A Vector Quantization Approach to Speaker Recognition

F. K. Soong
A. E. Rosenberg
L. R. Rabiner
B. H. Juang

AT&T Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT. In this study a vector quantization (VQ) codebook was used as an efficient means of characterizing the short-time spectral features of a speaker. A set of such codebooks were then used to recognize the identity of an unknown speaker from his/her unlabelled spoken utterances based on a minimum distance (distortion) classification rule. A series of speaker recognition experiments was performed using a 100-talker (50 male and 50 female) telephone recording database consisting of isolated digit utterances. For ten random but different isolated digits, over 98% speaker identification accuracy was achieved. The effects, on performance, of different system parameters such as codebook sizes, the number of test digits, phonetic richness of the text, and difference in recording sessions were also studied in detail.

I. Introduction

Automatic speaker recognition has long been an interesting and challenging problem to speech researchers [1-10]. The problem, depending upon 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 "none of the above" (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 (text-constrained) or text-independent (text-free). In the text-dependent mode the speaker is asked to utter a prescribed text. The utterance is then compared with some prestored pattern(s) of the same text. In the text-independent mode the speaker can, in theory, speak any speech material with no constraint. The unconstrained speech input is compared with model(s) which characterizes the speaker's features. In general, due to higher acoustic-phonetic variability of the text-independent input, longer training material is required to characterize (model) a speaker more reliably than in the text dependent mode. On the other hand, a fairly short, usually sentence long, utterance is adequate for text-dependent speaker recognition. In general, recognition procedures 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 non-linear time alignment between the utterances. There is no possibility of time alignment in true text-independent systems. Talker 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 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. Here, instead of using word-based VQ codebooks to characterize the phonetic contents of isolated words, we propose to use speaker-based VQ codebooks to characterize the variability of short-time acoustic features of speakers.

II. Speaker-based VQ Codebook Approach to Speaker Characterization and Recognition

A set of short-time raw feature vectors of a speaker can be used directly to represent the essential acoustical, phonological or physiological characteristics of that speaker 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 had to be found. In order to compress the original data to a small set of representative points, we used a VQ codebook with a small number of codebook entries.

The speaker-based VQ codebook generation can be summarized 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 $S = S_1 \cup S_2 \cup \cdots \cup S_M$. Each partition, S_i, forms a convex, 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

$$D = \frac{1}{I}\sum_{i=1}^{I}\min_{1\leq j\leq M} d(\mathbf{a}_i,\mathbf{b}_j) \qquad (3)$$

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

In this study we use short-time LPC vectors as feature vectors. The corresponding distortion measure to measure the similarity between any two feature vectors is the LPC likelihood ratio distortion measure. The likelihood ratio distortion between two LPC vectors $\mathbf{a}$ and $\mathbf{b}$ is defined as

$$d_{LR}(\mathbf{a},\mathbf{b}) = \frac{\mathbf{b}^T\mathbf{R}_a\mathbf{b}}{\mathbf{a}^T\mathbf{R}_a\mathbf{a}} - 1 \qquad (4)$$

where $\mathbf{R}_a$ is the autocorrelation matrix of speech input data associated with the vector $\mathbf{a}$. 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 depicted in Fig. 1. 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 (encoded) using the N codebooks corresponding to the N different speakers. The quantization

CH2118-8/85/0000-0387 $1.00 © 1985 IEEE

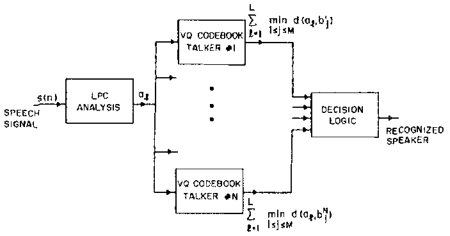

Fig. 1 Block diagram of the speaker-based VQ speaker recognition system.

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

$$D^i = \frac{1}{L} \sum_{\ell=1}^{L} \min_{1\leq j\leq M} d(\mathbf{a}_\ell, \mathbf{b}_j^i) \tag{5}$$

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

$$i^* = \underset{1\leq i\leq N}{\operatorname{argmin}} D^i \tag{6}$$

A speaker-verification system has a similar structure except 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 talker.

III. Data Base, LPC Analysis and Experiments

The proposed speaker-based VQ approach to speaker characterization is applicable to both text-dependent and text-independent speaker recognition. Since collecting a truly linguistically unconstrained database of many speakers is not a trivial task, we decided to use a much more constrained database as a first step to test the idea. Over a period of 2 months, a 100-talker (50 male and 50 female), digit vocabulary database was collected. Each of 100 talkers spoke 200 isolated digits (20 utterances per digit) over standard, local, dialed-up telephone lines in 5 different recording sessions. In each recording session the talker was asked to speak 4 sets of digit strings. Each string consisted of 10 randomly ordered, different isolated digits. The 200 isolated digits were split into two parts. The first one hundred digits were used for training (codebook generation) while the second one hundred digits were used for testing (recognition). The recognition experiments were performed on a digit-independent basis. In other words the speaker was allowed to say any random, isolated digit strings.

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 ms Hamming window was used to window the preemphasized speech data and an 8th order autocorrelation analysis was performed. The resultant 9 autocorrelation coefficients were then used to find the LPC feature vector which 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 recognition performance. In particular we varied:

(1) The number of digits in the test utterance.
Test tokens of one, two, four and ten random but different digits were used in the 100 talker identification experiments.

(2) The size of the codebook.
Six codebooks with 2, 4, 8, 16, 32 and 64 codebook vectors were used.

(3) The digits used in the test utterances.
Sets of 10 repeated digits were used as test tokens. The results were compared with the results obtained by using 10 random but different digits as test tokens.

(4) The frames used in the recognizer.
Voiced frames were extracted from the test utterance by using a simple but effective voiced/unvoiced detector. The voiced/unvoiced decision was made based on the frame energy and prediction gain. The results were compared with the results obtained when all speech frames were used.

(5) 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 intra-speaker variations and possible channel variations of different recording sessions affected the recognizer performance.

IV. Results and Discussion

4.1 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 Fig. 2. The codebook size is represented by the

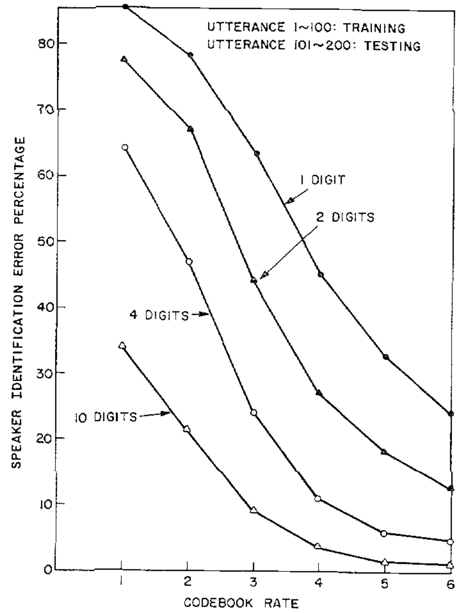

Fig. 2 Identification error percentage versus codebook rate for test sequences of 1, 2, 4, and 10 digits.

number of vector entries, M or by the corresponding rate, $M = 2^R$. Four curves corresponding to 4 different test token lengths (i.e., 1 digit, 2 digits, 4 digits 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 ten isolated digits, the error rate dropped from 34% to 1.5% when the codebook size was increased from 2 to 64 vectors. With a codebook of 64 vectors the error rate dropped from 24% to 1.5% when the test token length was increased from 1 to 10 digits.

To further illustrate the effects of codebook size on the recognition error rate Fig. 3 shows the averages, standard deviations and histograms, respectively, of inter-speaker and intra-speaker distortions of the 100 speaker identifications. Results using two different codebook (i.e. 8 vectors (rate 3) and 32 vectors (rate 5)) are given in this figure. The average distortions of both codebooks are fairly

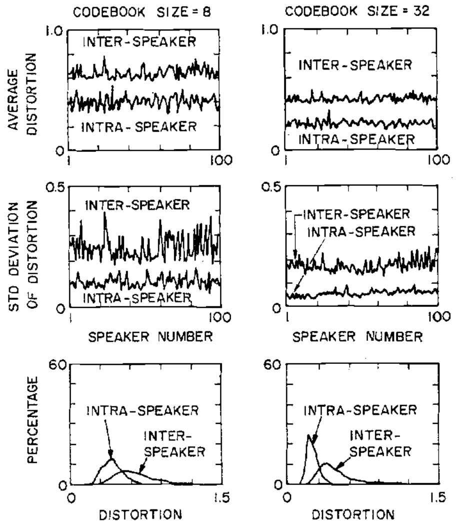

Fig. 3 The effects of different codebook sizes.

constant across all 100 talkers and only a small amount of random fluctuations are observed. It is seen that the separations between the inter- and the intra-speaker average distortions are roughly the same for the two different size codebooks, or, equivalently when the codebook sizes increase, both the inter- and intra-speaker average distortions decrease by more or less the same amount. Also we observe that the standard deviations of both the inter- and intra-speaker distortions are reduced when the codebook size is increased from 8 to 32. The inter- and the intra-speaker distortion histograms clearly show that when the separation between the inter- and intra-speaker average distortion are held relatively fixed, reduction of the standard deviation decreases the overlapping area between the two histograms. As a result of this reduction of overlapping area, the recognition error rate is reduced.

In Fig. 4 the averages, standard deviations and histograms of inter- and intra-speaker distortions of three different token lengths (1 digit, 4 digit and 10 digits) are shown. A codebook of 32 vectors was used for all cases. The average distortions are identical for all three token lengths, but the standard deviations of the distortions are significantly different. The intra-speaker standard deviation decreases much faster than the inter-speaker standard deviation as the number of digits in the test increases. The overlapping region between the two histograms in Fig. 4(c) is smaller for longer test tokens.

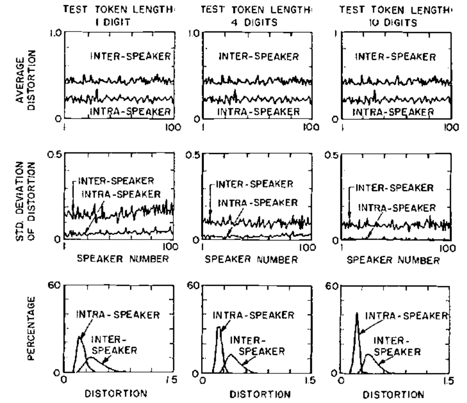

Fig. 4 The effects of different token lengths.

4.2 Effects of Repeated Digits and Different Digits

The speaker identification results obtained using 10 repeated digits are shown in Fig. 5. The digit "9" achieved the best results while the digits "3", "4" and "6", provided the worst scores. The good result obtained for digit "9" may be explained by the fact that it has a relatively long duration and it has a strong nasal-vowel coarticulation. The nasal-vowel coarticulation has long been known to be effective for speaker recognition purposes [14]. Since it is a relatively rapid event, the nasal-vowel coarticulation is unlikely to be modified consciously by the speaker. Therefore it serves as a reliable talker identification feature with inherently low intra-speaker variability.

The digits "6" and "8", each having a stop gap with silence frames, do not achieve good recognition scores. This should not be 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 large variance. As a consequence the average distortion of the true talker is corrupted by a random distortion making a misrecognition more likely.

It is interesting to note that the performance of 10 different digits, shown as the dashed line in Fig. 5, is better than the performance of any 10 repeated digits. This result indicates that statistically uncorrelated information could, and did, improve the speaker recognition performance.

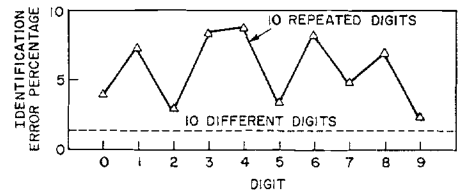

Fig. 5 Identification error percentage versus digit for repeated digits and different digits.

4.3 Effects of Using Only Voiced Speech

The results obtained using only frames of voiced speech are presented along with the results obtained using all speech frames in Fig. 6. In this figure, the speaker identification error rate obtained using only voiced frames is consistently worse than the error rate obtained using

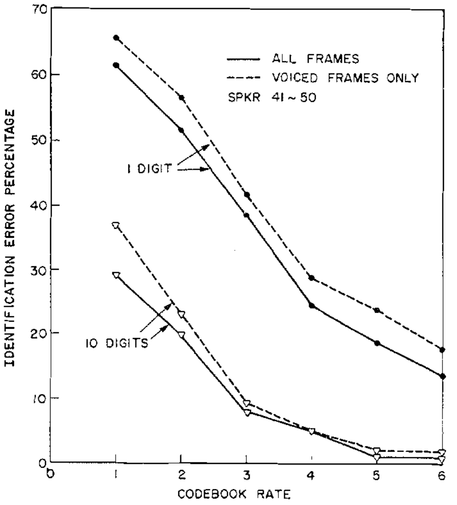

Fig. 6 Identification error percentage versus codebook rate for 1 and 10 digit test utterances using either voiced frames or all frames.

all speech frames for different codebook sizes and test token lengths, although for 10 digit test utterances the difference is small. These results are somewhat surprising because it is commonly believed that the voiced frames are more effective than the unvoiced frames in characterizing the acoustical features of a speaker. Also, the likelihood ratio distortion used in our experiments is most effective in measuring the spectral distortion (similarity) between voiced speech frames. The results can be explained as follows. In the nonparametric VQ codebook approach, all of the feature vectors (both voiced and unvoiced) are well represented. Since, in the training phase, we do not deliberately remove the unvoiced frames, we should use them in the recognition phase, too. In this new nonparametric VQ approach we not only eliminate the need to separate voiced frames from the input data, but also we improve the speaker recognition performance by using all the speech data.

4.4 Effects of Different Recording Sessions

The identification error rate plotted as a function of the recording session number is shown in Fig. 7. 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 forty utterances in the third recording session formed a rather homogeneous data set. Since the first twenty digits of this homogeneous set were used as part of the VQ codebook training set, the remaining twenty digits should be well represented by the resultant codebook. In the recognition phase when the remaining 3rd session utterances were used, the effects of intra-speaker 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 the channel difference and the intra-speaker variations degraded the recognizer performance. Since the performance of the 5th recording utterances was even worse than 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.

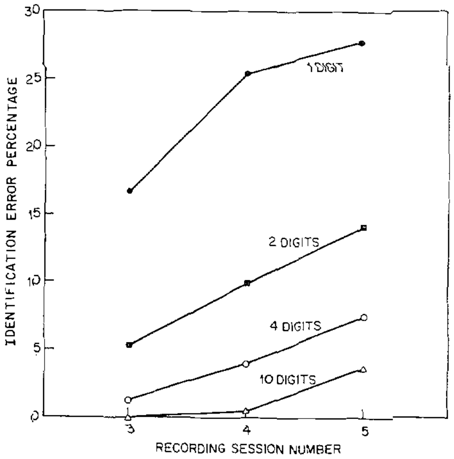

Fig. 7 Identification error percentage versus recording session number.

V. Summary and Discussion

We have proposed a speaker-based VQ codebook approach to speaker recognition application. The VQ codebook was used as an efficient means to characterize a speaker's feature space and was employed as a minimum distance classifier in the proposed speaker recognition system. The results of the 100 talker speaker recognition experiments are good; given a 10 digit long test token and a codebook of 64 vectors a recognition rate of over 98% was achieved. We summarize our results as follows:

(1) Both larger codebook size and longer test token length (more digits in the test utterance) can be used to improve the recognition performance. Ten different digits, when used as test tokens, outperformed any of the repeated digit test tokens. The digit "9" when used as a test token outperformed other digits.

(2) Phonetically rich test tokens give better performance than phonetically poor test tokens.

(3) It is recommended that all speech frames be used for both training and recognition in our approach.

(4) It is recommended that the VQ codebook be updated from time to time to alleviate the performance degradation due to different recording conditions and intra-speaker variations.

The proposed approach, although tested only on a highly linguistically constrained database (i.e., digits), is believed to be extensible to truly text-independent speaker recognition environment.

REFERENCES

[1] Pruzansky, S., "Pattern-Matching Procedure for Automatic Talker Recognition," *JASA*, Vol. 35, pp. 354-358, March 1963.

[2] Li, K. P., Dammann, J. E. and Chapman, W. D., "Experimental Studies in Speaker Verification Using an Adaptive System," *JASA*, Vol. 40, pp. 966-978, 1966.

[3] Pruzansky, S. and Mathews, M. V., "Talker-Recognition Procedure Based on Analysis of Variance," *JASA*, Vol. 36, pp. 2041-2047, 1964.

[4] Atal, B., "Automatic Recognition of Speakers from their Voices," *Proc. IEEE*, Vol. 64, pp. 460-475, April 1976.

[5] Markel, J. D., Oshika, B. and Gray, A. H., "Long-Term Feature Averaging for Speaker Recognition," *IEEE Trans. Acoust., Speech and Signal Processing*, Vol. ASSP-25, pp. 330-337, August 1977.

[6] Markel, J. D. and Davis, S. B., "Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-spaced Data Base," *IEEE Trans. Acoust., Speech and Signal Processing*, ASSP-27, No. 1, pp. 74-82, February 1979.

[7] Rosenberg, A. E. and Sambur, M. R., "New Techniques for Automatic Speaker-Verification," *IEEE Trans. Acoust., Speech and Signal Processing*, Vol. ASSP-23, pp. 169-176, 1975.

[8] Furui, S., "Cepstrum Analysis Technique for Automatic Speaker Verification," *IEEE Trans. on Acoust., Speech and Signal Processing*, ASSP-29, No. 2, pp. 254-272, April 1981.

[9] Li, K. P. and Wrench, E. H., "An Approach to Text-Independent Speaker Recognition with Short Utterances," *Proc. ICASSP*, pp. 555-558, 1983.

[10] Schwartz, R., Roucos, R. and Berouti, M., "The Application of Probability Density Estimation to Text-Independent Speaker Identification," *Proc. ICASSP*, pp. 1649-1652, 1982.

[11] Pan, K. C., Soong, F. K. and Rabiner, L. R., "A Vector Quantization Based Preprocessor for Speaker-Independent Isolated Word Recognition," submitted for publication.

[12] Shore, J. E. and Burton, D. K., "Discrete Utterance Speech Recognition Without Time Alignment," *IEEE Trans. on Inform. Theory*, Vol. IT-24, No. 4, pp. 473-491, July 1983.

[13] Linde, Y., Buzo, A. and Gray, R. M., "An Algorithm for Vector Quantization," *IEEE Trans. on Communications*, Vol. COM-28, No. 1, pp. 84-95, January 1980.

[14] Su, L-S, Li, K. P. and Fu, K. S., "Identification of Speakers by Use of Nasal Coarticulation," *JASA*, Vol. 56, pp. 1876-1882, December 1974.