Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Text Dependent Modes

A.E. Rosenberg
F.K. Soong

Speech Research Department
AT&T Bell Laboratories
Murray Hill, NJ 07974 USA

ABSTRACT. A vector quantization based talker recognition system is described and evaluated. The system is based on constructing highly efficient short-term spectral representations of individual talkers using vector quantization codebook construction techniques. Although the approach is intrinsically text-independent, the system can be easily extended to text-dependent operation for improved performance and security by encoding specified training word utterances to form word prototypes. The system has been evaluated using a 100-talker database of 20,000 spoken digits. In a talker verification mode, average equal-error rate performance of 2.2% for text-independent operation and 0.3% for text-dependent operation is obtained for 7-digit long test utterances.

1. Introduction

Vector quantization has provided an effective means for compressing the short term spectral representation of speech signals. (For a review of the techniques, see [1].) It has also been successfully applied to speech recognition, for example, by representing each word in a vocabulary by one or more spectral vector codebooks [3,4]. VQ can be similarly applied to talker recognition. In our basic VQ-based talker recognition technique [2], each talker is characterized by a VQ codebook constructed from a large set of short term spectral vectors obtained from a series of training utterances provided by the talker. If the set of training utterances is large and rich enough to adequately represent the talker's utterance repertoire, and, in turn, the talker model codebook is constructed to adequately represent the short term spectra of these utterances, the characterization can be said to be text independent. In addition to our previous study, this kind of text independent talker recognition has also been studied by Helms [5] and Shikano [6]. Related work has been reported by Li and Wrench [7], Dorsey and Bernstein [8], and Buck et al [9].

The work reported in this paper modifies and extends the previously reported work [2] in the following ways. The analysis has been modified to provide linear predictive coding (LPC) based cepstral coefficients together with a weighted Euclidean distance as the distortion measure. The basic text independent system has been extended to provide a text dependent operation option which requires an additional training procedure to obtain encoded word prototypes.

2. VQ-based talker recognition

The basic principle underlying the VQ-based talker recognition technique is the compression of a large set of short-term spectral vectors representing the spoken utterances of a talker into a small set of such vectors which, according to a specified criterion, provide a faithful and efficient representation of the original set of vectors. The compressed set of vectors is referred to as a "codebook". We use the minimum average distortion technique [10] to generate our talker model codebooks.

Suppose we are given N training vectors $r_1,r_2,...,r_N$. The space of these vectors is partitioned into Q disjoint sets of vectors whose centroids are denoted by $c_1,c_2,...,c_Q$. These are referred to as the codebook vectors. The partitioning is carried out such that the average distance of each training vector to its best matching, nearest neighbor, codebook vector is minimized. That is, the set of codebook vectors $c_1,c_2,...,c_Q$ is obtained such that

$$D = \frac{1}{N}\sum_{i=1}^{N} \min_{1\leqslant q\leqslant Q} d(r_i,c_q)$$

is minimized, where d is the distance between vectors in the space, and $\min_{1\leqslant q\leqslant Q} d(r_i,c_q)$ is the codebook distortion associated with r_i. Other criteria are possible for partitioning the the training vector space. In the so-called minmax or covering algorithm used by Helms [5] and Dorsey and Bernstein [8], the maximum distance of any training vector to its best matching codebook vector is minimized over all training vectors. This procedure produces a codebook whose vectors "cover" the entire region spanned by the training vectors. Codebook vectors tend to be evenly spaced. In contrast, the minimum average technique produces codebook vectors which tend to represent the density and clustering of the input training vectors.

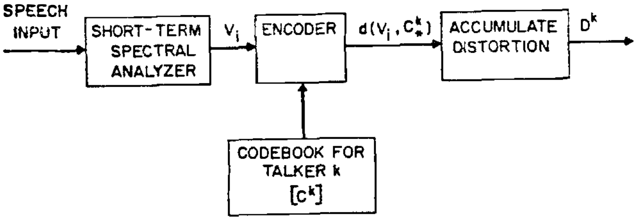

Fig. 1. Basic, text-independent operation of the VQ-based talker recognizer.

In operation, as shown in Figure 1, a set of test vectors, $v_1,v_2,...,v_M$, analyzed from the utterance of an unknown talker is encoded using the codebook, $\{c^k\}$, of a specified talker k. The distortion for each test vector is given by

$$d(v_i,c_*^k) = \min_{i\leqslant q\leqslant Q} d(v_i,c_q^k).$$

where c_*^k is the best matching codebook vector. The distortion accumulated over the test vectors and normalized by the number of test vectors, $D^k = \frac{1}{M}\sum_{i=1}^{M} d(v_i,c_*^k)$, is used to make a recognition decision.

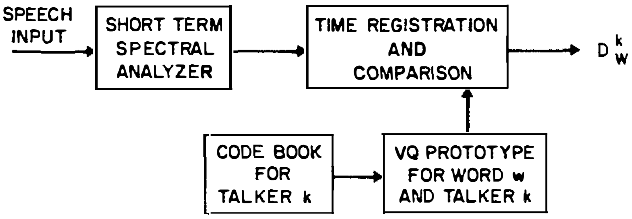
SPEECH INPUT
SHORT TERM SPECTRAL ANALYZER
TIME REGISTRATION AND COMPARISON
D_W^k
CODE BOOK FOR TALKER k
VQ PROTOTYPE FOR WORD w AND TALKER k

Fig. 2. Text-dependent operation of the VQ-based talker recognizer.

CH2243-4/86/0000-0873 $1.00 © 1986 IEEE

The text-independent system can be extended to text-dependent operation in the following way. After the talker model codebook has been generated for a particular talker, the talker is prompted to provide one or more utterances of each word from a specified vocabulary. These utterances are analyzed and encoded using the talker's codebook and stored as word prototypes in the form of a sequence of codebook vector indices for each prototype. In operation, an unknown talker is prompted to provide an utterance of a particular word w from the vocabulary. As shown in Figure 2, the utterance is analyzed to generate a sequence of test vectors. This sequence is time registered and compared with the sequence of reference word vectors associated with talker k for word w using standard dynamic programming time warping (DTW) techniques [11]. The overall distance between the input word and the prototype word is obtained by accumulating individual test-vector-to-reference-vector distances along the best warping path.

3. Talker Recognition Modes

Two talker recognition modes are considered in this study, talker identification and talker verification. In talker identification the test input for an unknown talker is compared with references for every talker in a given population. The reference talker associated with the smallest overall distance for the test input is taken to be the identified talker. In our evaluation test utterances from each talker are compared with references for every talker in the population. The number of incorrect identifications are tabulated. Identification performance is specified as the average rate of incorrect identifications over all test utterances and test talkers.

In talker verification a test utterance from an unknown talker is compared with the reference for the talker whose identity is claimed. If the overall distance for the test input is less than a specified threshold the identity claim is accepted; otherwise it is rejected. In our evaluation test utterances from every talker are compared with each reference. Threshold are not explicitly assigned. Instead, for each reference talker, the distance for which the rate of reference talker rejection ("customer" rejections) is the same as the rate of acceptance of all other talkers ("impostor" acceptances) is calculated. The associated rate is known as the verification equal-error rate. It is averaged over all test utterances and reference talkers to estimate performance in the verification mode.

4. Data Base, Analysis, and Experimental Setup

The data base used for evaluation of this system consists of 20,000 isolated digit utterances obtained from 100 talkers, 50 male and 50 female. The utterances were recorded over dialed-up local telephone lines using an ordinary telephone handset with the talkers seated in a sound booth. Each talker provided 20 repetitions of each digit in 5 recording sessions held over a period of up to two months. In each recording session the talkers were prompted to utter 4 series of the 10 digits, with the order of the digits randomized in each series.

The analog speech utterances, bandpass filtered from 200 to 3200 Hz, are sampled at a 6.67 kHz rate. The speech samples are preemphasized and an eighth order autocorrelation analysis is carried out over 45 msec Hamming windows every 15 msecs. Each frame of autocorrelation coefficients is converted into an eighth order LPC vector. Finally, two kinds of eighth order cepstral vectors are calculated from the LPC vectors: "static" cepstral vectors obtained directly from each LPC vector, and "dynamic" cepstral vectors calculated from a best linear fit to the first order change of the "static" vectors over 7 successive frames.

In the evaluation, the first eighty utterances (eight repetitions of each digit) for each talker are set aside for training. The remaining 120 utterances are used for testing. 80 digit utterances is equivalent to approximately 2900 frames or 43.5 secs. of speech.

Talker model codebooks, containing 16, 32 or 64 vectors (alternately referred to as 4, 5, or 6 bit codebooks), are constructed from 20, 40, or all 80 of the training utterances. Separate codebooks are constructed for static and dynamic cepstral vectors. The distortion measurement used to construct the codebooks is a weighted Euclidean distance between vectors represented as

$$d = \sum_{j=1}^{8} w_j^2 (r_j - r_j')^2$$

where r_j and r_j' are the j-th components of two cepstral vectors r and r'. w_j^2 is given by $1/s_j^2$, where s_j^2 is the variance of the j-th component of the training vectors averaged over all talkers and training utterances.

Word prototypes for text-dependent operation are constructed using the first 50 training utterances. After the construction of the talker model codebooks, word prototypes are obtained by encoding, frame by frame, these training utterances from each talker with the talker's codebooks. The word prototypes are stored as the sequence of the indices of the encoded vectors. Since a word prototype is obtained from each of these training utterances, the result is 5 word prototypes for each of the digits.

In text-independent operation, the static and dynamic cepstral vectors representing each frame are encoded with the static and dynamic codebooks for a specified talker. The resulting static and dynamic distortions are scaled by dividing each one by its average over all the training utterances. The total distortion is the sum of the scaled static and dynamic distortions.

In text-dependent operation, the distance calculation is similar, except that the distortion measurements are carried out with respect to codebook vectors specified by a word prototype instead of the best matching codebook vector. The overall distance for a designated digit is taken as the minimum distance over the number of prototypes available for that word.

Detailed descriptions of the cepstral representations and distance measurements are found in Soong and Rosenberg [12].

In the test evaluation, a trial consists of 1 to 10 digit utterances. Thus, the average trial length varies from 0.54 to 5.44 secs. The utterances for each trial are obtained in the order in which they were recorded. Therefore, each trial is a simulation of the utterance of a string of random digits. Results are tabulated separately for each trial length. Since there are a total of 120 test utterances, there are 120 single-digit trials, 60 2-digit trials, and so forth, up to 12 10-digit trials per talker.

5. Results and Discussion

All results are presented as error rates averaged over the test trials and talkers in the population for each test trial length. Although both simulated identification mode and verification mode experiments were carried out, to compress the presentation, except for the first set of results, only verification mode results are given in detail. Verification equal-error rate is (ideally) independent of talker population size and provides a readily interpretable measure of the separation between talkers.

5.1 Text-independent performance for different codebook sizes

Overall results for text-independent operation are shown in Fig. 3. Mean identification error rate performance over all 100 talkers is shown in (a), while mean verification equal-error rate performance is shown in (b). Three plots are shown in each case for each of the three codebook sizes investigated, 16, 32 and 64 vectors. The training conditions for each of these codebook sizes are respectively 20, 40, and 80 training utterances divided equally between the first two recording sessions. On average, therefore, there are approximately 45 training vectors for each codebook vector. It can be seen that performance increases as either the test trial length or codebook size increases. However, the rate of improvement drops off sharply as test trial lengths become greater than 5 or 6 digits. Also, the rate of improvement is smaller going from 32 to 64 vectors than from 16 to 32 vectors. For 64 vector codebooks, identification error rate improves from 21.7% for single-digit trials to 2.1% for 7-digit trials. The equal-error rate for verification is similarly reduced from 14.0% to 3.0%.

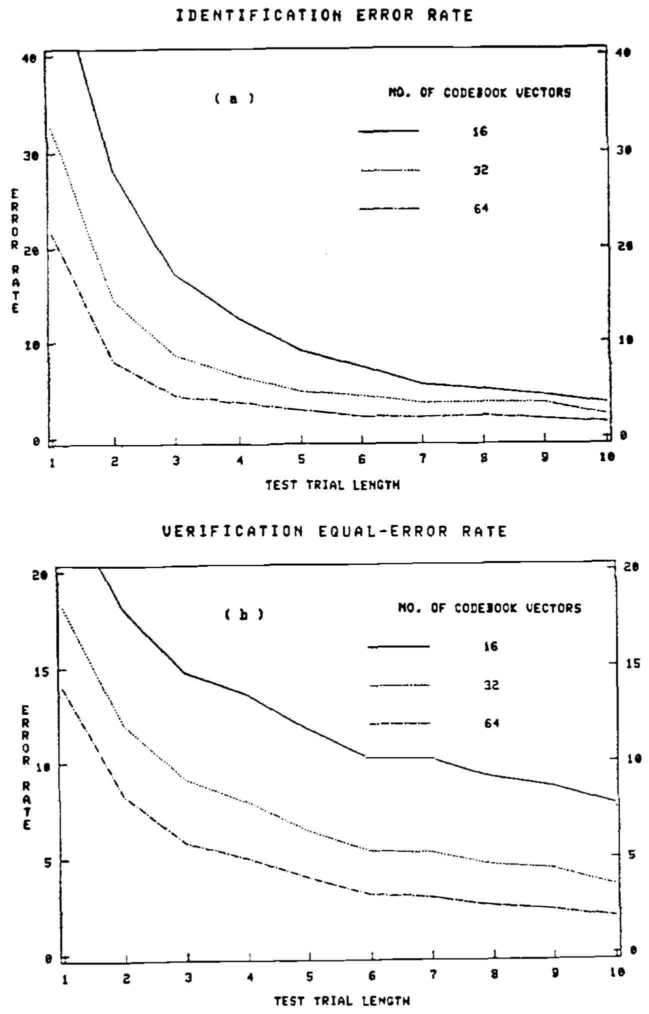

Fig. 3. Mean error rate over 100 talkers for text-independent operation as a function of test trial length for 3 codebook sizes; (a) shows identification error rate; (b) shows verification equal-error rate.

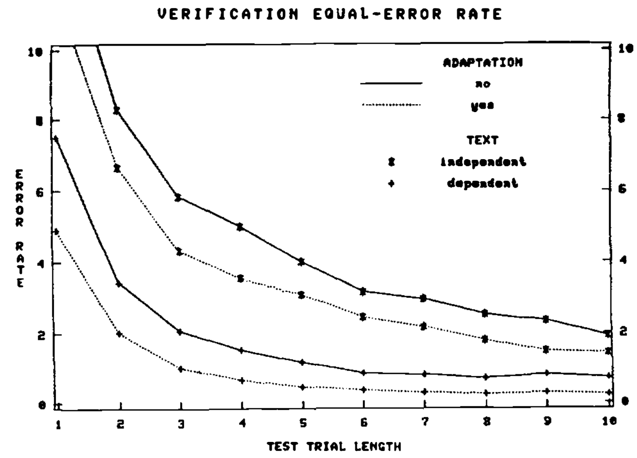

Fig. 4. Mean verification equal-error rate over 100 talkers as a function of test trial length for text-independent and text-dependent operation and without and with adaptation. The codebook size is 64 and there are 5 prototypes per word.

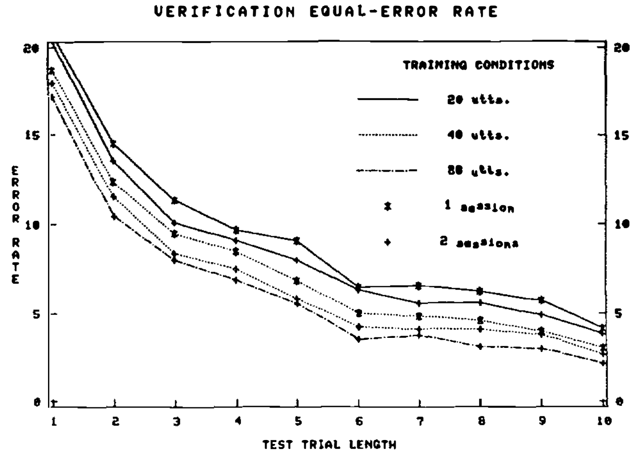

Fig. 5. Mean verification equal-error rate over 20 talkers for text-independent operation as a function of test trial length for various training conditions. The codebook size is 32 vectors.

5.2 Text-dependent operation

Verification performance in text-independent and text-dependent operation are compared in Fig. 4. In this example, 64-vector codebooks are used and there are 5 word prototypes per digit available for text-dependent operation. For the moment, consider only the results for which there is no adaptation. It can be seen that there is considerable overall improvement in performance for text-dependent operation which, however, diminishes somewhat for longer test trials. Equal-error rate is reduced from 7.5% for single-digit trials to 0.8% for 7-digit trials. It is generally expected that text-dependent performance should be consistently superior to text-independent performance for the same test input. In text-independent operation, the overall short-term spectral content of an utterance is compared with a short-term spectral model of a specified talker's utterances. By contrast, in text-dependent operation, the comparisons are substantially more constrained and precise. That is, the word prototype information constrains the sequence and subsets of codebook vectors which are compared.

5.3 Training condition variations

Some effects of varying training conditions on performance with text-independent operation are shown in Fig. 5. The experiments are carried out over a 20 talker (10 male, 10 female) subset of the population with approximately the same verification performance as that obtained over the whole population. The results are obtained using 32-vector codebooks. Five training conditions are examined. Codebooks are constructed using 20, 40, and 80 training utterances where, in each case, the training utterances are divided equally between the first two recording sessions. Also, for 20 and 40 training utterances, codebooks are also constructed from training utterances obtained from only the first recording session.

First, as expected, for the 2-session conditions, we observe that performance increases as the number of training utterances increases, although the improvement going from 20 to 40 utterances is substantially greater than from 40 to 80 utterances. For example, for 7-digit trial lengths, equal-error rate is 5.6% for 20 utterances, 4.1%

for 40 utterances, and 3.8% for 80 utterances. This result is consistent with a "rule of thumh" found in preliminary experiments. This is that codebook construction stabilizes with an average of 50 training vectors per codebook vector. 40 utterances is equivalent to approximately 1450 training vectors or 45 training vectors per codebook vector. Second, we observe a significant, though small, performance advantage for 2-session training over 1-session training. This observation is consistent with the results of other experimenters who have shown that recognition performance improves when the collection of training data is carried over more than one session.

5.4 Adaptation

Consistent with the notion of improving performance by collecting training utterances over a long period of time in order to account for natural variations in talker behavior is the notion of updating talker reference prototypes with current test inputs. By this means reference prototypes can be made to continue to reflect natural talker variations after their initial construction by combining them with current test inputs. For this evaluation we have experimented with two adaptation techniques, one for talker codebooks, the other for word prototypes.

For codebook adaptation we average current input vectors encoded by a particular codebook vector with that codebook vector to form an updated version of this vector. Suppose there are M_q input vectors $v_{qi}, i=1,2,....,M_q$, encoded by c_q. Then

$$c_{qnew} = (N_q c_q + \sum_{i=1}^{M_q} v_{qi}) / (N_q + M_q)$$

where N_q, the "cell occupancy", is the current total number of vectors, associated with c_q. The cell occupancy is updated as follows:

$$N_{qnew} = N_q + M_q - \overline{N}$$

where $\overline{N}$ is the average cell occupancy over the entire codebook. This procedure is carried out for each codebook vector c_q, $q=1,2,...,Q$.

For word prototype adaptation, we simply replace the worst matching prototype of a word by the encoded input of that word in the current trial.

In this evaluation, each talker's codebooks are adapted after every 10 test utterances (regardless of the number of test utterances per trial), while word prototypes are replaced after every test utterance. The effect of adaptation on verification performance is shown in Figure 4 for the 100 talker population using 6 bits codebooks and 5 word prototypes per digit. It can be seen that there is a significant and consistent decrease in error rate for both text-independent and text-dependent operation. For example, for single digit trials text-independent performance improves from 14.0% to 11.7%, (offscale in the figure), while performance for 7-digit trials improves from 3.0% to 2.2%. For text-dependent operation, the improvements are from 7.5% to 4.9% for single-digit trials, and from 0.8% to 0.3% for 7-digit trials.

It should be noted that these performance results for adaptation are idealized in the sense that reference data are adapted using current test inputs from the correct talker regardless of the outcome of the trials.

6. Summary and Conclusion

A VQ-based approach to automatic talker recognition has been presented and evaluated. The approach is based on constructing highly efficient and representative talker models using vector quantization codebook construction techniques from short-term spectral analysis of training utterances. The technique is intrinsically text-independent provided that the overall text content of test and training utterances is reasonably uniform. It differs from previously reported techniques for text-independent talker recognition in that it explicitly attempts to model representative short-term spectral characteristics whereas other techniques have characterized talkers in terms of long time average statistics of short-term spectra. The new technique, therefore, has the potential for making finer discriminations between talkers using shorter test utterances.

Our evaluation has shown that high performance talker recognition is attainable using highly compressed models of talkers, For example, with 6-bit (64-vector) codebooks and 7-digit test utterances (approximately 3.5 secs. long) a mean verification equal-error rate of 3.0% is obtained. A simple adaptation technique using current test inputs to modify talker model codebooks has been shown to help control intra-talker variability and reduce average equal-error rate to 2.2%.

An important feature of this talker recognition system is that it can be easily extended from text-independent to text-dependent operation. Specified training words or phrases from each talker are encoded using the talker's already established codebook. The storage requirements for the resulting prototypes are inherently modest since representing an encoded prototype using, for example, a 6-bit codebook requires only 6 bits per frame. For 6-bit codebooks with 5 word prototypes per digit and 7-digit test utterances the average equal-error rate is 0.8%. With adaptation, the average equal-error rate drops to 0.3%.

REFERENCES

[1] J. Makhoul, S. Roucos, and H. Gish, "Vector quantization in speech coding," *Proc. IEEE*, v. 73, pp. 1551-1588, 1985.

[2] F.K. Soong, A.E. Rosenberg, L.R. Rabiner, and B.H. Juang, "A vector quantization approach to speaker recognition," *Proc. ICASSP 85, IEEE Intl. Conf. on Acoust., Speech, and Signal Processing*, v. 1, pp. 387-390, 1985.

[3] J.E. Shore and D.K. Burton, "Discrete utterance speech recognition without time alignment," *IEEE Trans. on Inform. Theory*, v. IT-24, No. 4, pp. 473-491, 1983.

[4] K.C. Pan, F.K. Soong, and L.R. Rabiner, "A vector quantization based preprocessor for speaker-independent isolated word recognition," *IEEE Trans. on Acoust., Speech, and Signal Processing*, v. ASSP-33, pp. 546-560, 1985.

[5] R.E. Helms, "Speaker recognition using linear prediction vector codebooks," Ph.D. thesis, Southern Methodist University, 1981.

[6] K. Shikano, "Text-independent speaker recognition experiments using codebooks in vector quantization," (abstract), *J. Acoust. Soc. Am.*, Suppl. 1, v. 77, p. S11, 1985.

[7] K.P. Li and E.H. Wrench, Jr., "An approach to text-independent speaker recognition with short utterances," *Proc. ICASSP 83, IEEE Intl. Conf. on Acoust., Speech, and Signal Processing*, v. 2, pp. 555-558, 1983.

[8] E. Dorsey and J. Bernstein, "Inter-speaker comparison of LPC acoustic space using a minimax distortion measure," *Proc. ICASSP 81, IEEE Intl. Conf. on Acoust., Speech, and Signal Processing*, v. 1, pp. 16-19, 1981.

[9] J.T. Buck, D.K. Burton, and J.E. Shore, "Text-dependent speaker recognition using vector quantization," *Proc. ICASSP 85, IEEE Intl. Conf. on Acoust., Speech, and Signal Processing*, v. 1, pp. 391-394, 1985.

[10] Y. Linde, A. Buzo, and R.M. Gray, "An algorithm for vector quantization," *IEEE Trans. on Commun.*, v. COM-28, no. 1, pp. 84-95, 1980.

[11] L.R. Rabiner and S.E. Levinson, "Isolated and connected word recognition-theory and selected applications," *IEEE Trans. on Commun.*, v. COM-29, pp.621-659, 1981.

[12] F.K. Soong and A.E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *Proc. ICASSP 86, IEEE Intl. Conf. on Acoust., Speech, and Signal Processing*, pp. 17.5.1-17.5.4, 1986.