WEIGHTED CEPSTRAL DISTANCE MEASURES IN VECTOR QUANTIZATION BASED SPEECH RECOGNIZERS

Ted H. Applebaum, Brian A. Hanson and Hisashi Wakita

Speech Technology Laboratory
3888 State Street, Santa Barbara, CA 93105, USA

ABSTRACT

This paper extends the use of weighted cepstral distance measures to speaker independent word recognizers based on vector quantization. Recognition results were obtained for two recognition methods: dynamic timewarping of vector codes and hidden Markov modeling. The experiments were carried out on a vocabulary of the ten digits and the word "oh". Two kinds of spectral analysis were considered: LPC, and a recently proposed, low dimensional, perceptually based representation (PLP). The effects of analysis order and varying degrees of quantization in the spectral representation were also considered.

Recognition experiments indicate that the performance of the weighted cepstral distance with vector quantized spectral data is considerably different from that previously reported for unquantized data. Comparison of recognition rates shows wide variations due to interaction of the distance measure with the analysis technique and with vector quantization. The best recognition scores were obtained by the combination of weighted cepstral distance and low order PLP analysis. This combination maintained good recognition rates down to very low (16 or 8 codes) codebook sizes.

I. INTRODUCTION

Weighted cepstral distance measures have recently been shown to be superior to Euclidian distance in the cepstral domain for several speech recognition tasks which use *unquantized* linear predictive cepstral coefficients [1-5]. This paper concerns the use of a representative weighted cepstral distance measure, root power sums [1], with *vector quantized* representations of speech for *speaker-independent* word recognition. Use of vector quantization (VQ) based recognizers is considered because they reduce both the computational and memory requirements on the recognition processor [6,7]. The use in speaker independent recognition of weighted cepstral distances with vector quantized data has not previously been investigated.

Cepstral distance (CEP) is the Euclidian distance between two sets of cepstral coefficients. CEP distance expresses the difference between all pole model spectra when the cepstral coefficients are recursively derived from autoregressive coefficients, such as those obtained from linear prediction (LPC) analysis [8]. Alternatively, the cepstral coefficients may be obtained from perceptually-based linear prediction (PLP) analysis [9,10] which will be considered in this study since it has been shown that weighted cepstral distances work well with this analysis method for speaker dependent recognition [4,10]. *Weighted cepstral distance* [1,2] is simply the Euclidian distance between cepstral coefficients for which each term of the sum is multiplied by a predetermined weighting coefficient w_k:

$$distance = \sum_{k=1}^{N} (w_k (c_{Tk} - c_{Rk}))^2 \qquad (1)$$

When constant weighting is used, this reduces to the standard cepstral distance.

Triangular weighted cepstral distances comprise the subclass of weighted cepstral distance measures for which the weighting factor (w_k) increases linearly with the index (k). The *Root Power Sums* (RPS) distance measure [1] is a special case of triangular weighted cepstral distance measure for which cepstral weights equal the summation index. CEP (Euclidian distance in the cepstral domain) is the degenerate case of triangular weighted cepstral distance with $w_k = 1$. Other cepstral weights have been proposed including a raised sine [3] and the inverse of the standard deviations of the cepstral parameters [2]. It is argued in [11], however, that these yield similar recognition performance to RPS. We will use RPS throughout this paper as a representative of weighted cepstral measures, assuming that small differences in recognition rate obtained with other cepstral weightings will not invalidate the trends we observe for the RPS distance.

In [2,3,11] the tradeoffs between using various numbers of difference terms in eqn. (1) are considered for several weighting functions. Speaker dependent [11] and speaker independent [3] recognition experiments have shown that, for triangular weighted cepstral distance measures, recognition performance is best when the number of cepstral difference terms are approximately equal to the order of the all-pole model. However, these studies considered only fixed analysis orders. The effects of varying the order of the spectral analysis used with the RPS distance have recently been demonstrated to be important [10]. We will consider the effect of analysis order with the RPS distance, since it is an open issue for vector quantization based speech recognition.

Another important consideration in applying vector quantization-based recognition is optimal codebook size. The computational efficiency of the recognition depends strongly on the size of the vector quantization codebook. We will find a combination of analysis method and distance measure which yields high recognition rate at small codebook sizes.

II. VQ BASED RECOGNITION SYSTEMS

Isolated word recognition experiments were performed to evaluate the effectiveness of distance measures in two types of vector-quantization based, speaker independent speech recognizers. The recognizers, based on dynamic time warping and hidden Markov modeling, were of conventional design. The characteristics of the particular recognition systems, and of the database used to test them, are summarized below.

A. Database. The speech database comprised 1056 isolated words. It consisted of the digits "zero" through "nine" plus the word "oh" as uttered once each by 48 male and 48 female speakers of American English. The data were recorded in five widely separated American cities, chosen to sample the major dialects of the United States. The speech was tape recorded in office-like environments using a free standing microphone. The speech data were grouped by speaker into four "teams" of 12 male and 12 female speakers each. The recorded speech was lowpass filtered and digitized to 16 bit accuracy at a sampling rate of 10 kHz. Endpoints were determined from energy

CH2396-0/87/0000-1155 $1.00

and zero-crossing rate, and hand corrected by inspection of the digital spectrogram.

B. Analysis. Either the standard LPC or the PLP analysis technique was applied to the sampled data. Each analysis was done with a 20 msec Hamming window and an analysis step of 10 msec. A pre-emphasis factor of 0.98 was used in the LPC analysis. The analysis procedure created six alternative representations of the speech: LPC or PLP at 5th, 8th or 14th order. Fifth order PLP, which gives a two peak representation, was chosen as well as higher orders of PLP which have also been recently shown to have merit [4,10]. In addition to 14th order LPC (typical for 10 kHz sampling rate [8]), fifth and eighth order LPC were investigated for comparison with PLP and earlier studies on weighted cepstral distance [2,3].

C. Distance Measures. Both the standard Euclidean distance in the cepstral domain (CEP) and root-power sums (RPS) distance, as evaluated with eqn. (1) using $w_k = 1$ and $w_k = k$ respectively, were considered. Note that the summation in eqn. (1) begins with c_1, rather than the gain term c_0, so energy is not included in either the CEP or RPS distances. The distance measure enters into the recognition experiments in both systems through its use in vector quantization clustering and code assignment, and additionally in the dynamic time warping system where it is used as a similarity measure for comparing test and reference utterances.

D. Vector Quantization. Vector quantization consists of two stages: determining a codebook which is representative of the training data (clustering) and assigning a vector code to each speech frame (assignment). Clustering uses the distance measure to assign frames to clusters by the nearest neighbor rule, and to average frames in the clusters to compute centroids. Assignment also uses the distance measure to assign each test frame to a cluster centroid from the codebook.

The data from each team were down sampled to approximately 3500 speech frames, and clustered by the k-means [12] algorithm. Sets of vector quantization codebooks were created for each team of speakers, using all four combinations of analysis method (LPC or PLP) and distance measure (CEP or RPS). Each set contained codebooks of size 8, 16, 32, 64 and 128. Successively larger codebooks were generated by the cell splitting technique described in [13].

After codebooks were generated, each frame of analyzed speech was assigned a codebook index, as determined in a full search of the codebook. This code assignment used the same distance measure as was used in generating the codebook. For all distance measures considered in this paper, the complexity of assigning codes is linear with the order of analysis and with the size of the codebook.

E. Dynamic Time Warping Recognition System. The dynamic time warping (DTW) word recognition system used vector quantized test data and reference templates, as in the "double split" method [7] and the "full search two-sided quantization" method [14]. The spectral distance between every pair of codes in the codebook was computed and stored prior to recognition. Each team of 24 speakers was used as reference in turn, giving 24 reference templates per vocabulary word. This multi-template system used each reference utterance as a reference template, since optimum template selection was not considered in this study.

F. Hidden Markov Modeling Recognition System. The hidden Markov model (HMM) word recognition system [6] used ten-state left-to-right models with self-transitions and transitions to the first succeeding state. To train individual word models, the vector quantized data for all productions of that word uttered by the 24 speakers in one team were processed by ten iterations of the Baum-Welch algorithm. A post-processing step which set the lowest emission probability to 10^{-4} was included to compensate for the finite size of the training set. To recognize words, the vector quantized input utterance was compared to the models by the Viterbi algorithm.

III. VQ BASED RECOGNITION RESULTS

As summarized in the above, the experiments encompassed two analysis methods, two distance measures, five codebook sizes, and two recognition systems: one based on dynamic time warping and one based on hidden Markov modeling. Both recognition systems relied on vector quantized cepstral coefficients.

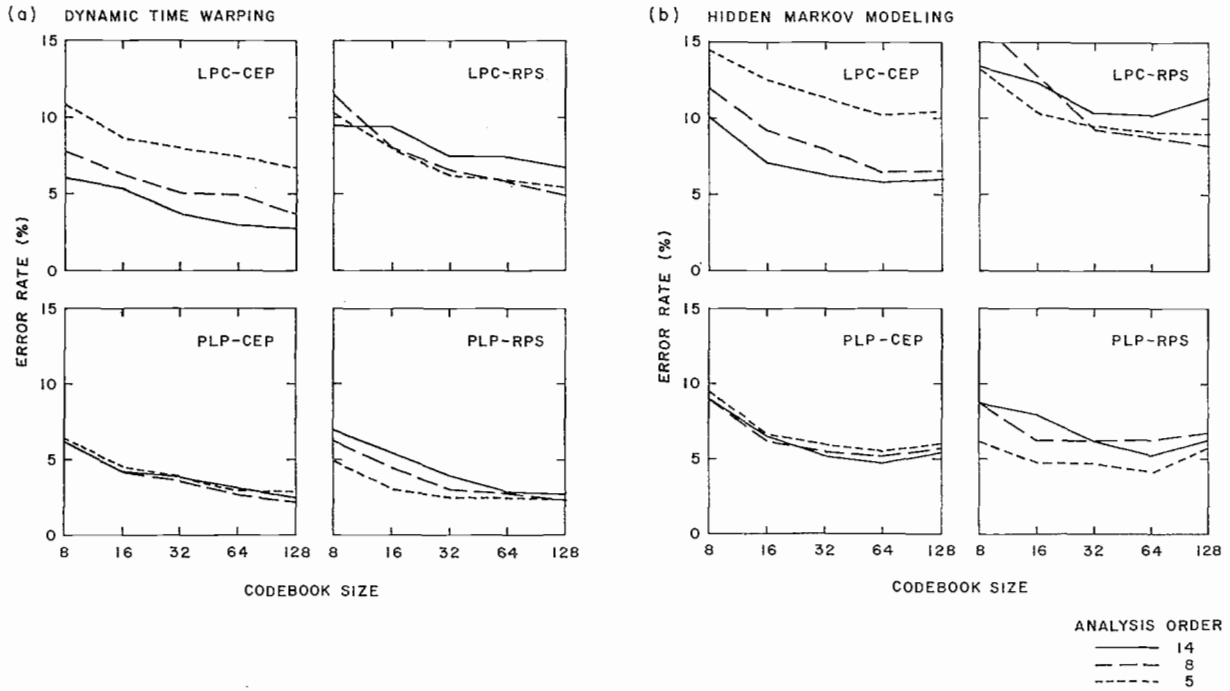

Fig. 1 Recognition error rates, averaged over all open test comparisons, as a function of vector quantization codebook size, analysis order, analysis method, and distance measure. Each data point represents 3168 recognition trials.

Each of the four teams of speakers was used in turn for training in each experiment, and the remaining three teams were used for testing. The data from the "training team" of 24 speakers was used to generate the codebooks. The same training data was used to create the reference templates in the DTW system, or Markov models in the HMM system.

Fig. 1 shows the recognition error rates averaged over the four training teams, plotted as a function of vector quantization codebook size. Each data point in Fig. 1 represents 3 test teams $\times$ 24 speakers/team $\times$ 4 training teams $\times$ 11 words = 3168 recognition trials. Each trial is the comparison of a test word to 264 reference templates (DTW) or 11 Markov models (HMM).

The error rates from the dynamic time warping recognition system are shown in Fig. 1(a). Results for each combination of analysis method and distance measure are presented separately in four graphs. In each graph the results improve with increasing codebook size. With both analysis techniques the best results for RPS distance are obtained from the lower orders (i.e. 5 or 8), whereas the best results for CEP distance are obtained from the higher orders. Comparing best results from each distance measure, the RPS results are clearly worse than those from the CEP distance for LPC. The best results overall come from PLP with either CEP or RPS distance.

The error rates from the hidden Markov modeling recognition system are shown in Fig. 1(b). Although absolute error rates are higher, the dependencies of error rate on analysis method and distance measure are similar to those observed above for the dynamic time warping system. As expected, due to the limited amount of training data, the dependency of error rate on codebook size for HMM is different from that of the DTW recognizer in that the best results are obtained at intermediate (16-64) values of codebook size.

For comparison with published results we also determined recognition rates with a DTW recognizer based on unquantized spectral data [15]. Fig. 2 contrasts the results from this recognizer with the comparable results from the VQ-based DTW recognition system for the finest and coarsest quantization. These experiments used each of three teams of speakers as reference in turn, and the remaining two teams as test. Recognition error rate increases in all cases as the quantization is made more coarse. The increase in error rate is most rapid for 8th order LPC combined with RPS distance.

IV. DISCUSSION

The results summarized in the previous section highlight the importance of interactions between analysis method and distance measure. More significantly, comparison of these results with earlier work [2,3] indicates the importance of the interaction of distance measures with vector quantization (VQ). Finally, the combination of RPS distance measure with PLP analysis has been shown to provide higher recognition rate and efficiency when compared to LPC analysis for vector quantization based recognizers. These points are elaborated below.

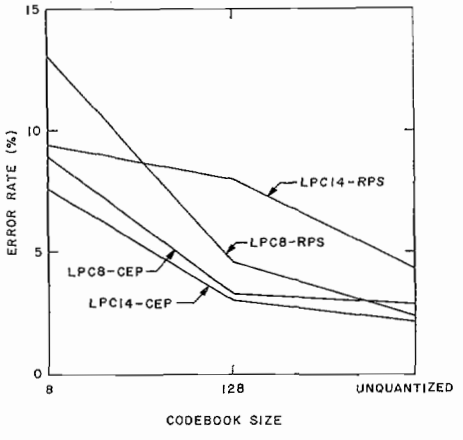

Fig. 2 Dynamic time warping recognition error rates for quantized and unquantized spectral data. Each data point represents 1584 recognition trials.

A. Interaction of Distance Measure and Analysis Method. Weighted cepstral distance measures have been shown [2,3] to have advantages in speaker independent recognition for 8th order LPC analysis. The effects of analysis order with the RPS form of weighted cepstral distance have been considered for speaker dependent recognition [4,10] and cross speaker recognition [5]. In the current study, the effect of analysis order on recognition rate is investigated for VQ-based speaker independent recognizers which use the RPS distance measure. In contrast to the standard CEP distance case, which achieves the best recognition at higher analysis orders, we have shown that RPS achieves better results at lower analysis orders (i.e. 5 or 8). Finally, the uniformly good recognition results obtained from the VQ-based recognizers with PLP analysis using *either* CEP or RPS distance contrast sharply with the results for LPC where CEP distance works well and RPS distance does not. This further emphasizes the importance of the interaction between distance measures and the analysis technique.

B. Interaction of Distance Measure with Vector Quantization. We have shown that for VQ-based recognizers the LPC analysis method achieves much better recognition rates with CEP distance than with RPS distance. However, several researchers [2,3] have found that 8th order LPC analysis achieves better recognition rates with RPS than with CEP. The major difference between the current work and these earlier studies is that the current work is based on vector quantized rather than unquantized speech spectra.

Fig. 2 contrasts the recognition scores from the unquantized recognizer (see section III) with the scores from the 128-code and the 8-code VQ-based DTW cases. These cases are presented as the closest and farthest, respectively, from the unquantized case. The figure shows a decrease in recognition performance as the quantization of the speech changes from continuous to a coarse representation. This experiment confirmed the previously reported superiority of RPS over CEP for unquantized data. However, while 8th order LPC with RPS distance (LPC8-RPS in Fig. 2) gives better performance than 8th order LPC with CEP distance in the "unquantized" case, LPC8-RPS degrades the fastest as quantization becomes coarser.

A possible explanation of the decrease in recognition rates with coarse quantization of the LPC analysis and RPS distance combination is found by considering the codebooks for very coarse (8 code) quantization. When using very small codebooks in speaker independent recognition, each centroid must represent a broad class of speech sounds. We hypothesize that for coarsely quantized speech, a smooth spectral representation is needed. Fig. 3 shows the spectra of the codewords for the coarsest (i.e. 8 code) quantization of 8th

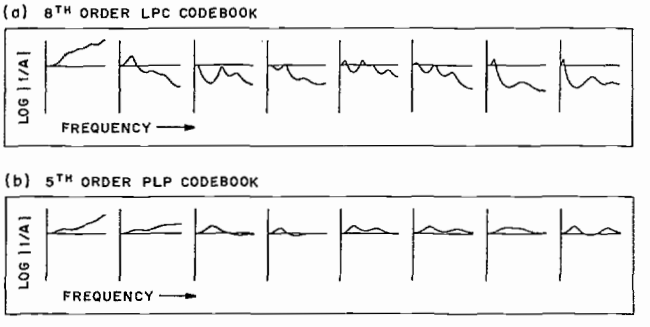

Fig. 3 Log magnitude spectra of the codewords for the (8 code) quantization from RPS distance with (a) 8th order LPC and (b) 5th order PLP. The horizontal axis of each subfigure is frequency from 0 to 5 kHz.

order LPC and 5th order PLP analyses combined with RPS distance. Despite the fact that they are averages of large clusters, codeword spectra for the 8th order LPC case appear to represent specific speech spectra (i.e. have narrow peaks). In contrast, the PLP codeword spectra in Fig. 3 are smooth. Our hypothesis is thus supported since with RPS distance, the error rate of 8th order LPC is twice of that of 5th order PLP.

Another view of the interaction between weighted cepstral distance and vector quantization is shown in Figs. 4(a) and (b). The small dots in this scatter plot show the first two cepstral coefficients (c_1 and c_2) for every frame of the training data from one team of 24 speakers. The big dots indicate the 8-code centroids obtained from 14th order LPC analysis with RPS or CEP distance. These two figures show that the codebook based on the standard CEP distance provides broader and more uniform coverage of the training data in the c_1-c_2 plane than does the RPS-based codebook. This is expected since the CEP distance emphasizes differences in lower cepstral coefficients, while the RPS distance emphasizes differences in higher coefficients. A similar figure (not shown) for the 13th and 14th cepstral coefficients indicates that the RPS-based codebook provides broader coverage of the training data in the c_{13}-c_{14} plane. Thus, RPS-based codebooks emphasize variation in high order cepstral coefficients which express the most rapid variations in the spectra (e.g. spectral peaks). This is desirable in speaker *dependent* recognition where no averaging is done across speakers and the spectral peaks for a particular speech sound are less variable. Since for speaker *independent* recognition we expect a much larger variation in spectral peaks we conjecture that, with a coarsely quantized representation, smoothing of the spectra will account for such variation.

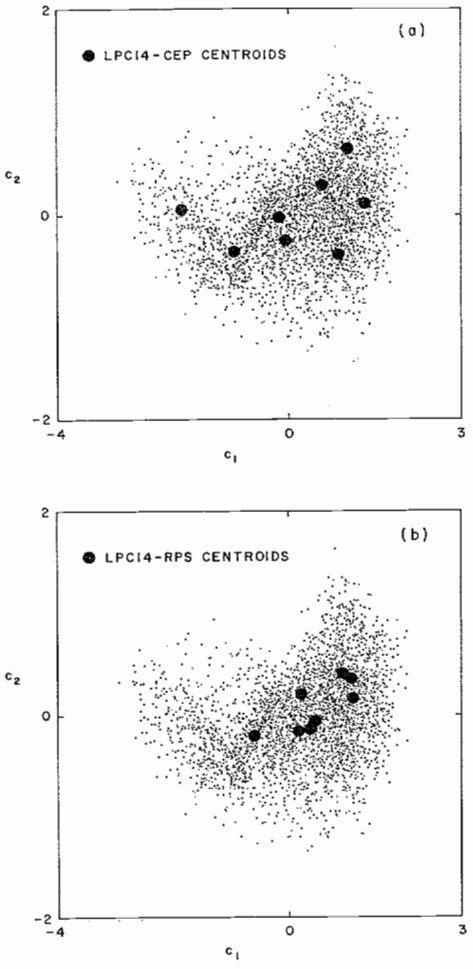

Fig. 4 Scatter plot in the plane of the first two cepstral coefficients. The small dots represent training frames and the big dots represent the centroids from the 8-code codebooks from LPC analysis using (a) cepstral and (b) RPS distances.

C. Efficient Recognition with PLP Analysis and RPS Distance. The hypothesized need for smoothing or averaging of spectra when using small codebooks for speaker independent recognition is supported by the higher recognition rates obtained for PLP analysis. In particular, 5th order PLP combined with RPS maintained good recognition rates for very small (i.e. 16 or 8) codebook size. These codebook sizes are smaller than the number of phonemes in the lexicon, which indicates that low order PLP can represent broad phonetic categories well. The success of this small codebook representation, combined with the nearly three-to-one reduction in the number of coefficients required by 5th order PLP versus 14th order LPC, suggests that an efficient speaker-independent recognition system may be implemented using PLP and weighted cepstral distance.

V. CONCLUSIONS

A representative weighted cepstral distance has been examined for two vector quantization based speaker independent word recognizers. Used with caution, the weighted cepstral distance gives good recognition results. Both the best and the worst recognition results in the reported experiments were obtained with the weighted cepstral distance measure. Our conclusions are summarized below.

1. Weighted cepstral distance interacts with the analysis method. When used with *PLP analysis*, RPS distance gives recognition results as good as, or better than CEP distance. When used with *LPC analysis*, RPS distance gives the worst recognition results of any combination of analysis method and distance measure we investigated.

2. Weighted cepstral distance interacts with vector quantization. When used with *unquantized* low (8th) order LPC analysis data, RPS distance gives better recognition results than CEP distance [2,3]. When used with *vector quantized* low order LPC analysis data, RPS distance gives worse recognition results than CEP distance. Recognition rate for the combination of RPS and LPC decreases rapidly as the quantization is made more coarse.

3. Weighted cepstral distance supports efficient recognition. The combination of RPS distance measure and PLP analysis continues to give good recognition results for very small (16 or 8 code) codebooks, and low analysis order.

ACKNOWLEDGEMENTS

The authors would like to thank Brad Richey and Ben Reaves for their careful work on the speech database.

REFERENCES

[1] Paliwal: Speech Communication, 1, 151-154, 1982.
[2] Tohkura: ICASSP-86, 761-764, 1986.
[3] Juang, et. al.: ICASSP-86, 765-768, 1968.
[4] Hanson, et. al.: ICASSP-86, 757-760, 1968.
[5] Hermansky, et. al.: ICASSP-86, 1971-1974, 1968.
[6] Rabiner, et. al.: BSTJ, 1075-1105, 1983.
[7] Sugamura, et. al.: J. Acoust. Soc. Japan 5, 243-252, 1984.
[8] Markel, et. al.: *Linear Prediction of Speech,* Springer-Verlag, New York, 1976.
[9] Hermansky, et. al.: Speech Comm. 4, 181-187, 1985.
[10] Hermansky: In this ICASSP-87 Proceedings.
[11] Hanson, et. al.: Submitted for publication in ASSP.
[12] Linde, et. al.: IEEE Trans. COM-28, 84-95, 1986.
[13] Sugiyama, et. al.: Trans. IECE Japan J67-A, 1984.
[14] Glinski: AT&T Tech. J., 64, 1033-1045, 1986.
[15] Sakoe, et. al.: IEEE Trans. ASSP-26, 43-49, 1978.