論文 / 著書情報
Article / Book Information

Title	A VQ-Based Preprocessor Using Cepstral Dynamic Features for Large Vocabulary Word Recognition
Author	SADAOKI FURUI
Journal/Book name	IEEE ICASSP1987, Vol. , No. , pp. 1127-1130
発行日 / Issue date	1987, 4
権利情報 / Copyright	(c)1987 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Reprinted from PROCEEDINGS OF ICASSP — International Conference on
Acoustics, Speech, and Signal Processing, Dallas, Texas, April 6-9, 1987

A VQ-BASED PREPROCESSOR USING CEPSTRAL DYNAMIC FEATURES FOR LARGE VOCABULARY WORD RECOGNITION

SADAOKI FURUI

NTT Electrical Communications Laboratories

Musashino-shi, Tokyo, 180 Japan

ABSTRACT

This paper proposes a new VQ (Vector Quantization)-based preprocessor for use in a method which reduces the amount of computation necessary in speaker-independent large vocabulary isolated word recognition. A speech wave is analyzed by time functions of instantaneous cepstrum coefficients and short-time regression coefficients for both cepstrum coefficients and logarithmic energy. A universal VQ codebook for these time functions is constructed based on a multi-speaker, multi-word database. Next, a separate codebook is designed as a subset of the universal codebook for each word in the vocabulary. These word-specific codebooks are used for front-end preprocessing to eliminate word candidates whose distance scores are large. A dynamic time-warping processor based on a word dictionary, in which each word is represented as a time-sequence of the universal codebook elements (SPLIT method), then resolves the choice among the remaining word candidates. Effectiveness of this method has been ascertained by recognition experiments using a database consisting of words from a vocabulary of 100 Japanese city names uttered by 20 male speakers.

I. INTRODUCTION

Recently, a VQ-based approach has been proposed to reduce the amount of computation in large vocabulary isolated word recognition [1]. In this method, word-based VQs are used for preprocessing to eliminate word candidates whose distortion scores are large, with a DTW processor then resolving the choice among the remaining word candidates. The VQ codebooks are constructed by clustering the spectral envelopes of the utterances of each word taking the speaker-variability into consideration. Since each codebook simply represents a set of instantaneous spectra, that is, since there is a lack of transitional spectral information, the ability of the VQ-based processor decreases as the vocabulary grows in size.

A technique for incorporating a temporal structure into the preprocessor has been proposed to alleviate this difficulty [2]. In this method, the probability density function of the time of occurrence for each vector in the codebook is estimated from a study of the training sequences. The spectral distance score of the VQ is combined with a temporal distance score for each frame in the word. Although this method improves performance, it is problematic for two reasons: processing can start only after endpoint detection, and the performance depends on the accuracy of endpoint detection. Furthermore, the method using word-based codebooks has the disadvantage that the amount of distance calculation increases in proportion to the vocabulary size.

This paper proposes a new method for improving the performance of preprocessing. The method is based on the incorporation of dynamic and instantaneous spectral features at every short interval, and on the introduction of a universal codebook which is commonly used as a basic codebook for all words. The usefulness of dynamic features in word recognition and speaker recognition has been ascertained in previous research [3]-[6].

II. FEATURE EXTRACTION AND DATABASE

The speech wave is passed through a low-pass filter having a cutoff frequency of 4 kHz and digitized at an 8-kHz rate. Word endpoints (beginning and ending positions) are detected, and linear predictive coding (LPC) analysis is performed on all frames within the word. The LPC analysis is a tenth-order analysis of 32-ms frames, spaced every 8 ms along the word. Each overlapping 32-ms section of speech is windowed using a Hamming window. The results of the LPC analysis are the set of LPC cepstrum coefficients and logarithmic energy (these will be called simply cepstrum coefficients and energy hereafter).

Time functions of the cepstrum coefficients and the energy over the N frame intervals are extracted every 8 ms, and their dynamic characteristics indicated over the interval are represented by regression coefficients. The speech wave is then represented by a set of cepstrum coefficients and regression coefficients of both cepstrum and energy every 8 ms. The energy value itself is not used since it is insignificant for phoneme perception.

A vocabulary of 100 Japanese city names uttered by 20 male speakers (2000 samples) was used for the test utterances, and the same vocabulary uttered by a different set of four male speakers was used for the training utterances. These four speakers, considered to represent the entire range of male voices, were selected from among 30 male speakers.

III. PREPROCESSING EXPERIMENT USING WORD-BASED CODEBOOKS

3.1. Method

As the first step in the experiments, the effectiveness of feature parameters for preprocessing was evaluated using word-based codebooks constructed independently for each word. As shown in Fig. 1, feature parameter sets (cepstrum coefficients and regression coefficients for both cepstrum and energy) of all speech frames uttered by the four training speakers are clustered for each word to produce a codebook representing the parameter sets of the word. The clustering is done using the method proposed in [7]. Since the parameter sets consist of different kinds of features, each feature parameter is multiplied by appropriate weighting factor obtained in [4] during the distance calculation for clustering. Each input utterance is vector-quantized by codebooks of all vocabulary words, and word candidates having relatively small distortion scores averaged over all input frames are selected.

Two methods were investigated for word candidate selection: Method A, in which a fixed number of word candidates having relatively small distortions are selected, and Method B, in which word candidates having distortions smaller than a previously determined threshold are selected. In the latter case, the threshold, θ, is determined based on either of the equations,

CH2396-0/87/0000-1127 $1.00

$$\theta = \begin{cases} \theta_0 \quad \text{(fixed)} & \text{[Method B-1]} \\ \theta_0 + \min_n\{d_n\} & \text{[Method B-2]}, \end{cases}$$

where d_n is the distortion using the codebook of the n-th word.

3.2. Effectiveness of Feature Parameters in Preprocessing

The first experiment was performed using Method A, under the conditions that the cepstrum (C) and regression coefficients of both cepstrum (C') and energy (E') were all used as feature parameters, and that the codebook size for each word was set at 64. The symbol ' indicates the regression coefficient. Figure 2 shows the preprocessing error rate, that is, the ratio indicating that the true word is excluded from the selected candidates, as a function of the number of frames for regression analysis. The number of frames was varied between 7 and 15 (between 56 and 120 ms). Although the error rate variation as a function of the number of frames is small when the number of frames is less than 13 and the number of candidates is between 2 and 10, the error rate has a minimum at 11 frames when only one candidate is selected.

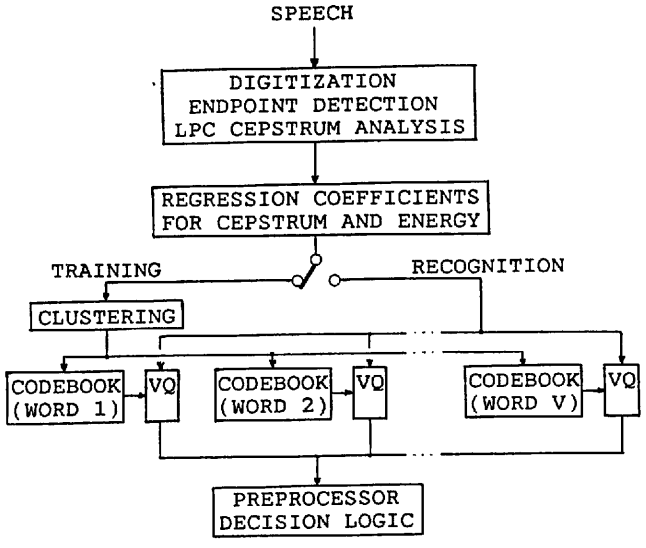

Fig. 1 - Block diagram of a VQ preprocessor based on word-based codebooks.

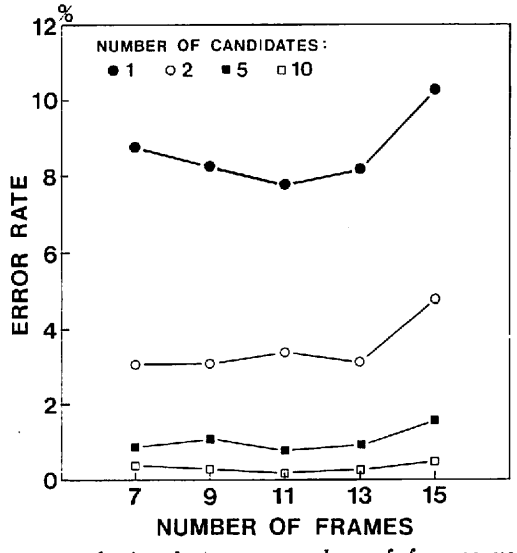

Fig. 2 - Relation between number of frames used for regression analysis and preprocessor error rate.

Preprocessing performances for various parameter combinations were compared in the second experiment under the condition that the number of frames was temporally fixed at 11 based on the results of the first experiment. Figure 3 shows the results for five parameter conditions: (C) and (C') only, and (C,E'), (C',E'), and (C,C',E') combinations. The resulting error rate is shown as a function of the number of selected candidates. The error rate derived using the regression coefficients of the cepstrum is 1/3 or less than that using the cepstrum itself. As is clearly indicated, the error rate is largely reduced by combining the regression coefficients of both cepstrum and energy with the cepstrum. When the (C,C',E') combination is used and 10 candidates are selected, the probability that the true word is included in the candidates is 99.8 %.

The experimental results for Method B-2 are plotted in Fig. 4. The results for three parameter conditions, (C), (C,E'), and (C,C',E'), are indicated by the preprocessing error rate and the ratio of the number of selected words to vocabulary size as a function of the threshold, θ_0. This shows that the error rate and the number of candidates can be greatly reduced by combining the regression coefficients with the cepstrum. When $\theta_0 = 0.2$, the probability that the true word is included in the candidates is 99.8 %, and the number of selected words averages 4.5 out of 100. These results are better than those obtained by Method A. Method B-1 was also examined for comparison. The results indicated that the number of selected words is 9.9, while the percentage of true word selection is only 92.4 % under the optimum threshold condition. This means that Method B-2 is far more effective than Method B-1. Accordingly, Methods A and B-1 were removed from the investigation.

The results for Method B-2, when the codebook size is varied between 8, 16, 32, and 64, are shown in Fig. 5. The difference between the results for 32 and 64 is clearly very small. Therefore, all subsequent experiments used codebooks having a size of 32.

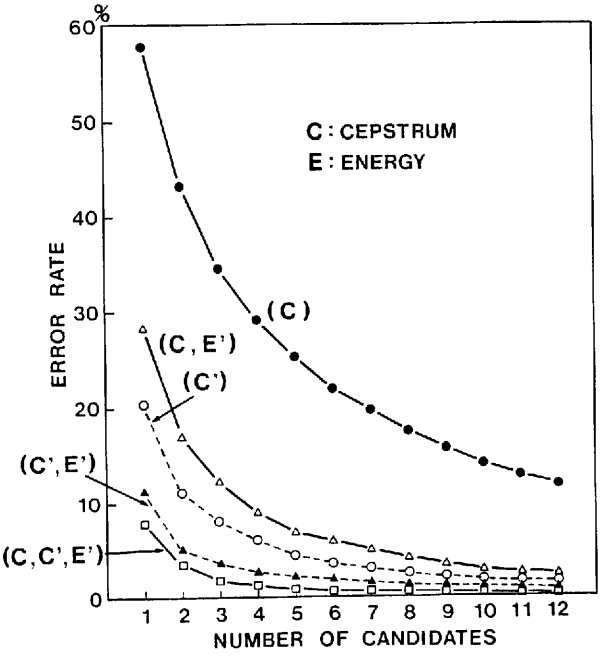

Fig. 3 - Comparison of preprocessor performances for five feature parameter conditions (Method A: Preselection by distance rank order).

IV. PREPROCESSING EXPERIMENT USING UNIVERSAL CODEBOOK

4.1. Method

In the above experiments, codebooks for preprocessing were constructed independently for each word. Although this method can efficiently reduce the number of candidate words, the number of calculations for the total system including preprocessing and postprocessing can be reduced only very slightly. This is because the distance between each input speech frame and each codebook element vector must be calculated for every vocabulary word; hence, the number of distance calculations increases in proportion to the vocabulary size.

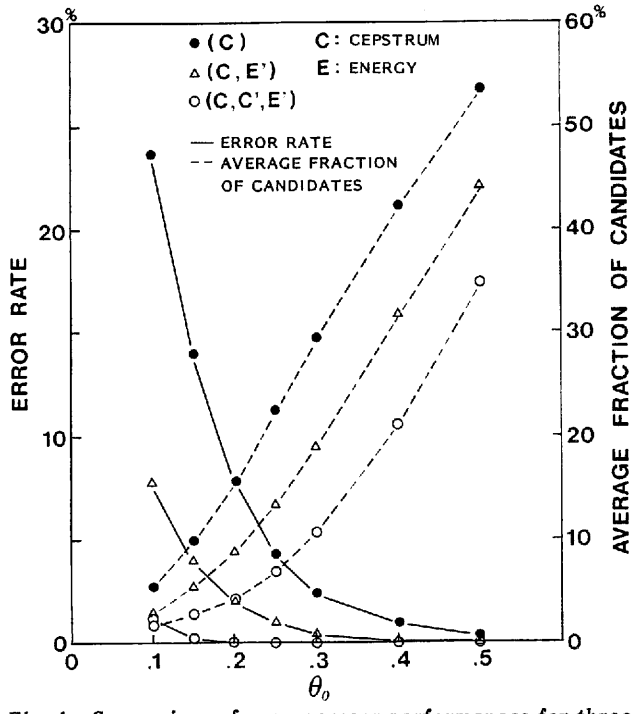

Fig. 4 - Comparison of preprocessor performances for three feature parameter conditions (Method B-2: Preselection by biased distance threshold).

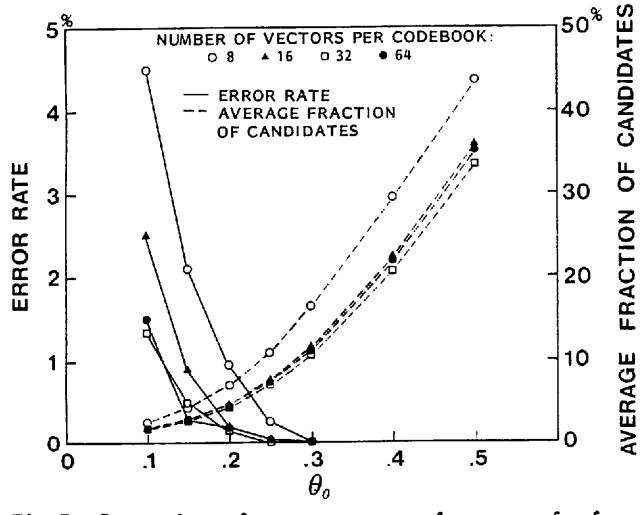

Fig. 5 - Comparison of preprocessor performances for four codebook-size conditions.

In this section, a new method diagrammed in Fig. 6 is investigated, in which a universal codebook is used for all vocabulary words. The universal codebook is constructed by clustering all element vectors of the word-based codebooks, which are in turn constructed for each word as described in the previous section. The codebook of each word for preprocessing is then built by selecting universal codebook elements which closely match the word-based codebook elements. With this method, distances are calculated between each input speech frame and each universal codebook element to construct a distance matrix. Since the codebook for each word is a subset of the universal codebook, the VQ distortion using the codebook of each word can be easily obtained by referring to the distance matrix in the same way as the SPLIT method [8].

The ratio of the number of distance calculations for the present method to that for the previous method is $M / (L \times 100)$, where L is the codebook size for each word, and M is the universal codebook size. For example, when L=32 and M=1024, the ratio is $1024/3200 \simeq 1/3$. Notably, the ratio decreases in inverse proportion to the vocabulary size.

The experimental results given in the previous section indicate that the preprocessing performance variation as a function of the number of frames for regression analysis is small in a range of less than 13 frames. On the other hand, recognition experiments based on the DTW matching method using the combination of cepstrum and regression coefficients for both cepstrum and energy without preprocessing [4] indicate that the optimum length is 7 frames. Based on both sets of results, the number of frames for polynomial expansion was set at 7 in all subsequent experiments.

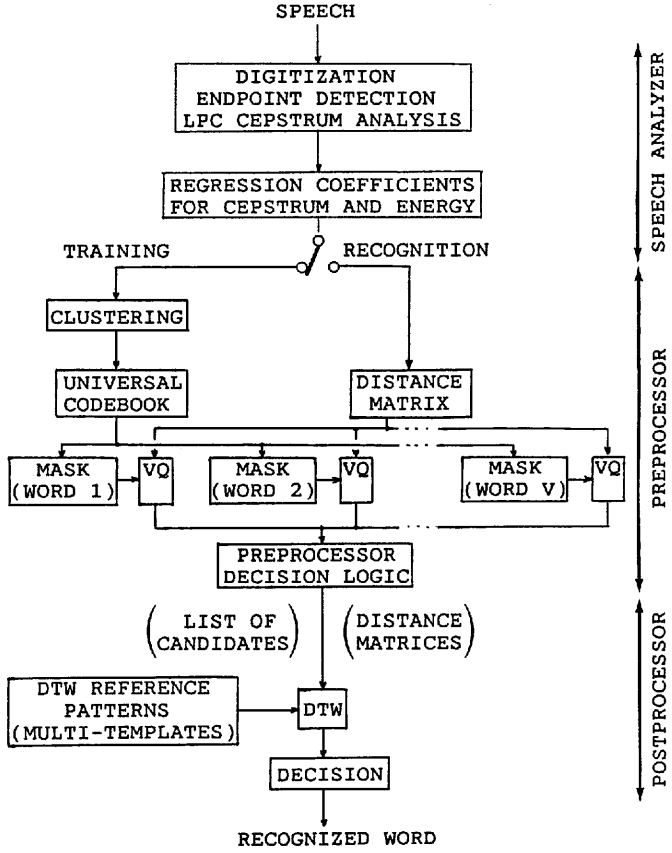

Fig. 6 - Block diagram of isolated word recognizer incorporating a VQ preprocessor based on a universal codebook and a SPLIT postprocessor.

4.2. Experimental Results

Figure 7 shows the results of preprocessing using the universal codebook as well as those using the word-based codebooks. The effect of the universal codebook size was investigated by varying it between 512 and 1024. When the universal codebook size is 1024, a preprocessing performance almost the same as that using word-based codebooks can be obtained. When θ_0 = 0.2, the number of candidate words is 6.8 on an average, and the true word inclusion percentage is 99.9 %. A precise analysis of the experimental results indicates that for a quarter of the input utterances (496 samples), the preprocessor eliminates all word candidates except one. For such cases, there is clearly no further processing required for word recognition.

V. COMBINATION OF PREPROCESSING AND DTW

5.1. Experimental Results

Overall recognition performance was compared between the method based on the combination of preprocessing and DTW processing, and the DTW-only method without preprocessing. DTW was performed based on the SPLIT method using a 1024-element universal codebook. Four reference templates which correspond to four training speakers were represented by the time sequences of universal codebook elements for each word. A distance matrix between input speech and universal codebook elements was used for both preprocessing and DTW. The word whose DTW distance was smaller than any other word was selected as the final decision.

Table 1 shows the recognition results for various experimental conditions. When a universal codebook and Method B-2 are used and θ_0=0.2, DTW is performed for only 6.8 % of the vocabulary words, as described in the previous section. Then, a recognition error rate of 2.0 %, which is the same as that obtained without preprocessing, can be finally achieved. The total amount of calculation necessary in this condition is almost 1/10 of that without preprocessing.

5.2. Discussion of VQ Method

In the above-mentioned experiments, three kinds of feature parameters (C,C',E') were clustered at once as elements of a single vector. On the other hand, in paper [6], cepstrum (C) and regression coefficients (C') were clustered separately and the VQ distortions associated with each parameter set were combined by a weighted summation during speaker verification experiments. An additional experiment was performed to compare the performances achieved by these two methods. The results indicated that the error rate for the separate clustering method is twice as large as the method proposed in this paper. This means that instantaneous and dynamic features convey phonetic information through mutual interaction.

VI. CONCLUSION

A new VQ-based preprocessor for selecting candidate words has been proposed for large vocabulary speaker-independent word recognition. Each word-specific codebook is built as a subset of a universal codebook. A feature parameter vector consists of instantaneous and transitional parameters for both cepstrum coefficients and logarithmic energy. The input utterance is passed through VQ processors using the codebook of each word, and the words whose VQ distortion scores are relatively small are selected as candidate words. A DTW processor based on the SPLIT method then resolves the choice among the word candidates.

Experiments confirmed two principal results. First, the number of candidate words for DTW processing is reduced to 1/15 of the vocabulary while maintaining the same recognition accuracy (98.0 %) as that obtained without the preprocessor. Second, the number of calculations is reduced to almost 1/10 of that of the method without preprocessing. Further investigations will include evaluation experiments using a large vocabulary database.

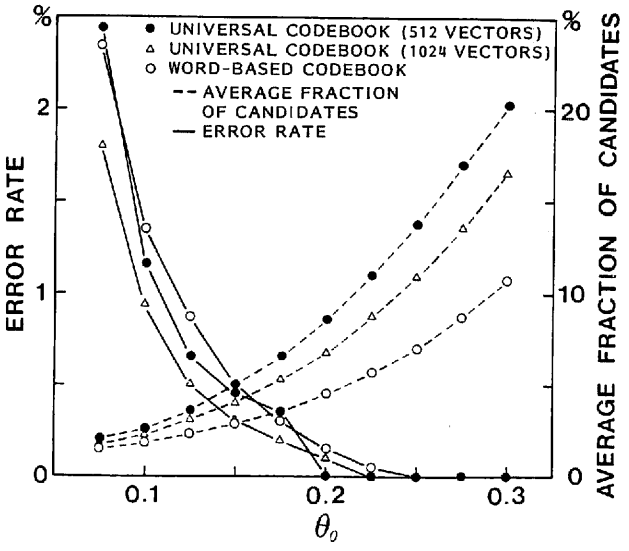

Fig. 7 - Comparison of preprocessor performances for three codebook conditions.

Table 1 - Experimental results for the word recognizer incorporating a VQ preprocessor and a DTW-based postprocessor.

		ERROR RATE (%) PRE-SELECTION ALONE	ERROR RATE (%) PRE-SELECTION & DTW	NUMBER OF CANDIDATES FOR DTW
DTW ALONE		-	2.0	100
WORD-BASED CODEBOOK	PRESELECTION BY BEST ONE	7.9	-	-
	TOP TEN	0.2	2.1	10
	$<\theta$ (θ_0=.2)	0.2	2.1	4.5
UNIVERSAL CODEBOOK	PRESELECTION BY BEST ONE	10.9	-	-
	TOP TEN	0.4	2.3	10
	$<\theta$ (θ_0=.2)	0.1	2.0	6.8

REFERENCES

[1] K-C.Pan,F.K.Soong and L.R.Rabiner, "A vector-quantization-based preprocessor for speaker-independent isolated word recognition," IEEE Trans. Acoust.,Speech,Signal Processing, ASSP-33,3, pp.546-560 (1985)

[2] A.F.Bergh,F.K.Soong and L.R.Rabiner, "Incorporation of temporal structure into a vector-quantization-based preprocessor for speaker-independent,isolated-word recognition," AT&T Tech.J., 64,5,pp.1047-1063 (1985)

[3] S.Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoust.,Speech,Signal Processing, ASSP-29,2,pp.254-272 (1981)

[4] S.Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. Acoust., Speech,Signal Processing,ASSP-34,1,pp.52-59 (1986)

[5] S.Furui, "Speaker-independent isolated word recognition based on emphasized spectral dynamics," Proc.IEEE Int.Conf. Acoust., Speech,Signal Processing,Tokyo,37.10,pp.1991-1994 (1986)

[6] F.K.Soong and A.E.Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," Proc. IEEE Int.Conf.Acoust.,Speech,Signal Processing,Tokyo,17.5, pp.877-880 (1986)

[7] Y.Linde,A.Buzo and R.Gray, "An algorithm for vector quantizer design," IEEE Trans.Commun.,COM-28,1,pp.84-95 (1980)

[8] N.Sugamura,K.Shikano and S.Furui, "Isolated word recognition using phoneme-like templates," Proc.IEEE Int.Conf. Acoust.,Speech, Signal Processing,Boston,16.3,pp.723-726 (1983)