TEXT-DEPENDENT SPEAKER RECOGNITION USING VECTOR QUANTIZATION

Joseph T. Buck, David K. Burton, and John E. Shore

Computer Science and Systems Branch, Information Technology Division
Naval Research Laboratory, Washington, DC 20375

ABSTRACT

An application of source coding to speaker recognition is described. The method is text-dependent – the text spoken is known, and the problem is to determine who said it. Each speaker is represented by a sequence of vector quantization codebooks; known input utterances are classified using these codebook sequences and the resulting classification distortion is compared to a rejection threshold. On a 16 speaker test population with an additional 111 imposters, this method achieved a false rejection rate of 0.8%, an imposter acceptance rate of 1.8%, and within the 16 speakers, an identification error rate of 0.0%.

BACKGROUND

Vector quantization (VQ) is a data compression principle that permits low rate speech coding [1, 2]. A VQ encoder is designed by obtaining a long training sequence of speech, dividing the speech into frames, performing linear predictive analysis on the speech, and using a clustering algorithm on this data to obtain a *codebook* of representative spectra, or *code words*[3]. The size of the codebook is normally a power of 2 i.e. – $N=2^R$, and it is called a *rate* R codebook. The codebook is designed to minimize the average distortion that results from representing this training sequence. The distortion measure used in the work reported here is the gain-normalized Itakura-Saito distortion [4].

VQ has successfully been applied to speech recognition by using VQ to reduce the amount of data to be processed [5, 6, 7]. In addition, we developed a method of isolated word recognition (IWR) in which a separate codebook is designed to represent each isolated word to be recognized [8]. An unknown word is encoded using each codebook, and the distortion caused by the encoding is measured. The codebook that encodes the unknown word with lowest distortion determines the classification. We subsequently generalized the method to represent each word by a time-ordered sequence of codebooks [9, 10]. We call each of these codebooks a *section codebook*, each sequence a *multisection codebook* (or simply a *codebook*), and the approach *multisection*.

To use multisection VQ codebooks for IWR, each word in the vocabulary is represented by a sequence of vector quantization codebooks. Each codebook sequence is trained with several utterances of a single vocabulary word. First, each training word is length normalized to L (24 in this paper) frames and linear predictive analysis is performed. The word is then divided into sections, each section consisting of n frames (n must divide L). The first n frames of each training utterance form the training sequence for the first section codebook. Codewords for this section are obtained using a clustering algorithm [3]; these spectra form the first section codebook. The next n frames from each utterance form the training sequence for the second section codebook, etc. This sequence of L/n section codebooks is a multisection codebook.

To classify an unknown utterance, it is length normalized and analyzed in the same way as the training words. For each multisection codebook, an average distortion value is obtained, as follows: the first n frames are *encoded* using the first section codebook (encoded means that the codeword matching the input frame with minimum distortion is found), the second n frames using the second section codebook, etc. The average distortion between the input word and its encoded representation is found for each multisection codebook; the codebook yielding the minimum distortion determines the classification. For details, see [9, 10]. The distortion measure used in most of our IWR work is the gain-optimized Itakura-Saito distortion [4], also known simply as the Itakura distortion.

All results reported in this paper were obtained using the autocorrelation method of linear predictive analysis and the following analysis conditions: Hamming windowing, analysis window width = 128 points, pre-emphasis = 94%, and analysis filter order = 10. The data was digitized at 8000 samples per second. In addition an energy threshold was used to remove nearly silent frames from the codebook generation and the classification process. All frames with an $R(0)$ (sum-of-squares of the data points) less than 250 were ignored.

SPEAKER RECOGNITION

This IWR technique is applicable to text-dependent speaker recognition as well. For speaker recognition, a codebook is designed for each speaker, rather than for each vocabulary word. The training sequence for each speaker consists of repetitions of a given utterance, which is the same for all speakers. This same utterance is later spoken by speakers to be recognized. The test utterance from the unknown speaker is encoded with the codebook for each speaker, and the speaker whose codebook encodes the utterance with minimum distortion is chosen. The only problem in using the multisection codebooks for speaker recognition is to formulate criteria for rejecting utterances as not matching any acceptable speaker.

In a preliminary experiment that did not attempt to reject speakers, eight female speakers, each speaking the word *zero* , were used. The data were obtained from a Texas Instruments data base that was prepared to test speaker-trained IWR systems [11]. Multisection codebooks for each speaker were designed from ten repetitions of the word *zero* ; each section was four frames long and was represented by a rate-3 codebook. The system was tested using sixteen utterances of *zero* by the same speakers (total 128); only one recognition error was made.

A. Rejection Thresholds

Speaker recognition systems attempt to discriminate one or more "acceptable" speakers from a larger population. Usually, no information is available for characteristics of specific "unacceptable" speakers. The main problem in using the multisection VQ approach in speaker recognition is to formulate criteria for

rejecting utterances as not matching any acceptable speaker.

To decide whether to reject an utterance (speaker), we associate a threshold with each speaker codebook. An unknown utterance is rejected if its distortion exceeds the threshold. To design thresholds, we estimate two Gaussian distributions for each speaker: the *in-class* distribution of distortions (obtained by encoding new utterances from that speaker in his or her codebook) and the *out-of-class* distribution of distortions resulting from utterances spoken by other speakers. We choose the threshold to equalize the overlap area of the two distributions, thus equalizing the expected numbers of imposter acceptances (false acceptances) and rejections of acceptable speakers (false rejections).

In more detail, the threshold computation is as follows. For each speaker, encode new training data with his or her codebook. Compute μ_i^{in}, the mean distortion resulting from encoding the training data from speaker i using the codebook for speaker i, and σ_i^{in}, the corresponding standard deviation. Also compute μ_i^{out}, the mean distortion resulting from encoding utterances *not* spoken by speaker i using the codebook for speaker i, and σ_i^{out}, the corresponding standard deviation. To equalize the number of false acceptances and false rejections, the threshold T_i is chosen to be an equal number of standard deviations away from each mean, giving

$$T_i = \frac{\mu_i^{\text{in}}\sigma_i^{\text{out}} + \mu_i^{\text{out}}\sigma_i^{\text{in}}}{\sigma_i^{\text{out}} + \sigma_i^{\text{in}}} \quad (1)$$

This method of threshold estimation assumes Gaussian distributions. Some previous studies, however, have shown that the logarithms of distortions are more nearly Gaussian than the distortions themselves [12], so we computed the thresholds based on the statistics of the log distortions as well as the statistics of the distortions themselves.

To identify an unknown speaker, first the codebook yielding the minimum distortion is found. Then this minimum distortion value is compared to the threshold associated with that codebook. If the distortion value exceeds the threshold, the speaker is rejected; otherwise the speaker is classified according to this minimum distortion codebook.

B. Evaluating The Method

To evaluate the method, we performed a second experiment using the eight female speakers from the TI data base, speaking a single digit (the experiment was repeated for each of the ten digits). For four of the speakers (ALK, CJP, DFG, and JWS), a rate 3, section size 4, multisection codebook was designed using the first six of the 26 utterances of the digit by that speaker. For each speaker, 10 utterances were available for computing μ_i^{in} and σ_i^{in}, and 70 utterances were available for computing μ_i^{out} and σ_i^{out}. Finally, 10 new utterances from each of the eight speakers were classified using the codebooks and thresholds. The results appear in the Table I.

Few identification errors (misclassification of an acceptable speaker) were made, but only one digit gave more than a 90% accuracy. To improve accuracy, we decided to have speakers say several different words and to base the classification decision on all results for all the words.

EXTENSION TO MULTIPLE WORDS

For each speaker, a separate codebook is designed for each word; if there are N speakers to be accepted and W words to be spoken, there are NW codebooks. For example, if speakers are requested to say *zero*, *three*, and *nine*, the *zero* utterance is encoded with the *zero* codebook for each speaker; the *three* utterance is encoded with the *three* codebook for each speaker, etc.

We investigated two ways of extending the method to multiple words. The first is simply a plurality rule. In this method, separate thresholds are computed for each word, and W different classification decisions are made. The decision made by a plurality of the classifiers is used as the overall decision. In case of ties, the speaker is rejected. For the second method, a *weight* is assigned to each of the W words. The weighted sum of distortions is treated as a single overall distortion, for use both in computing thresholds according to (1) and for classification.

If jointly normal distributions are assumed and the mean vector and covariance matrices are known, an optimal weight vector for each speaker exists, but finding it requires the solution of a system of nonlinear equations [13]. Instead, two heuristic methods were tried. The first method we call *Normalized Mean Weight.* In this method, the weight assigned to speaker i, word j (w_{ij}) is just the reciprocal of μ_{ij}^{in}, the average distortion of utterances of word j spoken by speaker i in C_{ij} (the codebook for speaker i, word j). This rule attempts to normalize the distortions. The second rule, which we call *Weighted Separation*, sets w_{ij} to $\mu_{ij}^{\text{out}}-\mu_{ij}^{\text{in}}$, where μ_{ij}^{out} is the average distortion in C_{ij} of utterances of word j spoken by speakers other than speaker i. This rule tries to weight more heavily words that give better discrimination.

Another rule to try would be the Bayes decision rule, assuming jointly normal densities for the distortions. Unfortunately, the Bayes rule requires good estimates of covariance matrices, which in turn require a good deal of training data; therefore the Bayes rule was not tried.

Table I. Speaker Recognition Results Using A Single Codebook Per Speaker

Digit Spoken	Identification Errors	False Acceptances	False Rejections	Total (of 80)
ZERO	0	13	0	13
ONE	0	14	3	17
TWO	0	6	2	8
THREE	0	16	3	19
FOUR	1	9	1	11
FIVE	1	16	4	21
SIX	1	13	2	16
SEVEN	1	10	8	19
EIGHT	0	6	5	11
NINE	0	4	3	7

Table II. Speaker Recognition Using Multiple Codebooks Per Speaker

Section Rate	Section Size	Digits Used	Plurality Rule		Normalized Mean Weight		Weighted Separation	
			avg. dist.	avg. log. dist.	avg. dist.	avg. log. dist.	avg. dist.	avg. log. dist.
2	4	037	6/1	7/1	11/1	12/1	10/8	10/8
2	4	0379	4/2	5/1	6/2	6/1	4/2	4/2
3	4	0379	3/1	3/1	4/3	4/1	4/2	2/2
3	4	023679	0/1	0/1	0/2	0/1	0/2	0/1
3	4	all	0/0	0/0	0/0	0/0	0/1	0/0
3	4	1279	1/3	2/2	0/3	0/1	0/5	0/4
3	6	all	0/0	0/0	0/0	0/0	0/0	0/0
3	6	1279	0/2	0/1	0/1	0/1	0/2	0/2
3	8	1279	0/0	0/0	0/0	0/0	1/3	1/3

A. Preliminary Experiments

The first set of multiple-word experiments was performed using the same data as for the single word experiments: we attempted to recognize the four female speakers ALK, CJP, DFG, and JWS, given training and test data from these and four other female speakers. The results appear in the Table II. Remember there are 40 possible false acceptances and a total of 40 possible identification errors plus false rejections. No identification errors occurred in any of the experiments; entries in the table are in the form fa/fr where fa is the number of false acceptances and fr is the number of false rejections.

Apparently, from Table II, larger section sizes (6 or 8), higher section rates (3), and larger digit sets improve accuracy. No decision rule emerged as a clear winner, and results using the average of the log distortions were about equal to the results using the average distortions. In the next section, we describe a larger speaker recognition test. Based on the results mentioned above, we chose a section codebook rate of 3, a section size $n = 8$ frames, all 10 digits as the recognition text, the plurality decision rule, and log distortions.

B. Evaluation Of Multiple Word Approach

The multiple word approach was evaluated using data from two different data bases, both collected by Texas Instruments Incorporated. The only difference that we are aware of in the data bases is the resolution of the A/D converters. The data base mentioned above was digitized with a 12-bit converter; on the second data base, a 16-bit converter was used.

Data for designing the digit codebooks for each speaker, computing the in-class distributions, and testing recognition accuracy came from the data base [11] used earlier in this paper. It contains 26 utterances of each digit by 16 speakers (8 male and 8 female). for each speaker - word combination, we used the first 9 utterances to build codebooks, the second 9 to estimate probability distribution parameters, and the last 8 to test the method (for a total of 128 in class test utterances). The out-of-class data came from a data base designed for evaluating speaker-independent recognition of the digits [14]. The out-of-class distributions were estimated using the training portion of this data base. We used 109 sets of digit utterances (one set from each of 54 male and 55 female speakers) in the out-of-class distribution estimates. The imposter data came from the separate testing portion of the same speaker-independent digit data base. We used two utterances from 55 male and 56 female speakers, for a total of 222 imposter utterance sets.

The results are shown in Table III. Each row comprises the results from one speaker in the recognition population; the

Table III. Speaker Recognition Confusion Matrix

	ALK	CJP	DFG	GNL	GRD	HNJ	JWS	KAB	MSW	REH	RGL	RLD	SAS	SJN	TBS	WMF	Reject
ALK	8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
CJP	.	8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
DFG	.	.	8	.	.	.	.	.	.	.	.	.	.	.	.	.	.
GNL	.	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	1
GRD	.	.	.	.	8	.	.	.	.	.	.	.	.	.	.	.	.
HNJ	.	.	.	.	.	8	.	.	.	.	.	.	.	.	.	.	.
JWS	.	.	.	.	.	.	8	.	.	.	.	.	.	.	.	.	.
KAB	.	.	.	.	.	.	.	8	.	.	.	.	.	.	.	.	.
MSW	.	.	.	.	.	.	.	.	8	.	.	.	.	.	.	.	.
REH	.	.	.	.	.	.	.	.	.	8	.	.	.	.	.	.	.
RGL	.	.	.	.	.	.	.	.	.	.	8	.	.	.	.	.	.
RLD	.	.	.	.	.	.	.	.	.	.	.	8	.	.	.	.	.
SAS	.	.	.	.	.	.	.	.	.	.	.	.	8	.	.	.	.
SJN	.	.	.	.	.	.	.	.	.	.	.	.	.	8	.	.	.
TBS	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	.	.
WMF	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	.
Other	.	.	.	.	.	1	.	1	.	.	2	.	.	.	.	.	218

columns correspond to the classification decisions. The final row represents the classification results from all the imposter attempts, and the final column represents rejections. The method made no identification errors, only 1 false rejection, but 4 false acceptances. All falsely accepted speakers were accepted as speakers of the correct sex. Although the thresholds for each word were designed to give equal false rejection and false acceptance error probabilities, the results using all 10 words and a plurality rule showed a bias toward false acceptances, generally considered the "worst" of the two error types. We see several ways of changing this bias.

The estimation of the parameters for each in-class distribution is based on only 9 training utterances. These estimates are susceptible to outliers in the training data, which can be caused by endpoint detection errors. We believe using more training data for the in-class distribution estimates would reduce this bias problem. An alternative that would change the bias from false acceptances to false rejections is to base the decision on a majority rule, instead of a simple plurality. None of the false acceptances in this test would have occurred if we had required at least 6 of the 10 word-based decisions to agree on a speaker; of course the number of false rejections would have increased.

SUMMARY

Several ways of using VQ source coding in speaker recognition were examined. We used as a discrimination function the average quantization distortion that results from encoding known utterances in codebooks designed to represent each member of a small population of speakers. No single word (digit) provided reliable discrimination, but by incorporating all 10 digits into the recognition process, we achieved a false rejection rate of 0.8% and a false acceptance rate of 1.8% on a combined data base that contained 16 acceptable speakers and 111 imposters. These results illustrate that specially designed VQ codebooks can be useful in speaker recognition systems.

ACKNOWLEDGEMENTS

We thank Rod Johnson for helpful discussions, and we thank Tom Schalk and Gary Leonard for their help in obtaining the data bases.

References

1. A. Buzo, A. H. Gray, Jr., R. M. Gray, and J. D. Markel, "Speech coding based upon vector quantization," *IEEE Trans. Acoust., Speech, Signal Processing* **ASSP-28**, pp. 562-574 (Oct. 1980).

2. R. M. Gray, A. H. Gray, Jr., G. Rebolledo, and J. E. Shore, "Rate-distortion speech coding with a minimum discrimination information distortion measure," *IEEE Trans. Inform. Theory* **IT-27**, pp. 708-721 (Nov. 1981).

3. Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.* **COM-28**, pp. 84-95 (Jan. 1980).

4. R. M. Gray, A. Buzo, A. H. Gray, Jr., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing* **ASSP-28**, pp. 367-376 (August 1980).

5. N. Sugamura, K. Shikano, and S. Furiu, "Isolated Word Recognition Using Phoneme-Like Templates," pp. 723-726 in *Proceedings of ICASSP 1983, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Boston, Mass. (April, 1983).

6. R. Pieraccini and R. Billi, "Experimental Comparison Among Data Compression Techniques In Isolated Word Recognition," pp. 1025-1028 in *Proceedings of ICASSP 1983, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Boston, Mass. (April, 1983).

7. K.-C. Pan , "Isolated Word Recognition Based Upon Vector Quantization Techniques," Master's Thesis, Massachusetts Institute of Technology, Cambridge, Mass. (1984).

8. J. E. Shore and D. K. Burton, "Discrete utterance speech recognition without time alignment," *IEEE Trans. Inform. Theory* **IT-29**, pp. 473-491 (July, 1983).

9. D. K. Burton, J. E. Shore, and J. T. Buck, "A generalization of isolated word recognition using vector quantization," pp. 1021-1024 in *Proceedings of ICASSP 1983, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Boston, MA (April, 1983). IEEE 83CH1841-6.

10. David K. Burton, John E. Shore, and Joseph T. Buck, "Isolated-Word Speech Recognition Using Multi-Section Vector Quantization Code Books," *IEEE Trans. Acoust., Speech, Signal Processing* (1985). to appear

11. G. R. Doddington and T. B. Schalk, "Speech recognition: turning theory to practice," *IEEE Spectrum* **Vol 18, No. 9**, pp. 26-32 (Sept. 1981).

12. Joseph T. Buck, "Vector quantization code book distortions as features for maximum likelihood classification of isolated words," pp. 9.3.1-9.3.5 in *Proceedings of 1984 IEEE Global Telecommunications Conference (GLOBECOM)*, Atlanta, GA (Nov. 1984).

13. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press (1972).

14. R. Gary Leonard, "A Database for Speaker-Independent Digit Recognition," *Proceedings of 1984 ICASSP Conference*, pp. 42.11.1-42.11.4 (March, 1984).