Designing a Reduced Feature-Vector Set for Speech Recognition by Using KL/GPD Competitive Training

Tsuneo NITTA and Akinori KAWAMURA

Multimedia Engineering Laboratory, TOSHIBA CORPORATION
70 Yanagi-cho, Saiwai-ku, Kawasaki 210 JAPAN
E-mail: nitta@sp.mmlab.toshiba.co.jp

ABSTRACT

The hybrid algorithm of SMQ (Statistical Matrix Quantization) and HMM shows high performance in vocabulary-unspecific, speaker-independent speech recognition, however, it needs lots of computation and memory at the stage of the segment quantizer of SMQ. In this paper, we propose a newly developed, two-stage segment quantizer with a feature extractor based on KL expansion and a classifier, that can be trained by using competitive training of KL/GPD. Result of experiments shows 1/30 - 1/40 reduction in both computation time and a memory size with the same performance that the old version of SMQ shows.

1. INTRODUCTION

We developed a speaker independent, large vocabulary word recognition unit based on the hybrid algorithm of SMQ and HMM and applied it to various types of application with multimodal user interface [1]. On the other hand, contemporary personal computers provide sufficient computing power to accommodate speech recognition with no additional hardware. SMQ/HMM shows high performance in vocabulary-unspecific, speaker-independent speech recognition, however, it needs lots of computation and memory at the segment quantizer of SMQ. In this paper, we propose a design methodologies of a reduced feature-vector set for the segment quantizer by using KL/GPD (Generalized Probabilistic Descent) competitive training.

The new segment quantizer is composed of two stages: a feature extractor based on KL expansion and a classifier. We had adopted an MCE (Minimum Classification Error)/ ALSM (Averaged Learning Subspace Method) algorithm for the segment quantizer of SMQ. This algorithm was proposed by Maeda (1980)[2] and Oja (1982)[3] independently. The new segment quantizer uses an MCE/GPD algorithm for the feature extractor and the classifier. This algorithm was proposed by Amari first (1967)[4] and extended by Katagiri and Juang (1991)[5].

An application of MCE/GPD to designing a feature extractor was also proposed by Watanabe et al. (1995)[6] in which they adopted Jacobi-rotation matrices. In this paper, we propose a simple and efficient algorithm for training the feature extractor in which an orthogonalized feature-vector set is directly modified by training data and the resultant feature-vector obliques each other but gives better classification performance.

This paper is organized as follows. Section 2 provides an overview of the SMQ/HMM-based system and section 3 explains high-speed SMQ based on KL/GPD. Finally, section 4 presents experimental results.

2. SMQ/HMM-BASED SPEECH RECOGNITION

The recognition system based on SMQ/HMM is composed of two stages: a phonetic decoding stage, which performs statistical matrix quantization (SMQ), and an HMM-based word recognition stage. When designing the segment codebook of SMQ, we adopted the algorithm of MCE/ALSM[7] .

2.1 Overview of the System

Figure 1 shows a block diagram of the speaker independent, large vocabulary speech recognition system based on SMQ/HMM. After converting to LPC mel-cepstrum, an input speech is transferred into a phonetic segment sequence by the segment quantizer of SMQ. We adopted 652 phonetic segments including various types of context, V, CV, VCV, etc. In the word recognizer, 235 diphone models were used for sub-word discrete HMMs. A diphone model has 3 loops and 4 states, and each state includes 652 output probabilities of phonetic segments and a transition probability.

2.2 MCE/ALSM Algorithm

The MCE/ALSM algorithm is given as follows.

STEP-1: A correlation matrix R_k is calculated from a training set $\{f_p\}$ of class k.

$$R_k = \sum_{p=1}^{P} f_p \, f_p^{\,T} \qquad (1)$$

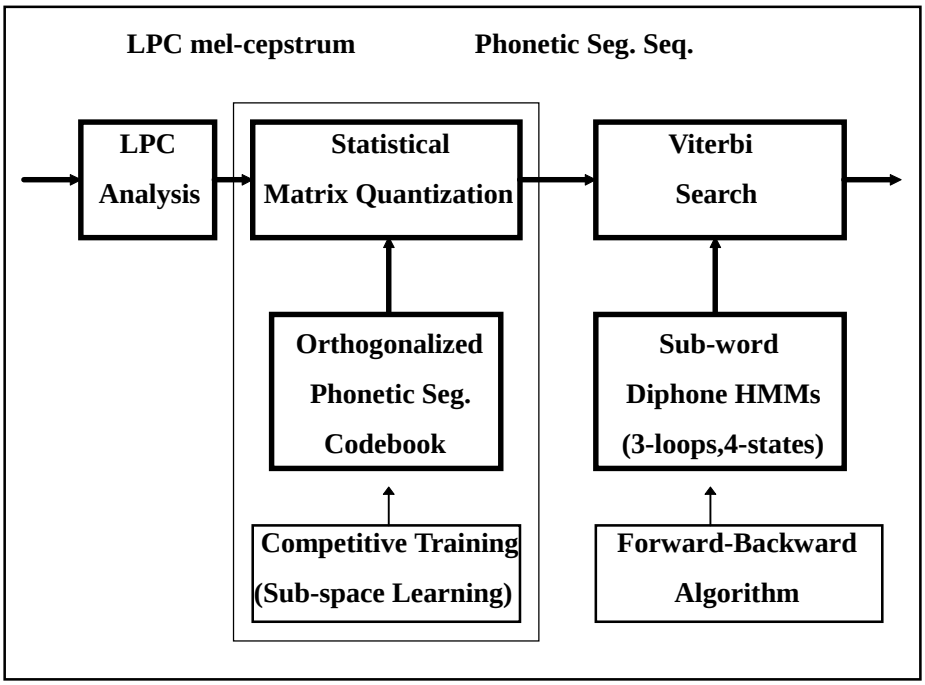

Fig. 1 Block Diagram of a Speech Recognition System Based on SMQ/HMM

Here, P is the number of training pattern and T denotes transposition. Next, an eigen vector set ϕ_{km} is derived from R_k by using KL transform.

STEP-2: A training pattern x is recognized by using the following equation:

$$S_k(x;\Lambda) = \sum_{m=1}^{M} (x \cdot \phi_{km})^2 \qquad (2)$$

where Λ is all the reference vector sets, M is the number of eigen vectors, and ($\cdot$) denotes inner product. The class C_i that x belongs to is decided by using the following rule.

$$C(x) = C_i \quad \text{if} \quad i=\underset{j}{\text{argmax}}\, S_j(x;\Lambda) \qquad (3)$$

STEP-3: The correlation matrix R_k is updated using the training pattern set $\{g_q\}$ of the same class k which was misrecognized to different classes, and the training pattern set $\{h_r\}$ of different classes which was misrecognized to the class k as follows:

$$R_k \leftarrow R_k + \sum_{q=1}^{Q} \mu\, g_q\, g_q^T - \sum_{r=1}^{R} \nu\, h_r\, h_r^T \qquad (4)$$

where μ and ν are coefficients, and Q and R are the number of training patterns in $\{ g_q \}$ and $\{ h_r \}$ respectively.

STEP-4: A new eigen vector set ($\phi_{km)}$ is derived from the updated correlation matrix ($R_{k)}$ by using KL transform.

STEP-5: STEP 2, 3, and 4 are iterated until the recognition accuracy no longer increases.

3. HIGH-SPEED SMQ BASED ON KL/GPD

Computation in SMQ needs about 0.5 million multiply and add operation every 8 msec (= 60 MIPS) and 500 kB codebook memory. To execute this operation on a speech recognition board, we had developed a phonetic segment engine chip. Reduction of computation time and memory in SMQ needs not only combining phonetic segments but also a new quantization algorithm.

3.1 Combination of Phonetic Segments

652 phonetic segments are combined into 168 phonetic classes (compression ratio = 4:1). The reduced classes consist mainly of Cv, V, v1v2, vc, etc.

3.2 SMQ Based on KL/GPD Algorithm

The new segment quantizer is composed from a feature-vector extractor based on KL expansion and a phonetic segment classifier as shown in figure 2. The feature extractor compresses an input LPC mel-cepstrum pattern x with the dimension of 192 (16 mel-cepstrum $\times$ 12 points) into a feature-vector y with the dimension of 24 - 64 every 8 msec. The orthogonalized phonetic segment vector set is designed from a data set including all the 652 phonetic segments by using KL transform.

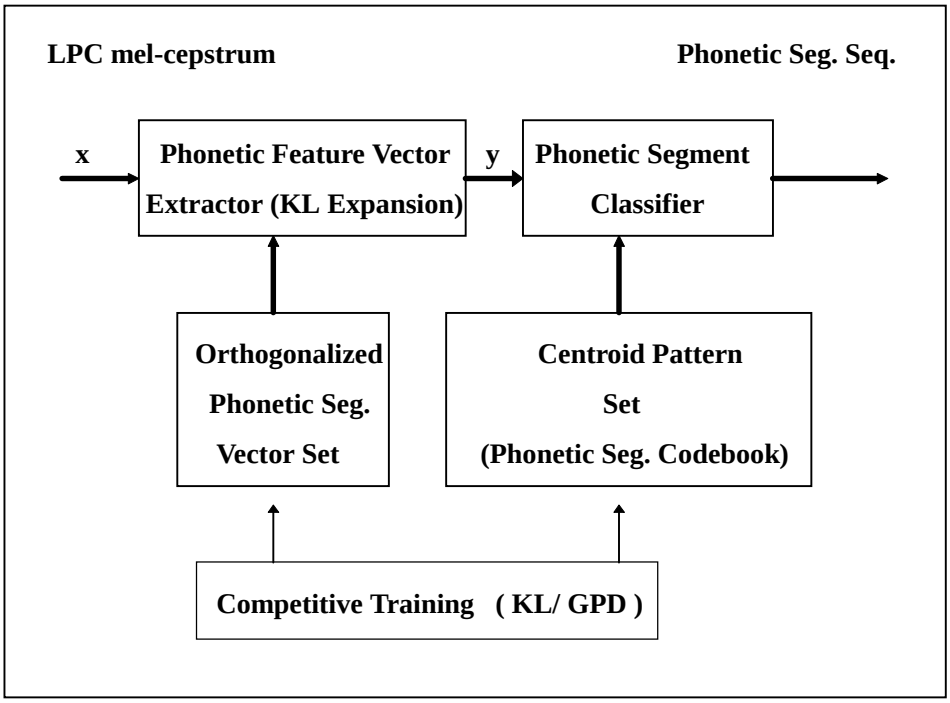

Fig. 2 Configuration of High-speed SMQ Based on KL/GPD

In the classifier, the feature-vector y is compared with 116 centroid patterns. In this paper, we use Euclidean distance as the classification measure. The MCE/GPD algorithm for training the classifier is given as follows:

$$g_k(y;\Lambda) = \| y - r_k \|^2 \quad (5)$$

$$y = W x \quad (6)$$

where $g_k(y;\Lambda)$ is the distance between the feature-vector y and the reference vector set Λ, and r_k is the reference vector of class k. W (W: w_1, w_2, …, w_m) is a set of feature extraction matrix and is given by KL expansion. The decision rule is as follows:

$$C(x)=C_i \quad \text{if} \quad i=\underset{j}{\text{argmin}}\, g_j(y;\Lambda) \quad (7)$$

where C_i is the class that x belongs to.

Here, we define the measure of classification error ($d_k(y)$) and the loss function ($L_k(y)$) caused by misclassification as follows:

$$d_k(y)=g_k(y;\Lambda) - [\{1/(M-1)\} \sum_{j\neq k} g_j(y;\Lambda)^{-\eta}]^{-1/\eta} \quad (\eta > 0) \quad (8)$$

$$L_k(y) = L(d_k(y)) \quad (9)$$

where $L(d_k(y))$ is the sigmoid function ($L(d) = 1 / (1+e^{-\alpha d+\beta})$).

In the competitive training process, the reference vector are given by minimizing the loss function through the steepest descent algorithm as follows ($\eta\rightarrow\infty$):

$$r_i \leftarrow r_i + \delta r_i \quad (10)$$

$$\delta r_i = -\varepsilon \partial L_k / \partial r_i$$

$$= 2\varepsilon L_k (1 - L_k)(y - r_k) \quad \text{for } i=k$$

$$- 2\varepsilon L_k (1 - L_k)(y - r_j) \quad \text{for } i=j \quad (11)$$

where ε is a positive constant. and $j=\underset{i\neq k}{\text{argmin}}\, g_i(x)$.

The KL/GPD algorithm together with the re-combination of phonetic segments reduces both of computation time and memory size in SMQ to 1/30 (feature dimension = 48) or 1/40 (32).

3.3 Competitive Training of a Feature Extractor

The feature extraction matrix W (W: w_1, w_2, …, w_m) can also be trained to improve the value of the loss function. The updating process is as follows:

$$W \leftarrow W + \delta W \quad (12)$$

$$\delta W = -\varepsilon \partial L_k / \partial W$$

$$= 2\varepsilon L_k (1 - L_k)(r_k - r_j) x^T \quad (13)$$

where $j=\underset{i\neq k}{\text{argmin}}\, g_i(x)$

After training, the initial orthogonalized feature vector set $\{w_1, w_2, \ldots, w_m\}$ is obliqued each other.

4. EXPERIMENTAL RESULT

4.1 Speech Database

3 sets of data were used for training and evaluation.

D-1: was used for designing the feature-vector extractor. 652 full Japanese phonetic segments were manually extracted from 250 phonetically balanced words uttered by 15 male and 15 female speakers each.

D-2: was used for sub-word HMM. D-2 consists of 492 isolated words including all the Japanese VCV context. 20 male speakers uttered these words once.

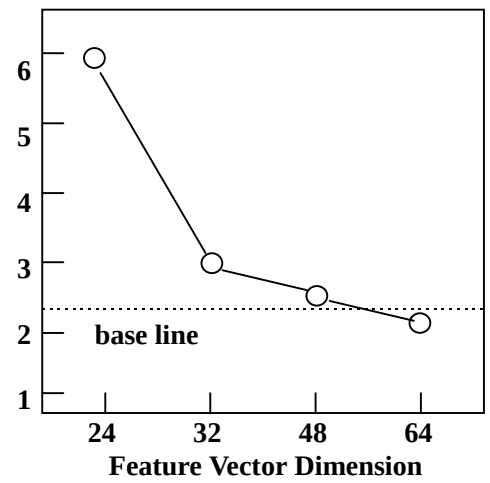

Fig. 3 Feature Vector Dimension vs. Word Recognition Error Rate

D-3: was used for word recognition test. 5 unknown male speakers uttered the same word list with D-2.

4.2 Experimental Result

The original pattern vector x with the dimension of 192 is converted to the feature-vector y with the dimension n. The horizontal axis of figure 3 shows n. In the figure, "base line" means our conventional method based on SMQ/HMM by using MCE/ALSM in which we used 652 phonetic-segment classes with the variable feature dimension from 64 to 192 and each class had 8 sub-spaces.

First, we applied MCE/GPD only to the classifier. ○ in figure 3 shows the experimental result. The feature-vector with the bigger number of dimension than 32 gives comparatively high performance, however, the smaller dimension of 24 stays in low accuracy.

Next, we applied MCE/GPD to the feature extractor as well as the classifier. The experiment was done to the worst case (feature dimension = 24). Table 1 shows the result. The error rate is depressed by 2 %.

5. CONCLUSION

The proposed method based on KL/GPD competitive training can achieve 1/30 - 1/40 reduction in both computation time and memory size with the same performance that the old version of SMQ shows. This method realizes a software only solution for vocabulary unspecific, speaker independent speech recognition based on SMQ/HMM and is bundled with Toshiba PCs as "Toshiba Speech System" software for Japanese.

TABLE - 1 Effect of Competitive Training for Feature Extractor

	Error Rate
Before Training	6.0 %
After Training	4.0 %

REFERENCES

[1] H.Matsu'ura, Y.Masai, J.Iwasaki, S.Tanaka, H. Kamio, and T.Nitta, "Multimodal, Keyword-based Spoken Dialogue System - MultiksDial", IEEE Proc. ICASSP94, pp.33-36 (1994).

[2] K. Maeda, Japanese Patent No. Showa-63-31831 (application date: March, 1980).

[3] M.Kuusela and E.Oja, "The Averaged Learning Subspace Method for Spectral Pattern Recognition", IEEE IJCPR'82, pp.134-137 (1982).

[4] S.Amari, "A Theory of Adaptive Pattern Classifiers", IEEE Trans. on Elec. Computers, vol.EC-16, No.3, pp.299-307 (June 1967).

[5] S.Katagiri, C.-H.Lee, and B.-H.Juang, "New Discriminative Training Algorithms Based on the Generalized Probabilistic Descent Method", IEEE Neural Networks for Signal Processing, pp.299-308 (1991).

[6] H.Watanabe, T.Yamaguchi, and S.Katagiri, " Discriminative Metric Design for Pattern Recognition ", IEEE Proc. ICASSP95, pp.3439-3442 (1995).

[7] E.Oja, "Subspace Method of Pattern Recognition ", Research Studies Press (1983).