APPLYING MATRIX QUANTIZATION TO ISOLATED WORD RECOGNITION

David K. Burton

Computer Science and Systems Branch, Information Technology Division,
Naval Research Laboratory, Washington D.C. 20375

ABSTRACT

A new approach to isolated word recognition is examined. This approach is based on an extension of vector quantization speech coding, called matrix quantization speech coding, that was developed by Tsao and Gray. In this new approach, a codebook containing a set of time-ordered-sequences of speech spectra represents each vocabulary word. A word is recognized by encoding it with each codebook and classifying the input word according to the codebook that yields the smallest distortion. On the digits, this approach achieved a speaker independent recognition accuracy greater than 98%. The approach is described, experimental results are presented, and comparisons with vector quantization based approaches are given.

INTRODUCTION

In this paper we describe a new method of isolated word recognition. The method is based on a narrow bandwidth speech coding technique that represents each 100 or so milliseconds of speech as one of a set of time-ordered-sequences of linear predictive coefficients (LPC) [1]. This speech coding technique is a generalization of another narrow bandwidth technique called vector quantization speech coding [2]; because in this generalization blocks of vectors are used, it is called *matrix quantization* speech coding.

Vector quantization (VQ) is an information-theoretic data compression technique. In VQ, a vector from an information source is represented by one of a prestored set of *codewords*. This set of codewords is called a *codebook*. A source vector is encoded in the codebook by finding the codeword that minimizes the distortion between itself and the source vector. Data compression is achieved by transmitting or storing the codeword's index in the codebook rather than the parameters of the original source vector.

In speech coding by VQ [2], the power spectrum corresponding to each frame (20 millisecond or so segment) of speech is represented by a vector of short-time autocorrelations and coded for transmission as one of a prestored set of characteristic spectra. The codebook of characteristic spectra is designed using an iterative clustering procedure on a long training sequence comprised of conversational speech [3]. Recently, Tsao and Gray developed a natural extension of VQ speech coding [1]. This extension considers each sequence of autocorrelation vectors (or equivalently LPC vectors) as a single symbol from a source alphabet. Each vector of source symbols (sequence of autocorrelations or LPC vectors) is an element to be coded in a codebook that contains representative power spectrum sequences. This extension, called matrix quantization (MQ), is motivated by rate-distortion theory, just as the VQ of individual samples is motivated.

We present results from using matrix quantization for isolated word recognition. In addition to describing this new approach and reporting recognition accuracies, we compare the accuracy and computational cost of the MQ approach with those of our earlier VQ approaches [4, 5, 6].

BACKGROUND

Since the introduction of VQ speech coding, VQ has been incorporated into isolated word recognition in many ways [7, 8, 9, 10, 5, 11, 12, 13, 4]. Previously, we incorporated VQ into isolated word recognition in two separate but related ways [4, 5]. In both approaches we represent each word in the recognition vocabulary by a specially designed VQ codebook. In this section, we briefly describe our previous isolated word recognition approaches and MQ speech coding.

A. Previous Approaches

Our original isolated word recognition approach recognizes words without using any time sequence information [4]. A VQ codebook containing $N=2^R$ codewords is designed for each vocabulary word from a training sequence comprised of several repetitions of that word. We call R the *codebook rate*. An unknown word is recognized by dividing it into frames, encoding it in each of the word codebooks, and classifying it according to the codebook that represents the unknown word with the smallest average distortion. We now call this approach *single section*.

A more recent approach incorporates time sequence information into the recognition process [5]. First each training utterance for a word is divided into L equal-size frames, and then the frames in each training utterance are grouped into equal-length sections containing n frames. (We call n the *section length*.) From the first section of each of the training utterances, we design a small VQ codebook, called a *section codebook*, containing 2^{R_s} codewords. The next n frames from each utterance form the training sequence for the second section codebook, and so on. We call R_s the *section codebook rate* and the sequence of L/n section codebooks a *multisection codebook*. After designing a multisection codebook to represent each vocabulary word, an unknown word is recognized by dividing it into n-frame sections and encoding it on a section-by-section basis in each of the multisection codebooks. As before, the word is classified according to the codebook that represents the unknown word with the smallest average distortion.

B. Matrix Quantization Speech Coding

In MQ speech coding [1], the speech signal is first divided into frames of about 20 milliseconds and the spectrum shape in

each frame is represented by an M^{th} order autoregressive model that is found by using the autocorrelation method of linear predictive analysis on the sample autocorrelations. Let $\mathbf{x}$ represent the LPC coefficients corresponding to a frame of speech. The sequence of speech spectrum shapes in each 100 millisecond or so speech segment is coded as the index corresponding to one of a prestored set of spectrum-shape sequences called *codeword matrices* (or simply codewords); the set of codeword matrices is a codebook.

More precisely, let $\mathbf{X} = [\, x_1, x_2, ..., x_K \,]$ represent the power-spectrum sequence corresponding to an input sequence of K speech frames, and let the codebook $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_N\}$ represent the collection of codewords, where each codeword $\mathbf{c}_i = [c_{i1}, c_{i2}, ..., c_{iK}]$ is a sequence of speech spectra. Now the input speech segment $\mathbf{X}$ is coded for transmission by finding the codeword $\mathbf{c}_i$ that best represents $\mathbf{X}$ in the following sense

$$D(\mathbf{X}, \mathbf{c}_i) = \min_j D(\mathbf{X}, \mathbf{c}_j), \tag{1}$$

where the distortion between a speech segment and the j^{th} codeword is

$$D(\mathbf{X}, \mathbf{c}_j) = \sum_{l=1}^{K} d(x_l, c_{jl}),$$

d is an appropriate vector distortion measure, and j ranges from $1 \cdots N$. The distortion measure d used in [1] is the Itakura-Saito distortion measure [14].

A matrix quantization codebook $\mathbf{C}$ is designed to minimize the average distortion that results from encoding a long training sequence of conversational speech. If a training sequence contains Q speech frames $\{t_1, t_2, ..., t_Q\}$, the codebook $\mathbf{C}$ is found that minimizes

$$\sum_{p=1}^{Q-K+1} D(\mathbf{T}_p, \mathbf{c}_j), \tag{2}$$

where $\mathbf{c}_j$ is the codeword resulting from encoding the speech segment $\mathbf{T}_p$, and

$$\mathbf{T}_p = [t_p, t_{p+1}, ..., t_{p+K-1}].$$

The MQ codebook design algorithm [1] is a generalized version of the VQ design algorithm developed by Linde et al [3]. This algorithm finds a codebook that is at least locally optimum for representing the training sequence. Again, the size of the codebook $\mathbf{C}$ is a power of 2 – i.e., $N = 2^{R_M}$.

ISOLATED WORD RECOGNITION

In this section, we describe our new isolated word recognition approach and specify distortion measures, LPC parameters, and the data base that were used in the experiments.

A. Approach

To use MQ in isolated word recognition, we design a separate MQ codebook for each word in the recognition vocabulary. Each vocabulary-word codebook is designed from a training sequence comprised of many repetitions of a word. An input word is classified by dividing it into sequences of K frames, encoding each K frame sequence in each of the word codebooks, and computing the average distortion by which each codebook represents the input word. The codebook yielding the smallest average distortion determines the input word's classification.

In particular, suppose V is the number of words in the recognition vocabulary. Then there are V codebooks $\mathbf{C}_i$, i = 1, ..., V. Let $\mathbf{c}_{ij}$, j = 1, ..., R_M be the codeword matrices in $\mathbf{C}_i$. Now, suppose an input word contains N frames, and the power spectrum corresponding to frame S_r ($r = 1, ..., N$) is x_r. Finally, let Δ_i be the average distortion resulting from encoding the unknown word with $\mathbf{C}_i$,

$$\Delta_i = \frac{1}{N} \sum_{r=0}^{ceil\,[\frac{N}{K}]} \min_j D(\mathbf{X}_{rK+1}, \mathbf{c}_{ij}), \tag{3}$$

where *ceil* [Z] means the smallest integer greater than or equal to Z, and where $(ceil\,[\frac{N}{K}]) \times K - N$ flat spectrum, low energy frames are added to the end of the word. The unknown word is then classified as the r^{th} vocabulary word, where

$$\Delta_r = \min_i \Delta_i .$$

Note that the classification procedure *jumps* through the word to be recognized. That is, the K vector sequence to be encoded moves ahead K frames each time. We also evaluated a classification procedure in which the K-frame sequence *slides* through the unknown word. We attach K–1 low energy, flat spectrum frames to both ends of the unknown word before sliding the K frame window through the data. We call the former approach jump classification and the latter slide classification.

B. Distortion Measures

In generating VQ codebooks for voice coding, two distortion measures are effective [2, 15]: the *Itakura-Saito* (d_{IS}) and the *gain normalized Itakura-Saito* (d_{GN}). We used both the d_{GN} and d_{IS} distortion measures in (2) to design MQ codebooks.

For the classification distortion measure in (3), we used the d_{IS} and the *gain optimized Itakura-Saito* distortion measure, which is also known as the log likelihood distortion measure. Properties of all three distortion measures are discussed in [14].

C. LPC Parameters

LPC parameters for both codebook generation and utterance classification were generated using the autocorrelation method of linear predictive analysis with Hamming windowing. We chose analysis conditions for compatibility with the Navy's 2.4-kbs LPC-10 system[16]: frame update size = 180 points, analysis window width = 130 points, filter order = 10, and pre-emphasis = 94%.

D. Data Base

Our experiments were conducted using a data base that was prepared by Texas Instruments, Inc. (TI) [17]. The data base contains data from 111 adult males, 114 adult females and 101 children (50 boys and 51 girls), and it is divided into two separate parts: a training portion and a testing portion. The training portion contains data from 55 males, 57 females, and 51 children; the testing portion contains data from the rest of the speakers. The data was collected in an acoustically treated sound room and digitized at 20,000 samples per second using a 16-bit A/D converter. We received the data after it had been down sampled to 8000 samples per second. Although the data base contains the 10 digits (*zero* through *nine*) and the word *oh* for all the speakers, for compatibility with other work we used only the isolated digits spoken by the adults. The data base contains two utterances of each digit by each adult speaker for a total of 2240 training and 2260 test digits.

RESULTS

This section presents results from using MQ in isolated word recognition and compares them with single section VQ and multisection VQ results. Results are for the 2260 test digits in the TI data base; codebooks were designed from the training portion of that data base.

Table I. Digit Recognition Accuracy Using Matrix Quantization: Codebook Rate = 4, Codebook Design Distortion = d_{GN}, And Classification Distortion = d_{GO}.

Classification Procedure	# Of Class.	Matrix Size (K) 2	3	4	6	8	12
Jump	2260	91.3	93.4	93.6	93.9	94.4	94.4
Slide	2260	92.7	94.7	95.0	96.7	97.3	97.3

A. Matrix Quantization Results

In our first set of experiments, we examined the effect of the matrix size K on recognition accuracy. We used d_{GN} as the codebook design distortion measure in (2), d_{GO} as the classification distortion measure in (3), and a codebook rate R_M = 4; we varied K from 2 to 12. The results are shown in Table I for both jump- and slide-classification procedures. For all values of K, slide classification achieved a higher recognition accuracy than jump classification. Both classification procedures did best for matrix sizes of 8 and 12, and the best recognition accuracy was 97.3%.

Based on the recognition accuracies shown in Table I and because the computational complexity grows linearly with K for the slide-classification procedure, we fixed K at two values: 6 and 8. We then evaluated the MQ approach using the d_{IS} distortion measure for both codebook design and utterance classification. The results show, see Table II, classification using d_{IS} on d_{IS} codebooks is always significantly worse than classification using d_{GO} on d_{GN} codebooks. Also, increasing R_M to 5 significantly improves the recognition performance of all approaches. Once again, slide classification did better than jump classification. The best recognition accuracy was 98.3%, achieved using slide classification on rate-5, d_{GN} codebooks.

B. Multisection VQ Results

We tested the multisection approach using L = 24, n = 4, various section codebook rates (R_S), and the two distortion measure combinations used in the MQ tests. In addition, we also pre-normalized each utterance to have the same average power, and then used d_{IS} in codebook design and utterance classification. We hoped this normalization would reduce the variability of the gain term for similar spectrum shapes, and this would in turn improve the recognition accuracy when using the d_{IS} distortion measure. Results are in Table III (top of next page).

Clearly, from Table III, power normalization improves the recognition performance when using the d_{IS} distortion measure. The power normalized results, however, are only about as good as those using d_{GO} as a classification distortion measure on d_{GN} codebooks. The power normalized approach achieved the best multisection recognition accuracy – 98.1%.

C. Single Section VQ Results

Table IV contains the results for various codebook rates (R) and distortion measure combinations. Using single section codebooks, the power normalized approach is clearly superior to both the other distortion measure approaches. The highest accuracy obtained using the single-section approach, however, was only 94.3%. A dramatic decrease from the accuracies measured using the multisection and the MQ approaches.

Table IV. Digit Recognition Accuracy Using Single Section Vector Quantization Codebooks.

Distortion Measures	# Of Class.	Codebook Rate = 5	Codebook Rate = 6
d_{IS}	2260	90.5	92.6
Power Normal. d_{IS}	2260	93.4	94.3
d_{GO}	2260	91.7	93.5

DISCUSSION

Two ways to compare the performance of the various approaches are accuracy alone and accuracy for a given computational cost. Considering only recognition accuracy, the ranking of the approaches from most to least accurate is as follows: MQ slide classification (98.3%), multisection VQ (98.1%), MQ jump classification (95.8%), and single section VQ (94.3%).

For a fixed computational cost, however, the ranking is different. For all four approaches, the computation requirements are dominated by the computation of the spectrum distortions. For all but the MQ slide classification, the number of distortion computations per input speech frame per vocabulary word is equal to the codebook rate – R, R_S, or R_M. Because in the MQ slide classification the K-vector sequence slides through the unknown word one frame at a time, each frame in the unknown word is classified K times, instead of just once. Thus, slide classification requires $K \times R_M$ distortion computations per input speech frame per vocabulary word. Fixing the number of distortion computations per input frame at 32, the ranking of the approaches is as follows: multisection VQ (97.7%), jump MQ (95.8%), single section VQ (93.4%), and slide MQ (92.7% for R_M = 4 and K = 2).

Table II. Digit Recognition Accuracy Using Matrix Quantization Codebooks.

Classification Procedure	# Of Class.	Codebook Rate = 4, K = 6 d_{IS}	d_{GO}	K = 8 d_{IS}	d_{GO}	Codebook Rate = 5, K = 6 d_{IS}	d_{GO}	K = 8 d_{IS}	d_{GO}
Jump	2260	86.2	93.9	80.2	94.4	89.3	95.7	87.2	95.8
Slide	2260	93.7	96.7	94.3	97.3	95.6	97.9	95.9	98.3

Table III. Digit Recognition Accuracy Using Multisection Vector Quantization Codebooks: L = 24 And n = 4.

Distortion Measures	# Of Class.	Section Codebook Rate = 4	Section Codebook Rate = 5	Section Codebook Rate = 6
d_{IS}	2260	96.6	97.1	97.7
Power Normal. d_{IS}	2260	97.0	97.4	98.1
d_{GO}	2260	96.9	97.7	97.4

Based on rate-distortion theory, we expected MQ codebooks to represent words better than VQ codebooks, and we hoped this would result in higher recognition accuracies when using MQ codebooks. As expected, for equal computational requirements ($R_M = R = 5$), MQ jump classification achieved a recognition accuracy 2.4% better than single section VQ.

The multisection VQ approach incorporates temporal information that the single section VQ approach ignores, and this improves the accuracy and reduces the computational requirements of the classification approach. The MQ slide classification incorporates temporal information that the jump classification ignores; the result is an improved recognition accuracy. Unlike the multisection VQ approach, however, the increased accuracy is achieved at the cost of a large increase in computational requirements. Because of the inherent advantage MQ has over VQ, properly incorporating temporal information into MQ classification should result in significantly better recognition performance than that achieved using multisection VQ. To be useful in isolated word recognition, however, the temporal information must be incorporated in a way that reduces the computational requirements of the MQ approach.

ACKNOWLEDGEMENTS

I thank Joe Buck, Rod Johnson, and John Shore for helpful discussions and comments about this paper, Joe Buck for writing most of the software, and Gary Leonard for his help in obtaining the data base.

References

1. C. Tsao and R. M. Gray, *Matrix quantizer design for LPC speech using the generalized Lloyd algorithm*, draft, 1984.
2. 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).
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. 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).
5. 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.
6. 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
7. R. Hamabe, Y. Yamada, M. Murata, and T. Namekawa, "A speech recognition system using inverse filter matching technique," *Proc. Annual Conf. Inst. of Television Engineers*, Kyushu University (June 1981). (in Japanese)
8. J. E. Shore and D. Burton, "Discrete utterance speech recognition without time normalization," pp. 907-910 in *Proceedings of ICASSP 1982, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Paris, France (May, 1982). IEEE 82CH1746-7.
9. A. Buzo, C. Riviera, and H. Martinez, "Discrete utterance recognition based upon source coding techniques," pp. 539-542 in *Proceedings of ICASSP 1982, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Paris, France (May, 1982). IEEE 82CH1746-7.
10. R. Billi, "Vector quantization and Markov source models applied to speech recognition," pp. 574-577 in *Proceedings of ICASSP 1982, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Paris, France (May, 1982).
11. 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).
12. 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).
13. L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition," *The Bell Systems Technical Journal* **Vol. 62, No. 4**, pp. 1075-1105 (April, 1983).
14. 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).
15. B.-H. Juang, D. Y. Wong, and A. H. Gray, Jr., "Distortion performance of vector quantization for LPC voice coding," *IEEE Trans. Acoust., Speech, Signal Processing* **ASSP-30**, pp. 294-303 (April, 1982).
16. T. E. Tremain, "The government standard linear predictive coding algorithm: LPC-10," *Speech Technology* **1**, pp. 40-49 (April 1982).
17. R. Gary Leonard, "A Database for Speaker-Independent Digit Recognition," *Proceedings of 1984 ICASSP Conference*, pp. 42.11.1-42.11.4 (March, 1984).