論文 / 著書情報
Article / Book Information

Title	A VQ-Based Preprocessor Using Cepstral Dynamic Features for Speaker-Independent Large Vocabulary Word Recognition
Author	SADAOKI FURUI
Journal/Book name	IEEE Trans. on ASSP, Vol. 36, No. 7, pp. 980-987
発行日 / Issue date	1988,
権利情報 / Copyright	(c)1988 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.

A VQ-Based Preprocessor Using Cepstral Dynamic Features for Speaker-Independent Large Vocabulary Word Recognition

SADAOKI FURUI, MEMBER, IEEE

Abstract—**This paper proposes a new VQ (Vector Quantization)-based preprocessor which reduces the amount of computation in speaker-independent large-vocabulary isolated word recognition. New features introduced are the use of a universal codebook in the VQ-based preprocessor and the use of multiple feature sets including cepstral dynamic features. 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 multispeaker, multiword 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 (DTW) 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. Recognition experiments using a database consisting of words from a vocabulary of 100 Japanese city names uttered by 20 male speakers confirmed the effectiveness of this method. That is, the number of candidate words for DTW processing is reduced to 1/15 of the vocabulary while maintaining the recognition accuracy (98 percent). The total amount of calculation necessary in this condition is almost 1/10 of that without preprocessing.**

I. INTRODUCTION

IT is expected that the need for speaker-independent word recognition for vocabularies as large as thousands of words will increase in the near future. Several approaches have already been proposed for recognizing isolated words, speaker-independently; the most popular and successful systems are based on either the multiple-template model [1] or the HMM (Hidden Markov Model) [2]. The former method has the advantage that the templates of spectral sequences can be easily constructed using a multiple-speaker database based only on distance relationships between training utterances, even if the statistical model of speaker variability is unclear. Additionally, the temporal information of the spectral sequence in each word is maintained precisely in the reference templates, unlike the HMM. However, this method has the disadvantage that the amount of calculation for recognition and memory size for storing the reference templates become large as the vocabulary increases, since each word has multiple templates, and each of them consists of spectral parameters sampled at short intervals.

To cope with this problem, the SPLIT method [3], which is based on vector quantization (VQ) and dynamic time warping (DTW), was proposed. This method is highly effective, since the memory size and the amount of distance calculation depend very little on the vocabulary size. However, even in this method, the calculation for the DTW increases with vocabulary size, and it cannot be neglected compared to distance calculation when the vocabulary is very large.

Recently, a new approach was proposed to solve this problem, in which the word-based VQ's [4] are used for preprocessing to eliminate word candidates whose distortion scores are large; a DTW processor then resolves the choice among the remaining word candidates [5]. The VQ codebooks are constructed by clustering the spectral envelopes of the utterances of each word taking speaker variability into consideration. However, since each codebook simply represents a set of instantaneous spectra, and contains no transitional spectral information, the effectiveness 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 [6]. 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 end-point detection, and the performance depends on the accuracy of end-point 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 sampled at short intervals, and on the introduction of a universal codebook which is used as a basic codebook for all words. The usefulness of dynamic features in word recognition and speaker recognition has been

Manuscript received July 29, 1987; revised December 30, 1987.

The author is with NTT Human Interface Laboratories, 3-9-11 Midori-cho, Musashino-shi, Tokyo, 180 Japan.

IEEE Log Number 8821154.

ascertained in previous research [7]–[9], and the usefulness of VQ, which includes dynamic features, has also been investigated [10].

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. The beginning position for each word is detected based on short-time energy, and linear predictive coding (LPC) analysis is performed on all frames within the word. The LPC analysis is a 10-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 raw energy value is not used because 1) the energy value cannot be used without normalization, since only the relative energy value is significant for phoneme perception; and 2) although the normalized energy is found to be a beneficial parameter [11], [12], normalization needs the maximum energy value, which can only be determined after the word ending position is detected. On the other hand, normalization is not necessary for the energy regression coefficients.

After regression analysis, the frame interval is converted from 8 to 16 ms by averaging the time functions of adjacent frames for both cepstrum and regression coefficients, in order to reduce the number of calculations at the preprocessing and DTW processing stages. The word ending position is finally detected based on the energy value.

One-hundred Japanese city names uttered by 20 male speakers (2000 samples) were used as the test vocabulary, and the same vocabulary uttered by a different set of four male speakers was used for training. These 4 speakers, considered to represent the entire range of males voices, were selected from among 30 male speakers.

III. VQ-Based Preprocessing Method Using Multiple Feature Sets

Feature parameter sets (cepstrum coefficients and regression coefficients for both cepstrum and energy) are concatenated into one long source vector for each speech frame, and the vectors of all 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 LBG-algorithm proposed in [13]. Since the parameter sets consist of different kinds of features, each feature parameter is multiplied by the appropriate weighting factor obtained in [8] during the distance calculation for clustering, as follows.

$$d(x, y) = \sum_{i=1}^{p} (C_i^x - C_i^y)^2 + w_1 \sum_{i=1}^{p} (C_i'^x - C_i'^y)^2 + w_2 (E'^x - E'^y)^2, \quad (1)$$

where x and y indicate individual frames, w_1 and w_2 represent weighting factors, and C_i, C_i', and E' represent cepstrum, cepstrum regression, and energy regression coefficients, respectively. The symbol $'$ indicates the regression coefficient. The order of LPC analysis p is 10, as described in the previous section.

Each input utterance is vector quantized using the codebook of each vocabulary word, 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 all word candidates having distortions smaller than a previously determined threshold are selected. In the latter case, the threshold θ is determined by either

$$\theta = \theta_0 \text{ (fixed)} \qquad \text{[Method B-1]}$$
$$\theta = \theta_0 + \text{Min}\{d_n\} \qquad \text{[Method B-2]}, \quad (2)$$

where d_n is the distortion using the codebook of the nth word. The threshold θ is fixed for all input utterances with Method B-1, whereas it varies according to the minimum distortion for each input utterance with Method B-2.

IV. Preprocessing Experiment Using Word-Based Codebooks

A. Method

As the first step, the effectiveness of feature parameters for preprocessing was evaluated using word-based codebooks constructed independently for each word. Fig. 1 shows a block diagram of the VQ processor [14].

B. Effectiveness of Feature Parameters in Preprocessing

The first experiment was performed using Method A. The cepstrum (C) and regression coefficients of both cepstrum (C') and energy (E') were used as feature parameters, and the codebook size for each word was set at 64. The preprocessing error rate, that is, the probability that the true word is not selected, as a function of the number of frames used for regression analysis is shown in Fig. 2. 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.

Preprocessing performances for various parameter

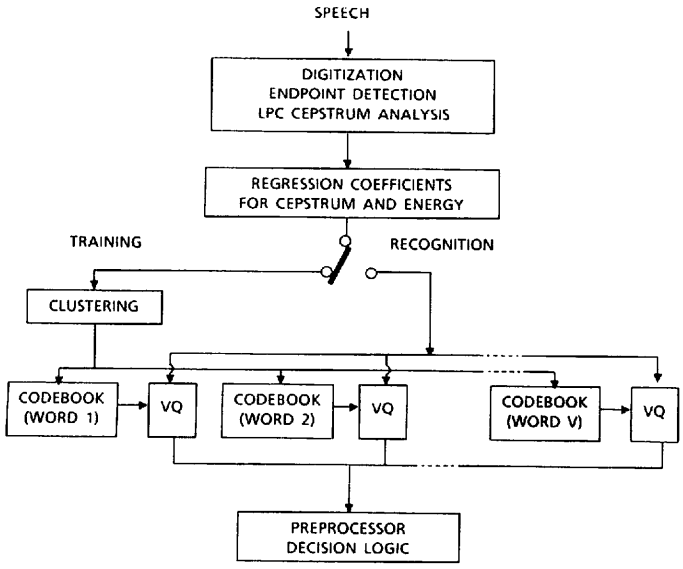


Fig. 1. Block diagram of a VQ preprocessor based on word-based codebooks.

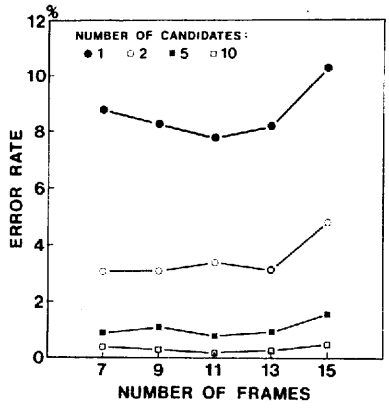


Fig. 2. Relation between number of frames used for regression analysis and preprocessor error rate (Method A: preselection by distance rank order).

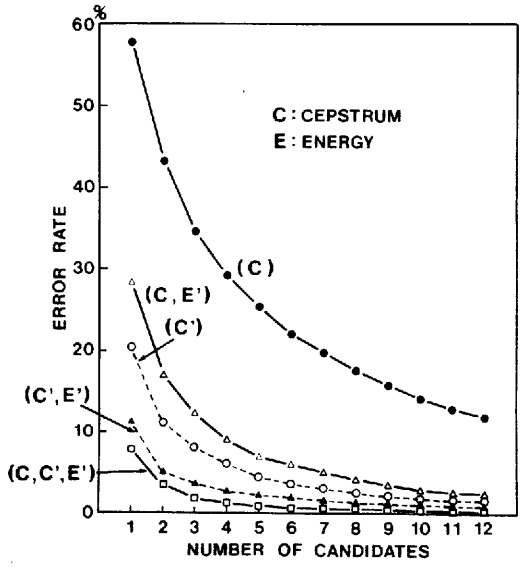


Fig. 3. Comparison of preprocessor performance for five feature parameter conditions (Method A: preselection by distance rank order).

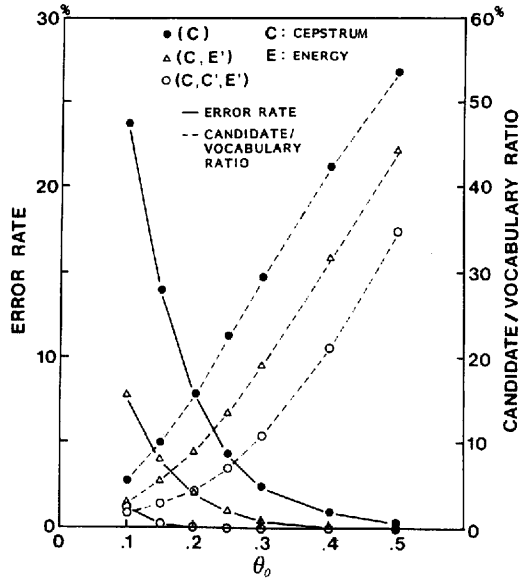


Fig. 4. Comparison of preprocessor performance for three feature parameter conditions (Method B-2: preselection by biased distance threshold).

combinations were compared in the second experiment, in which the number of frames was temporally fixed at 11 based on the results of the first experiment. The resulting error rate is shown in Fig. 3 as a function of the number of selected candidates for five combinations of parameters: (C) and (C') only, (C, E'), (C', E'), and (C, C', E'). 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 percent.

The experimental results for Method B-2 are plotted in Fig. 4. For three parameter conditions (C), (C, E'), and (C, C', E'), the preprocessing error rate and the ratio of the number of selected words to vocabulary size as a function of the threshold, θ_0, are indicated. This shows that the error rate and the number of candidates are 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 percent, and the number of selected words averages 4.5 out of 100. These results are better than those obtained by Method A. From this position on, C, C', and E' are all being used as recognition parameters, unless otherwise stated.

Fig. 5 shows the results for Method B-1. When $\theta_0 = 1.2$, which is nearly the optimum condition of Method B-1, the number of selected words is 9.9, while the percentage of true word selection is only 92.4 percent. By comparing these results to Fig. 4, it can be concluded that Method B-2 is far more effective than Method B-1. Based on these results, Method B-1 was removed from the investigation thereafter.

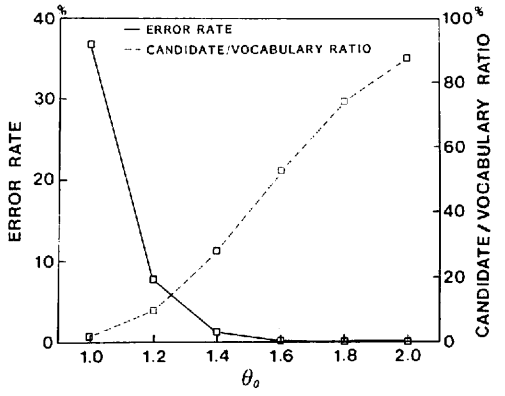


Fig. 5. Preprocessor performance for Method B-1: preselection by unbiased distance threshold.

The results of varying the codebook size in Method A are shown in Fig. 6. Fig. 7 presents the results for Method B-2. These results clearly show that the difference between the results for codebook sizes of 32 and 64 is clearly very small. Therefore, all subsequent experiments used codebooks having a size of 32. By comparing the results shown in Fig. 6(a) and (b), it can be concluded that, if the codebook size is larger than 16 and the number of candidates is larger than 2, the preprocessing performance with 7-frame regression analysis is almost the same as that with 11-frame analysis.

V. Preprocessing Experiment Using a Universal Codebook

A. 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.

In this section, a new method, diagrammed in Fig. 8, 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, as shown in Fig. 9. Since the codebook for each word is a subset of the universal codebook, the VQ distortion using the codebook of each word is easily obtained by referring to the distance matrix in the same way as the SPLIT method [3].

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

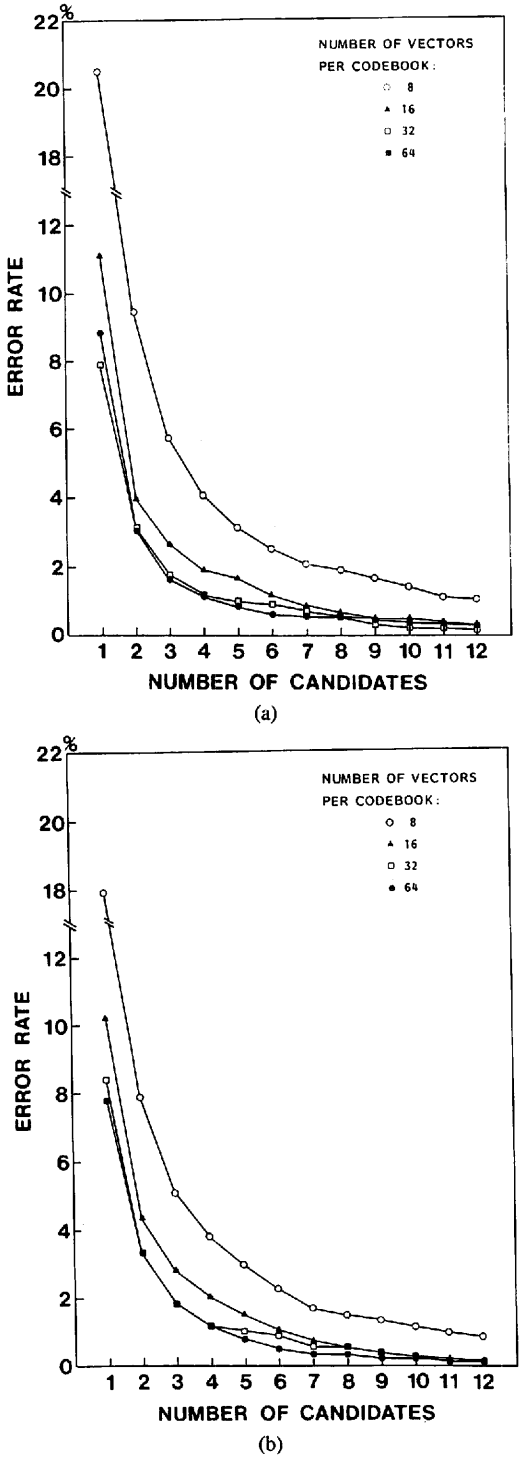


Fig. 6. Comparison of preprocessor performance for four codebook sizes; (a) 7 frames used for regression analysis, and (b) 11 frames (Method A: preselection by distance rank order).

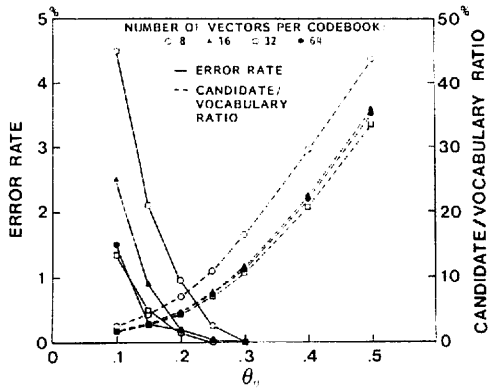


Fig. 7. Comparison of preprocessor performance for four codebook sizes (Method B-2: preselection by biased distance threshold; 7 frames used for regression analysis).

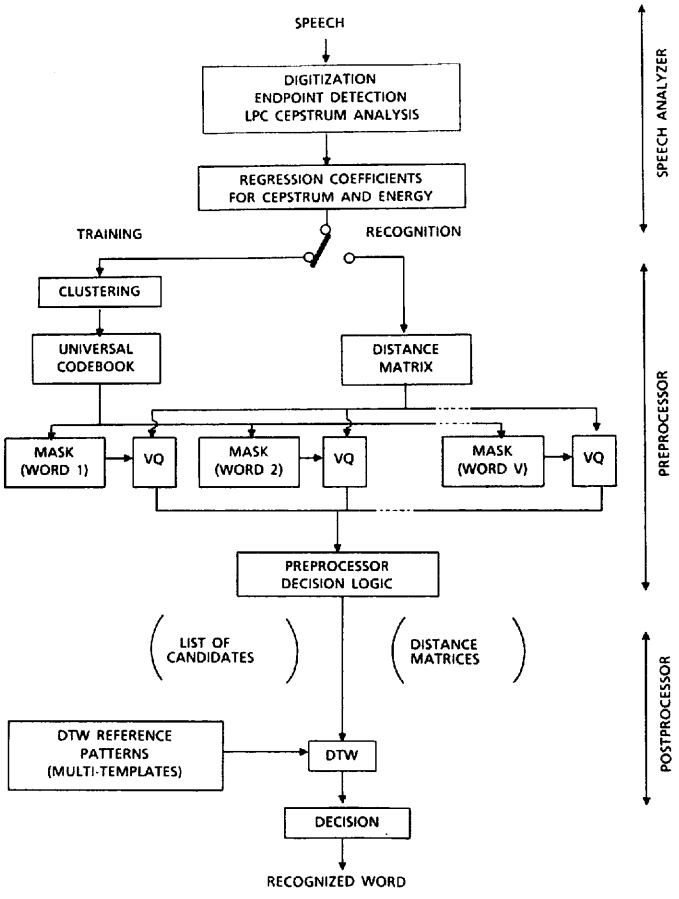


Fig. 8. Block diagram of isolated word recognizer incorporating a VQ preprocessor based on a universal codebook and a SPLIT postprocessor.

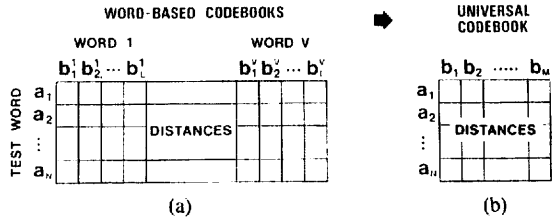


Fig. 9. Distance computation in the VQ preprocessor.

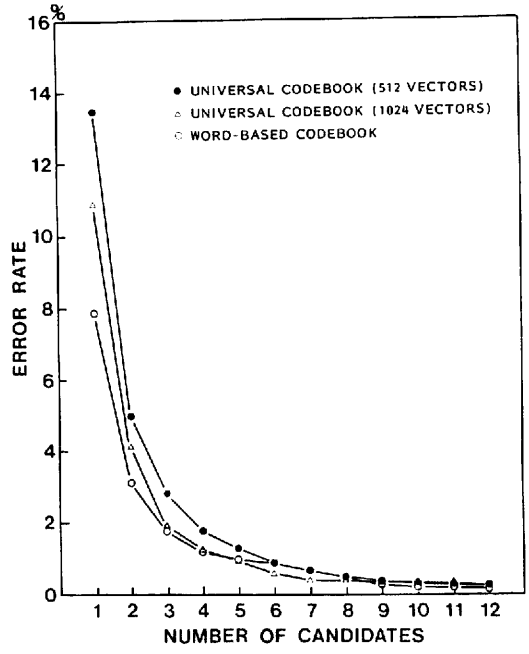


Fig. 10. Comparison of preprocessor performance for three codebook conditions (Method A: preselection by distance rank order).

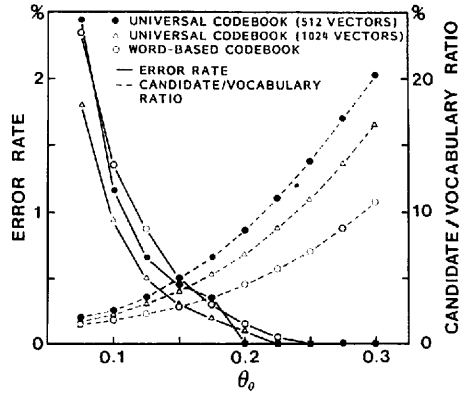


Fig. 11. Comparison of preprocessor performance for three codebook conditions (Method B-2: preselection by biased distance threshold).

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 for 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 [8] 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.

B. Experimental Results

The results of preprocessing using the universal codebook and word-based codebooks for Methods A and B-2, respectively, are shown in Figs. 10 and 11. The effect of the universal codebook size was investigated by varying it between 512 and 1024. When it is 1024, a preprocessing performance almost equal to that using word-based codebooks can be obtained. When it is 1024, and 10 candidates are selected using Method A, the probability that

the true word is included in the candidates is 99.6 percent. When Method B-2 is used and $\theta_0 = 0.2$, the number of candidate words is 6.8 on the average, and the true word inclusion percentage is 99.9 percent. A precise analysis of the experimental results indicates that for 1/4 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.

VI. Combination of Preprocessing and DTW

The overall recognition performance of the method combining preprocessing and DTW processing was compared to that of the DTW method only. DTW was performed between the input utterance, which is the time sequence of the feature parameter set, and multiple reference templates of each candidate word 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 the universal codebook elements for each word. A distance matrix between input speech and universal codebook elements was used for both preprocessing and DTW. The staggered array method [8] was used for DP matching in DTW. The word whose DTW distance was smaller than any other word was selected as the final decision.

The recognition results for various experimental conditions are shown in Table I. When a universal codebook and Method B-2 are used and $\theta_0 = 0.2$, DTW is performed for only 6.8 percent, that is, 1/15 of the vocabulary words, as described in the previous section. Then, a recognition error rate of 2.0 percent, which is the same as that obtained without preprocessing, can be finally achieved. The total amount of calculation necessary in this condition is about 1/10 of that without preprocessing.

VII. Discussion of VQ Method

In these experiments, three kinds of feature parameters (C, C', E') were clustered as elements of a single vector. In other words, the parameter vectors were concatenated into one long source vector; and a single codebook, rather than several independent codebooks, was built. On the other hand, in previous work [10], 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.

Although advantage of the joint clustering method in pattern recognition over the separate clustering method is easily justified by information theory, additional experiments were conducted to compare the actual recognition performances achieved by these two methods. In additional experiments, separate clustering was performed, and the following distance measures were used instead of (1):

$$d_1 = d(C') + \alpha d(C, E')$$
$$d_2 = d(C) + \beta d(C', E'), \quad (3)$$

TABLE I
Experimental Results for a Word Recognizer Incorporating a VQ Preprocessor and a DTW-Based Postprocessor

		Error Rate (percent)		
		Preselection Alone	Preselection 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
	$<\theta$ ($\theta_0 = 0.2$)	0.2	2.1	4.5
Universal Codebook	Preselection By:			
	Best One	10.9	—	—
	Top Ten	0.4	2.3	10
	$<\theta$ ($\theta_0 = 0.2$)	0.1	2.0	6.8

where α and β are combination factors. Using d_1, C and E' are clustered jointly and C' is clustered separately. On the other hand, using d_2, C' and E' are clustered jointly and C is clustered separately. Recognition experiments were performed by varying these combination factors. All codebook sizes were fixed at 32.

The error rates for the condition that only one candidate is selected are presented in Fig. 12. These results indicate that the error rate for the separate clustering method is roughly twice that of the method proposed in this paper, even if the combination factors are set at optimum values.

Statistical distributions of VQ distortion scores for four preprocessing methods are shown in Table II. Mean values (μ) and standard deviations (σ) are presented separately for two conditions: when input speech and codebook are the same word (within-word condition) and when they are different words (between-word condition). Suffixes w and b indicate within- and between-word conditions, respectively. All these values are normalized by the mean value of within-word distortion distribution for each method. The following evaluation value is also shown in this table for each preprocessing method:

$$\zeta = \frac{\mu_b - \mu_w}{(\sigma_b^2 + \sigma_w^2)^{1/2}}. \quad (4)$$

It has been reported that this value is correlated with the recognition accuracy [15].

The results presented in Table II show that the between-word distortion distribution varies depending on the clustering methods. Distribution of between-word distortions produced by the separate clustering method ($d(C') + \alpha d(C, E')$) is closer to the within-word distortion distribution than the between-word distortion distribution produced by the joint clustering method ($d(C, C', E')$).

These results indicate that instantaneous and dynamic features convey phonetic information through mutual interaction. In other words, spectral information in each word is more accurately characterized by adding transitional information to each instantaneous spectrum, than

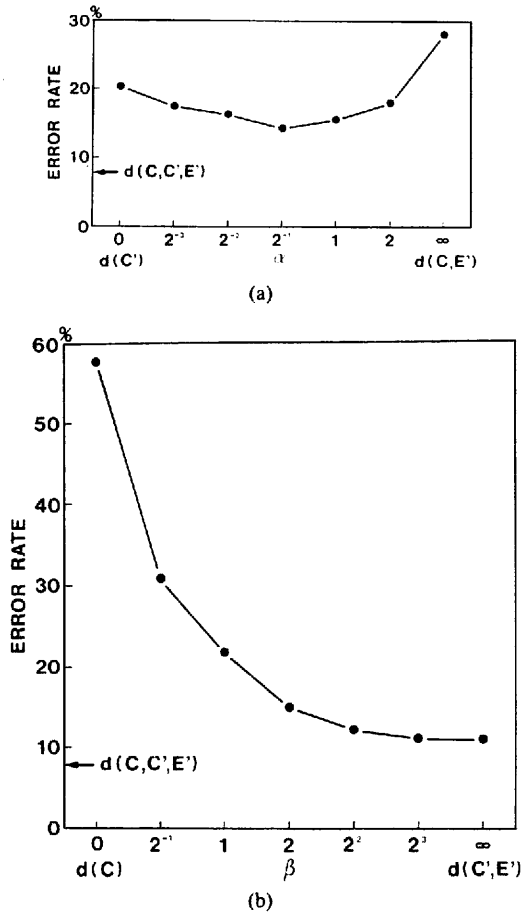


Fig. 12. Preprocessor performance versus factors α and β combining two distortions associated with separate VQ codebooks; (a) combination of $d(C')$ and $d(C, E')$, and (b) combination of $d(C)$ and $d(C', E')$.

TABLE II
STATISTICAL DISTRIBUTIONS OF WITHIN- AND BETWEEN-WORD VQ DISTORTION SCORES FOR VARIOUS PREPROCESSING METHODS

	Within-Word		Between-Word		
	μ_w	σ_w	μ_b	σ_b	ζ
$d(C, C', E')$	1.00	0.19	1.88	0.42	1.88
$d(C') + \alpha d(C, E')$	1.00	0.19	1.68	0.39	1.55
$d(C')$	1.00	0.21	1.58	0.38	1.34
$d(C, E')$	1.00	0.22	1.91	0.62	1.38

$\alpha = 2^{-1}$

by simply combining statistical distributions of instantaneous and transitional spectra.

Although the results are not positive, the separate VQ method may be advantageous over the joint VQ method when the number of training samples is restricted. This is because the mutual interaction between two separate features cannot be easily estimated using a small number of samples. Analogously, because of the difficulty of estimating cross-terms, the weighted Euclidian distance measure is frequently substituted for the Mahalanobis (covariance weighted) distance measure in pattern recognition by setting the off-diagonal terms of the covariance matrix at zero.

VIII. Conclusion

A new VQ-based preprocessor for selecting candidate words has been proposed for large vocabulary speaker-independent word recognition. In this method, each word is characterized by a VQ codebook which is constructed by clustering the feature parameter vector sets over all frames of the word. Each cluster center is then replaced by an element of a universal codebook. Therefore, 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 transitional information is represented by regression coefficients at every short period of roughly 50-ms long. Optimum size is 32 for the codebook for each word and 1024 for the universal codebook. 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. In the SPLIT method, each word is represented in the word dictionary as multiple time sequences of the universal codebook elements.

The above scheme was evaluated using a database consisting of words from a vocabulary of 100 Japanese city names uttered by 20 male speakers. 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 percent) as that obtained without the preprocessor. Second, the amount of calculation is reduced to almost 1/10 that of the method without preprocessing.

Several previous papers reported that speech utterances are extended and compressed mainly at steady periods in each word or sentence [16]. If this is true, this preprocessing method seems to be robust for variation in speaking rate, since the variation does not increase the VQ distortion in the preprocessing. Further investigations will include evaluation experiments using a large vocabulary database.

Acknowledgment

The author wishes to thank the two anonymous reviewers whose constructive suggestions have improved the quality of this paper.

References

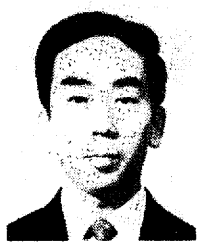
[1] L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker-independent recognition of isolated words using clustering techniques," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 336–349, 1979.

[2] 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," *Bell Syst. Tech. J.*, vol. 62, no. 4, pp. 1075–1105, 1983.

[3] N. Sugamura, K. Shikano, and S. Furui, "Isolated word recognition using phoneme-like templates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Boston, MA, 1983, pp. 723–726.

[4] J. E. Shore and D. K. Burton, "Discrete utterance speech recognition without time alignment," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 473–491, 1983.

[5] 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*, vol. ASSP-33, pp. 546-560, 1985.
[6] 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.*, vol. 64, no. 5, pp. 1047-1063, 1985.
[7] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 254-272, 1981.
[8] ——, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 52-59, 1986.
[9] ——, "Speaker-independent isolated word recognition based on emphasized spectral dynamics," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, Japan, 1986, pp. 1991-1994.
[10] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, Japan, 1986, pp. 877-880.
[11] L. R. Rabiner, M. M. Sondhi, and S. E. Levinson, "A vector quantizer combining energy and LPC parameters and its application to isolated word recognition," *Bell Syst. Tech. J.*, vol. 63, no. 5, pp. 721-735, 1984.
[12] K. Aikawa and K. Shikano, "Spoken word recognition using vector quantization in power-spectrum vector space," *Trans. Inst. Electron. Commun. Eng. Jap.*, vol. J68-D, no. 3, pp. 316-322, 1985.
[13] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84-95, 1980.
[14] S. Furui, "A VQ-based preprocessor using cepstral dynamic features for large vocabulary word recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Dallas, TX, 1987, pp. 1127-1130.
[15] A. E. Rosenberg, "Recognition error measurements from parameterized distance distribution," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tampa, FL, 1985, pp. 870-873.
[16] S. Saito, "Fundamental research on transmission quality of Japanese phonemes," Ph.D. dissertation, Nagoya Univ., 1961.



Sadaoki Furui (M'79) was born in Tokyo, Japan, on September 9, 1945. He received the B.S., M.S., and Ph.D. degrees in mathematical engineering and instrumentation physics from Tokyo University, Tokyo, Japan, in 1968, 1970, and 1978, respectively.

After joining the Electrical Communications Laboratories, Nippon Telegraph and Telephone Corporation in 1970, he studied the analysis of speaker characterizing information in the speech wave, its application to speaker recognition as well as interspeaker normalization in speech recognition, and the vector-quantization-based speech recognition algorithm. He is currently Head of the Speech and Hearing Research Department at NTT Human Interface Laboratories. From December 1978 to December 1979 he was with the Staff of the Acoustics Research Department at Bell Laboratories, Murray Hill, NJ, as a Visiting Researcher working on speaker verification.

Dr. Furui is a member of the Acoustics, Speech, and Signal Processing Society Technical Committee on Speech, the Acoustical Society of Japan, and the Institute of Electronics, Information, and Communication Engineers of Japan.