterizing a speaker than unvoiced frames. Furthermore, it is true that the all-pole, model-based LPC spectral distortion measure is more reliable and meaningful for voiced speech frames. However, the results indicate that the speaker identification error

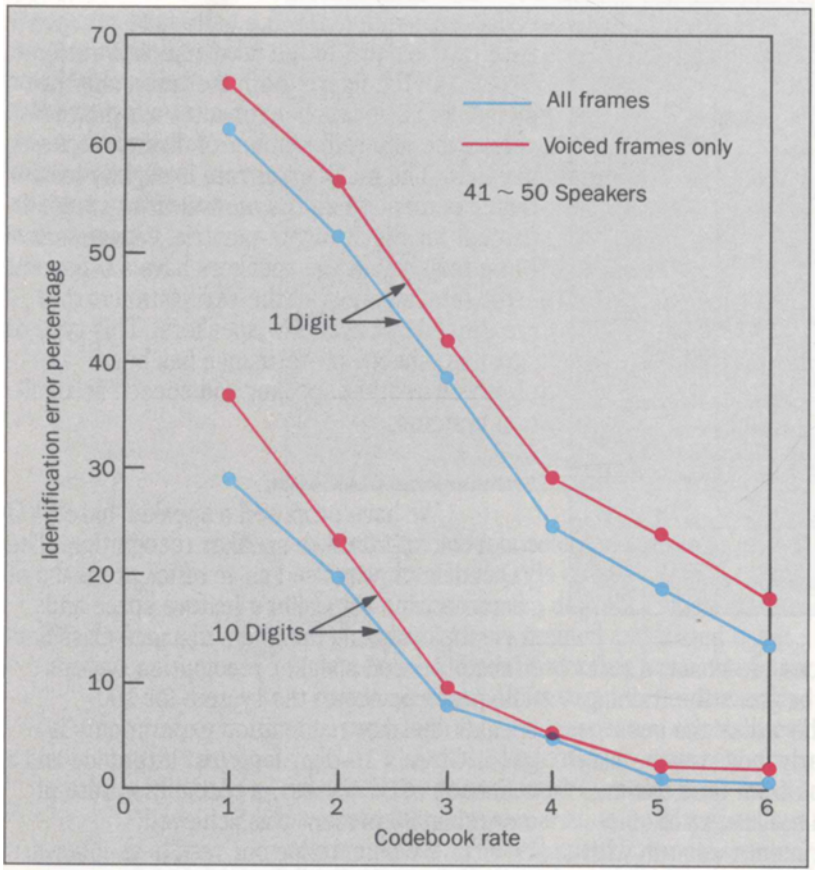

Figure 9. Identification error percentage versus codebook rate for 1- and 10-digit test utterances using either voiced frames or all frames.

rate obtained using only voiced frames is consistently worse than the error rate obtained using all speech frames for different codebook sizes and test utterance lengths (Figure 9). Although, for 10-digit test utterances the difference is smaller. The implication can be explained as follows.

In the nonparametric VQ codebook approach, all feature vectors (both voiced and unvoiced) are well represented. Because in the training phase we do not deliberately remove the unvoiced frames, we also use them in the recognition phase to achieve the same degree of characterization efficiency. The results clearly show that in this VQ approach it is not wise to discard an unvoiced speech segment because it carries less information. Thus, this VQ approach to speaker recognition does not require separating voiced and unvoiced frames. A straightforward VQ design is sufficient.

Effects of Different Recording Sessions. The identification error rate plotted as a function of the recording session number is shown in Figure 10. The codebook was generated from the first 100 utterances, or equivalently, utterances recorded in the 1st, the 2nd, and the first half of the 3rd session. The remaining 100 utterances were grouped according to their corresponding recording session numbers to form three different test sets. The first test set gave a significantly better result than the other two sets. This can be explained as follows.

The 40 utterances in the 3rd recording session formed a rather homogeneous data set. Because the first 20 digits of this homogeneous set were used with the 80 isolated utterances recorded in the 1st and 2nd session as the VQ codebook training set, the remaining 20 digits should be well represented by the resultant codebook. In the recognition phase, when the remaining 3rd session utterances were used, the effects of intraspeaker variation and channel differences were negligibly small and the distortion (quantization error) was as low as the training data.

On the other hand, the digit utterances recorded in the 4th and the 5th sessions were less correlated with the training utterances than the 3rd session, and possible channel difference and the intraspeaker variations could and did degrade the recognizer performance. Because the performance of the 5th recording utterances was even worse than

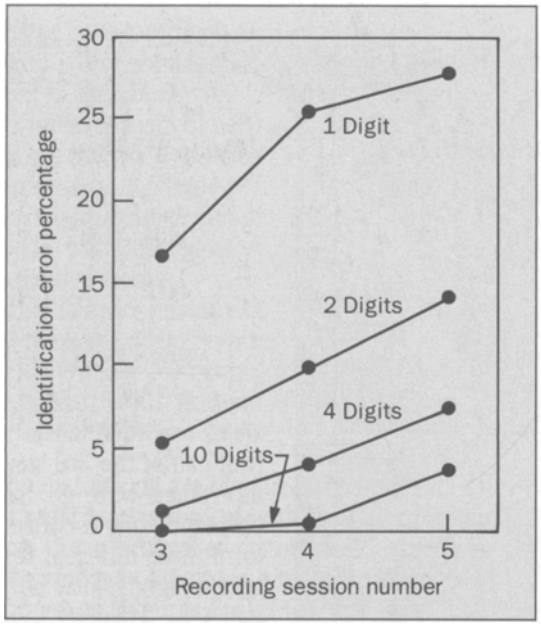

Figure 10. Identification error percentage versus recording session number.

the 4th recording utterances, we observe that the longer the separation between the training and the test recordings, the worse the performance. This result clearly indicates the need to update the VQ codebook from time to time.

Speaker Verification Results. A set of speaker verification experiments was run with the same speaker-based VQ framework. For each speaker, the test set consisted of his or her own 100 isolated-digit utterances plus 100 isolated-digit utterances of every other speaker (impostor). Speaker-based VQ codebooks of 64 entries were used. The verification threshold was computed *a posteriori*. That is, the cross-over point of the true speaker's distance probability distribution and the pooled impostors' distance probability distribution was used as the rejection/acceptance threshold. Thus, the verification curve shown in Figure 11 is an equal-error-rate curve (i.e., the false acceptance rate is equal to the false rejection rate).

In the figure, both the mean and the median of verification error rates are plotted versus the different number of distinct digits per trial. The mean error rate is slightly less than 2 percent while the median error rate is 0 percent for the 10-digits-per-trial experiment. More than half of the speakers have a 0-percent error rate and most of the verification errors are caused by only a few speakers. This type of "goat vs. sheep" performance has been observed in other speaker and speech recognition systems.

Summary and Discussion

We have proposed a speaker-based VQ codebook approach to speaker recognition. The VQ codebook was used as an efficient means of characterizing a speaker's feature space and was employed as a minimum distance classifier in the proposed speaker recognition system. The performance of the system for 100-speaker speaker recognition experiments is good. Given a 10-digit-long test utterance and a codebook of 64 vectors, a recognition rate of more than 98 percent was achieved.

We summarize our results as follows:

- Both larger codebook size and longer test token length (more digits in the test utterance) can improve recognition performance. The digit "9" when used as a test token outperformed any other digit.
- Using statistically less correlated information in a test utterance (e.g., using different instead of repeated digits in a test utterance) improves the speaker recognition performance.
- Performance improves when all speech frames are used for both training and recog-

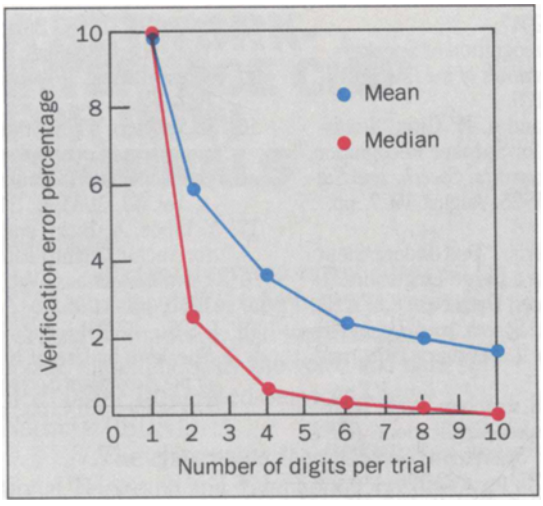

Figure 11. Mean and median of equal error rate (false rejection/ false acceptance) versus the number of digits per trial in speaker verification experiments.

nition than when a selected subset (e.g., only voiced frames) is used.

- System performance degrades when speaker VQ codebooks are trained and tested using data from different recording sessions. Some codebook adaptations seem to be necessary to cope with these intraspeaker variabilities.

The most distinctive feature of the proposed speaker-based VQ model is its multiple representation or partitioning of a speaker's spectral space. The VQ speaker model, while allowing some amount of overlap between different speaker's codebooks, is quite capable of discriminating impostors from a true speaker because of this distinctive feature. Furthermore, the finer acoustic and phonetic representations provided by the feature space partitioning have also made the proposed system more suitable for those text-independent speaker recognition applications where only a relatively short test utterance is available.

The proposed approach, although tested only on a highly phonetically constrained database (i.e., digits), is applicable to a truly text-independent speaker recognition environment. Of course, in a truly text-independent speaker recognition application, the training speech material has to be phonologically and linguistically much richer than the isolated-digit database used in this evaluation. Also, a VQ codebook with more than 64 entries may be necessary to characterize the acoustical and phonological details of truly unconstrained speech input.

The performance of the proposed text-independent speaker recognition system, although very good, is inevitably inferior to that of a text-dependent one because it lacks any temporal (dynamic) structure of the short-time spectrum. To compensate for this and to further improve system performance, the system can be easily modified to incorporate the temporal information when used in a text-dependent mode. In such a modification the specified reference utterance is first vector quantized and prestored as a sequence of VQ codebook indices. Then in the recognition phase, the test utterance is dynamically time aligned and compared with the prestored reference and a recognition decision can be made based on the accumulated distance along the optimally aligned path.

References

1. S. Pruzansky, "Pattern-Matching Procedure for Automatic Talker Recognition," *Journal of the Acoustical Society of America*, Vol. 35, March 1963, pp. 354-358.
2. K. P. Li, J. E. Dammann, and W. D. Chapman, "Experimental Studies in Speaker Verification Using an Adaptive System," *Journal of the Acoustical Society of America*, Vol. 40, 1966, pp. 966-978.
3. S. Pruzansky and M. V. Mathews, "Talker-Recognition Procedure Based on Analysis of Variance," *Journal of the Acoustical Society of America,*

Vol. 36, 1964, pp. 2041-2047.
4. B. S. Atal, "Automatic Recognition of Speakers from their Voices," *Proceedings of the IEEE*, Vol. 64, April 1976, pp. 460-475.
5. J. D. Markel, B. Oshika, and A. H. Gray, "Long-Term Feature Averaging for Speaker Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-25, August 1977, pp. 330-337.
6. J. D. Markel and S. B. Davis, "Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-spaced Data Base," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-27, No. 1, February 1979, pp. 74-82.
7. A. E. Rosenberg and M. R. Sambur, "New Techniques for Automatic Speaker-Verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-23, 1975, pp. 169-176.
8. S. Furui, "Cepstrum Analysis Technique for Automatic Speaker Verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-29, No. 2, April 1981, pp. 254-272.
9. K. P. Li and E. H. Wrench, "An Approach to Text-Independent Speaker Recognition with Short Utterances," *Proceedings, ICASSP*, 1983, pp. 555-558.
10. R. Schwartz, S. Roucos, and M. Berouti, "The Application of Probability Density Estimation to Text-Independent Speaker Identification," *Proceedings, ICASSP*, 1982, pp. 1649-1652.
11. K. C. Pan, F. K. Soong, and L. R. Rabiner, "A Vector Quantization Based Preprocessor for Speaker-Independent Isolated Word Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-33, No. 3, June 1985, pp. 546-559.
12. J. E. Shore and D. K. Burton, "Discrete Utterance Speech Recognition Without Time Alignment," *IEEE Transactions on Information Theory*, Vol. IT-24, No. 4, July 1983, pp. 473-491.
13. R. E. Helms, "Speaker Recognition Using Linear Prediction Vector Codebooks," Ph.D. thesis, Southern Methodist University, 1981.
14. E. Dorsey and J. Berstein, "Inter-Speaker Comparison of LPC Acoustic Space Using a Minimax Distortion Measure," *Proceedings, ICASSP*, 1981, pp. 16-19.
15. J. T. Buck, D. K. Burton, and J. E. Shore, "Text-Dependent Speaker Recognition Using Vector Quantization," *Proceedings, ICASSP*, 1985, pp. 381-384.
16. K. Shikano, "Text-Independent Speaker Recognition Using Codebooks in Vector Quantization," *Journal of the Acoustical Society of America*, Suppl. 1, Vol. 77, 1985, p. S11.
17. Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantization," *IEEE Transactions on Communications*, Vol. COM-28, No. 1, January 1980, pp. 84-95.
18. L-S Su, K. P. Li, and K. S. Fu, "Identification of Speakers by Use of Nasal Coarticulation," *Journal of the Acoustical Society of America*, Vol. 56, December 1974, pp. 1876-1882.

Biographies (continued)

Instituteof Technology and a Ph.D. in electrical engineering from the University of Pennsylvania. Mr. Juang joined AT&T in 1982 and does research on speech recognition, speech coding and enhancement, and stochastic processes. He has a B.S. in electrical engineering from National Taiwan University, and M.S. and Ph.D. degrees in electrical and computer engineering from the University of California at Santa Barbara. Mr. Rabiner joined AT&T in 1962 and is responsible for speech and speaker recognition research. He holds B.S., M.S., and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology.

(Manuscript received January 10, 1986)