# SPEAKER RECOGNITION MODEL USING TWO-DIMENSIONAL MEL-CEPSTRUM AND PREDICTIVE NEURAL NETWORK

*Tadashi KITAMURA and Shinsai TAKEI*

Dept. of Intelligence and Computer Science
Nagoya Institute of Technology
E-mail: kitamura@ics.nitech.ac.jp

## ABSTRACT

This paper describes a speaker recognition model using Two-Dimensional Mel-Cepstrum and predictive neural network. The speaker model consists of two networks. The first one is a self-organizing VQ map(Kohonen's feature map). The second part is the predictive network and learns transitional patterns on the feature map of each speaker's model. TDMC consists of averaged features and dynamic features of the two-dimensional mel-log spectra in the analyzed interval. The measure for speaker recognition is obtained by using a combination of the VQ distortion on the feature map and the prediction error on the predictive network. In the study, text-independent speaker identification experiments for 8 speakers were carried out. The experimental results have shown that a combination of a feature map and a predictive network is very effective, and that the proposed model using TDMC shows the robustness for time interval.

## 1. INTRODUCTION

For speaker recognition, power, pitch, spectral envelopes are very important to represent the characteristics of speakers. They have instantaneous features and transitional features of speech. Therefore, it has been considered that dynamic features as well as instantaneous features represented by power, pitch and spectral envelopes etc. play an important role. In general it is not so difficult to extract vocal tract information. Therefore, spectral envelopes are well used in speaker recognition. Some speaker recognition methods combining these features were reported to be effective[1]citefurui. Furthermore, the VQ-based methods using a codebook for each speaker and the HMM-based methods have been reported to been robust against utterance variation and give good a performance. However, it has been reported that when the training data is small, the performance of the VQ method decreases less than the HMM method [5].

This paper describes a speaker recognition model using a static feature model and dynamic feature model. The first one is a VQ feature map. In the training step, it learns automatically by a self-organizing learning algorithm and creates a feature map from the characteristic vectors of input speech. The second part consists of a three layer perceptron using a back-propagation algorithm. Transitional patterns of excited outputs on the feature map for each speaker are different. Therefore, the predictive network learns transitional patterns on the feature map of each speaker. In the recognition step, the first part of the model outputs a VQ distortion between the feature map and input speech vector. The second part also outputs a prediction error between the predicted position of the model and the current position of the excited unit. The final measure for speaker recognition is obtained using a combination of the VQ distortion and the prediction error. In order to evaluate the proposed model, a conventional mel-cepstrum (One-Dimensional Mel-Cepstrum) and Two-Dimensional Mel-Cepstrum (TDMC) parameters are compared. TDMC is defined as the Two-Dimensional Fourier transform of mel-frequency scaled logarithm spectra in the frequency and time domains. In the study, text-independent speaker identification experiments are carried out. The effectiveness to use a combination of the different networks and to use TDMC parameter as speech parameter is discussed.

## 2. SPEAKER MODEL

Fig.1 shows a block diagram of a proposed model for each speaker. This model consists of two parts representing static and dynamic spectral features of speech. The first one is a self-organizing neural network by Kohonen and an input vector is quantized into the excited unit. The second one is a predictive network and it predicts the next position on the map using a sequence of the past excited units.

### 2.1. Feature Map

In the learning step of VQ-based speaker recognition, a VQ codebook is generated by classifying the feature vector of input speech into a specified unit. In the recognition step, an average VQ distortion $E_{vq}^{(k)}$ for speaker $k$ is calculated using input speech. The speaker of input speech is considered as a speaker which gives a minimum value of $E_{vq}^{(k)}$. In this study a self-organizing algorithm by Kohonen is used to generate a VQ codebook.

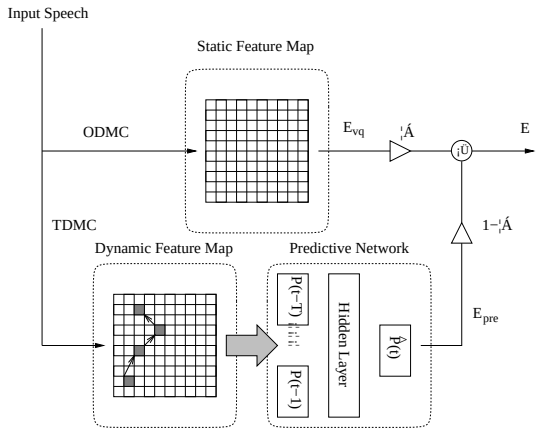


**Figure 1:** Block diagram of speaker recognition model combining a feature map and a predictive network.

## 2.2. Predictive Network

When a sequence of feature vectors of input speech is input to the feature map, the transitional pattern of the excited units are obtained. The transitional pattern for the relatively short sequence of each speaker has different pattern. Therefore, the network which predicts a current position by using the several past positions of the excited units is reported to be effective[3, 2, 7]. The neural network used here have a three layered feedforward neural network and learns using the backpropagation training algorithm. Each layer is fully interconnected with the higher layer. In this study, a sigmoid nonlinearity is used to calculate output because the transitional pattern is not linear.

The feature vector $\mathsf{p}(t)$ of a predictive network at time $t$ is defined by equation (1).

$$\mathsf{p}(t) = (x(t), y(t), \Delta x(t), \Delta y(t), \Delta c_o(t)) \qquad (1)$$

The output position vector $\mathsf{P}(t)$ of the output layer is defined by equation (2).

$$\mathsf{P}(t) = (x(t), y(t)) \qquad (2)$$

where $x(t), y(t)$ are elements of the position vector $\mathsf{P}(t)$, $\Delta x(t), \Delta y(t)$ are elements of the speed vector at time $t$, $\Delta c_o(t)$ is a speed of the zero order cepstrum which represents the power of speech. The predictive network predicts the current position $\mathsf{P}(t)$ using the past $T$ feature vectors, $\mathsf{p}(t-1), \mathsf{p}(t-2), \cdots, \mathsf{p}(t-T)$. Then it learns by the BP algorithm so that it minimizes the error between the predicted position $\widehat{\mathsf{P}}(t)$ and current position $\mathsf{P}(t)$ of excited unit on the feature map. In order to carry out the learning procedure effectively, the position on the feature map is normalized into [0, 1] and then coded by a Gray Code(a reflected binary code). The learning procedure is finished when the mean square error gives a minimum value for the another learning data set to stop learning procedure. Four sentences are used to stop the learning procedure. The desired outputs are set to 0.9 or 0.1 for the position of excited unit on the map, respectively. The momentum a and the learning rate $\eta$ were set to 0.7 and 0.3, respectively. A mean prediction error $E^{(k)}_{\mathsf{pre}}$ of speaker $k$ is used for recognition.

## 2.3. Identification Measure

In the speaker identification procedure, the feature vector of input speech is quantized on the static feature map and the VQ distortion is calculated. The position of the excited unit on the dynamic feature map is obtained as well. Then the predictive network outputs the current position using a sequence of the several past position vectors and gives the prediction error $E^{(k)}_{pre}$. Finally, the VQ distortion $E^{(k)}_{vq}$ and the prediction error $E^{(k)}_{pre}$ are combined into a distortion error $E^{(k)}$ in equation (3).

$$E^{(k)} = \alpha E^{(k)}_{vq} + (1-\alpha) E^{(k)}_{pre} \qquad (0 \le \alpha \le 1) \qquad (3)$$

where $\alpha$ is a combination factor. The input speaker is considered as a speaker $k$ who gives the minimum distortion error.

# 3. EXPERIMENTS OF PREDICTIVE NETWORK

## 3.1. Speech Data and Learning Condition

In order to study the number of frames or blocks for prediction, some experiments using only the predictive network were carried out. Speech database consists of 5 training sentences and 20 testing sentences of 8 female speakers. The speech data were recorded in the same session. The length of the training speech data is about 15 second.

In this study 15 orders of ODMC parameters and 60 orders of TDMC parameters are used respectively. Speech data is sampled at 10 kHz and ODMC parameter is analyzed using a 25.6 ms blackman time window and a frame period is 6.4 ms. TDMC parameter is analyzed using several frames of ODMC. Table 1 shows the analysis conditions of TDMC. It has been found that TDMC is very useful for speaker-independent word recognition. Especially, the dynamic features on TDMC is very useful for speech recognition[4]. In this study, TDMC weighted by standard deviation of each component of TDMC is used for optimization of TDMC parameters. Table 2 shows the training conditions of a feature map. The number of the input units and hidden units of the predictive network are changed according to the number of the necessary length for prediction. The output layer of the network has 10 units corresponding to the elements $\hat{x}(t), \hat{y}(t)$ of the predicted osition vectors $\widehat{\mathsf{P}}(t)$.

## 3.2. Predictive Network

In order to examine the optimal data length for prediction, identification experiments were carried out for 0.1, 0.36, 1.32, 1.96 second of input speech length, respectively. From the

**Table 1:** Analysis Condition of TDMC

| | |
|---|---|
| Block Size | 8 Frames |
| Block Length | 51.2 ms |
| Block Period | 12.8 ms |
| TDMC Area | Real : $0 \leq p \leq 2$, $1 \leq q \leq 12$<br>Imaginary: $1 \leq p \leq 2$, $1 \leq q \leq 12$) |

**Table 2:** Learning Condition of feature map

| | |
|---|---|
| Units | 100 |
| Momentum | 0.2 |
| Update Area | 3.0 |
| Iteration | 100 times |

experimental result for ODMC, it has been found that the number of prediciton frames of 8 gives the best result. It is equal to about 50 ms length. The experimental result for TDMC is shown in Fig.2. As in Fig.2, the maximum value of the recognition rates is obtained for $T$=2 blocks, which is about 60 ms length. When $T$ is 2 blocks, the recognition rates of 100 % are obtained for 1.32 and 1.96 second length of input speech. From these experimental results, it is considered that TDMC is more effective for training the neural network because the used area of TDMC has the spectral features smoothed with respect to time and frequency.

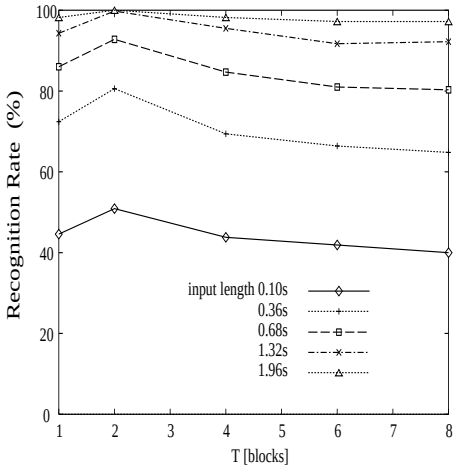


**Figure 2:** Recognition rate and the number of blocks for predicition of predictive network.

# 4. EXPERIMENTS

## 4.1. Combination Experiments

It has been found that a combination of several different features gives improvement for speech recognition[1, 6, 4]. Therefore, we examined the effectiveness of a combination of a feature map and a predictive network. The experiments for a combination of the feature map and the predictive network using ODMC and TDMC parameters were carried out. The relation between recognition rates and a combination factor $\alpha$ when the length of input speech is changed to 0.1, 0.36, 1.32, 1.96 second, were examined, respectively. From the experimental results for ODMC it has been found that $\alpha$ around 0.2 to 0.4 gives a slightly better result but this combination does not give much improvement.

Fig.3 shows the result for TDMC(TDMC$vq$ - TDMC$pre$). As in Fig.3, it has been found that when the combination factor *alpha* is about 0.1, the best ressults are obtained. When the input length are 0.36s and 0.68s, the optimal combination gives 10% and 4% better recognition rates than TDMC$pre$($\alpha$=0.0) only. Furthermore, the system gives 100% recognition rates for the input length of 1.32s and 1.96s.

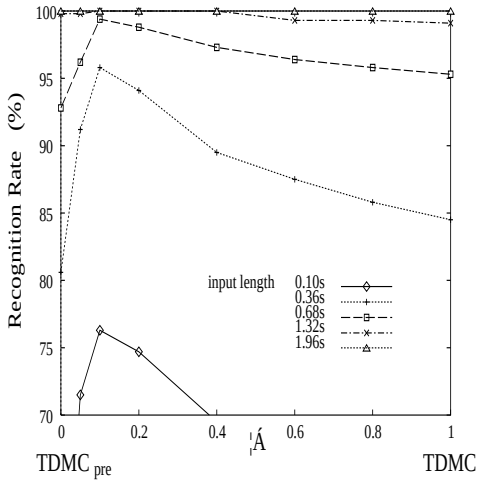


**Figure 3:** Recognition rate and a combination factor $\alpha$ for TDMC$vq$ and TDMC$pre$.

When only ODMC$vq$ is used, that is $\alpha$=0, ODMC$vq$ gives the better results than TDMC$vq$. When a combination parameter $\alpha$ of 0.0 is used, that is, only the predictive network is used, TDMC$pre$ gives the better result than ODMC$pre$. Therefore, experiments using a combination of ODMC$vq$ and TDMC$pre$ were carried out. Fig.4 shows the experimental result for several lengths of input speech. This combination gives the better results than that of ODMC$vq$ -ODMC$pre$ or TDMC$vq$ -TDMC$pre$. When $\alpha$ is about 0.1, the model gives the best recognition rresult for each input length. Recognition rate of 99% was obtained even for the input length of 0.68s. As mentioned above, the proposed system uses a feature map as static feature and a predictive network as a dynamic feature, respectively. Experimental results have shown that the combination of ODMC$vq$ and TDMC$pre$ is optimal.

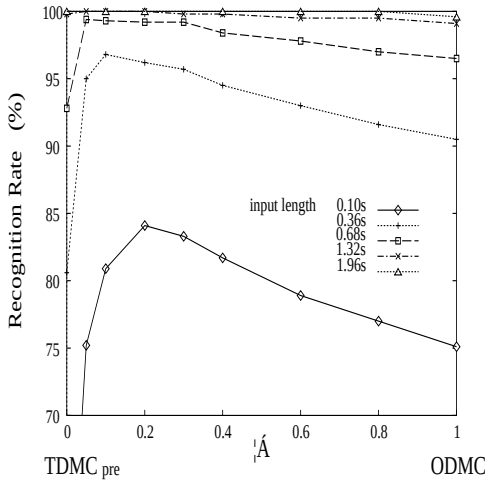


**Figure 4:** Recognition rate and a combination factor $\alpha$ for TDMC*vq* and TDMC*pre*.

## 4.2. Robustness for Time Variation

In order to examine the robustness for time interval of input speech, the some experiments using another database were carried out. The used speech database consists of 8 male speakers. Each speaker recorded every one month for six months and all sentences are same as in section 4. Five sentences were used for the training of the model and the remaining twenty sentences were used for testing. From the previous experiments, the combination of TDMC*vq* - TDMC*pre*, ODMC*vq* -TDMC*pre* and TDMC*vq* -TDMC*pre* were used for the experiments. Fig.5 shows the experimental result for 1.96s length of input speech. As in Fig.fig8, the combination of ODMC*vq* -TDMC*pre* gives the best result. Therefore, an utilization of TDMC*pre* is considered to be very effective for speaker recognition.

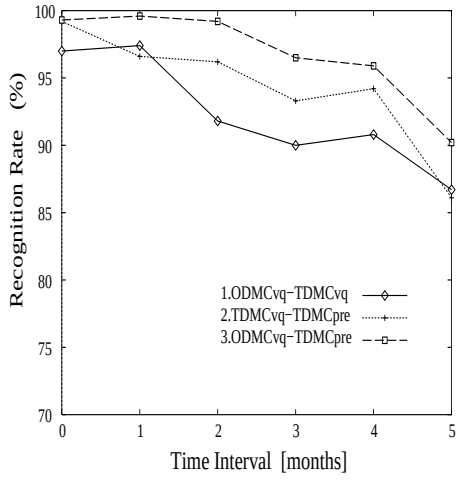


**Figure 5:** Recognition rate and a combination factor $\alpha$ for the different combinations.

# 5. CONCLUSION

In this paper we have proposed a speaker recognition model using Two-Dimensional Mel-Cepstrum and a predictive network. Each speaker model consists of two parts. The first one is a self-organizing VQ map and outputs VQ distortion. The second one is a predictive network and learns transitional patterns on the feature map of the each speaker's model. It give a prediction error between the predicted position and the current position of the excited unit of the dynamic feature map. The combination of the VQ distortion and prediction error is used to identify input speaker. From experimental results of text-independent speaker identification, it has been shown that the proposed gives a good model for speaker recognition and TDMC parameter is very effective.

# 6. REFERENCES


1. F.K.Soong and A.E.Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. *Proc. of ICASSP86*, 86, 1986.
2. H. Hattori. Text-independent speaker recognition using neural networks. *Tech. Report of IEICE*, 91:SP91–90:71–78, 1991.
3. K. Iso. Speech recognition using neural predicition model. *Tech. Report of IEICE*, 89:SP89–23:81–87, 1989.
4. T. Kitamura, E. Hayahara, and Y. Shimazaki. Speaker-independent word recognition using dynamic and averaged features based on a two-dimensional mel-cepstrum. *Proc. of ICSLP*, 90:25–5:1129–1132, 1990.
5. T. Matsui and S. Furui. Comparison of text-independent speaker recognition method using vq-distortion and discrete/countinous hmms. *Tech. Report of IEICE*, 91:SP91–89:65–70, 1991.
6. S.Furui. Comparison of spekear recognition methods using statistical features and dynamic features. *IEEE Trans. ASSP*, 29:342–350, 1986.
7. K. Sumida and T. Kitamura. Speaker identification using feature map of speech and prediction error. *Tech. Report of IEICE*, 92:SP92–73:1–6, 1992.