Speaker Independent Phoneme Recognition Based on Fisher Weight Map

Takashi Muroi, Tetsuya Takiguchi, Yasuo Ariki
Department of Computer and System Engineering
Kobe University, 1-1 Rokkodai, Nada, Kobe, 657-8501, JAPAN
muroi@me.cs.scitec.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

Abstract

We have already proposed a new feature extraction method based on higher-order local auto-correlation and Fisher weight map (FWM) at Interspeech2006. This paper shows effectiveness of the proposed FWM in speaker dependent and speaker independent phoneme recognition. Widely used MFCC features lack temporal dynamics. To solve this problem, local auto-correlation features are computed and accumulated by weighting high scores on the discriminative areas. This score map is called Fisher weight map. From the speaker dependent phoneme recognition, the proposed FWM showed 79.5% recognition rate, by 5.0 points higher than the result by MFCC. Furhermore by combing FWM with MFCC and ΔMFCC, the recognition rate improved to 88.3%. In the speaker independent phoneme recognition, it showed 84.2% recognition rate, by 11.0 points higher than the result by MFCC. By combining FWM with MFCC and ΔMFCC, the reecognition rate improved to 89.0%.

1: Introduction

In speech recognition, MFCC (Mel-Frequency Cepstrum Coefficient) is widely used which is a cepstrum conversion of a sub-band mel-frequency spectrum within a short time. Due to the characteristic of short time spectrum, MFCC lacks temporal dynamic features and degrades the recognition rate. To overcome this defect, the regression coefficients of MFCC (delta, delta delta MFCC) are usually utilized[1][2], but they are indirect expression of temporal frequency changes such as formant transition or high frequency plosives.

More direct expression of the temporal frequency changes will be a geometrical feature in a two-dimensional local area, for example within 3 frames by 3 frequency bands area, on the temporal frequency domain[3][4]. In order to locate such two-dimensional geometrical features, auto-correlation within a local area is effective because it can enhance the geometrical features. Originally this type of feature extraction was proposed in the field of facial emotion recognition [5]. Otsu computed 35 types of local auto-correlation features within a two-dimensional local area at each pixel on an image and accumulated them within some discriminative areas where the typical features among all emotions were well expressed. The map showing this discriminative areas was called Fisher weight map and Otsu employed a discriminant analysis to find this Fisher weight map.

We have already proposed a method to find the geometrical discriminative features and discriminative areas of phonemes on the temporal-frequency domain of speech signals by using the Fisher weight maps and showed the effectiveness by vowel recognition[6]. In this paper, effectiveness of the proposed discriminative feature is verified through speaker dependent and speaker independent 25 phoneme recognition experiments.

In section 2 of this paper, we describe an extraction flow of the geometrical discriminative features for phoneme recognition. In section 3 and 4, auto-correlation coefficients based on the local features and the Fisher weight maps are described. In section 5, speaker dependent and speaker independent phoneme recognition experiments are shown.

2: Extraction flow of geometrical discriminative features

Fig.1 shows an extraction flow of geometrical discriminative features and phoneme recognition. At first, speech waveforms are converted into time-frequency domain by short-time Fourier transformation. At this point, a time sequence of short-time spectra (frames) is obtained. Then a moving window with consecutive several frames is put on the time sequence of short-time spectra, forming a windowed time-frequency matrix. Local features of 35 types are computed at each position (time, frequency) within this window, forming a local feature matrix H with the number of positions $\times$ 35 types of local features.

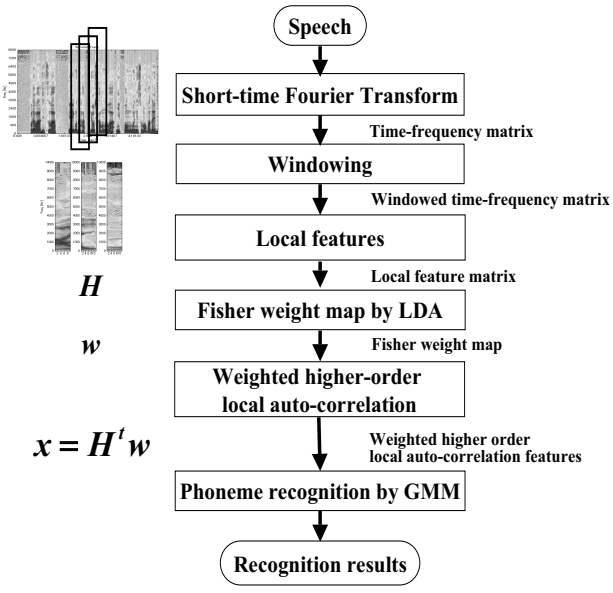

Figure 1. Flow of new feature extraction.

Finally Fisher weight map w is produced by applying linear discriminant analysis (LDA) to the local feature matrix H. Geometrical discriminative features are obtained as weighted higher-order local auto-correlation by summing up the local features weighted by the Fisher weight map for each type of local features, forming 35 dimensional vector x for a window. By moving this window, a sequence of 35 dimensional vectors of geometrical

discriminative features are obtained.

In a phoneme recognition, phoneme GMMs are trained at first. Then the test speech data is converted into a sequence of 35 dimensional vectors of geometrical discriminative features and phoneme likelihood is computed using the trained phoneme GMMs.

3 Local features and weighted higher order local auto-correlations

3.1 Local features

Two-dimensional geometrical and local features are observed on the time-frequency matrix shown on the left in Fig.2. On the right hand side, 3×3 local patterns are shown to capture the local features. The upper pattern is for continuation in a time direction, the middle for continuation in a frequency direction and the lower for transition. The flag "1" indicates the multiplication of the spectrum on the position.

A local feature within the k-th local pattern at a position r is formalized as follows;

$$h_r^{(k)} = I(r)I(r + a_1^{(k)}) \cdots I(r + a_N^{(k)}) \tag{1}$$

where $I(r)$ is the power spectrum at the position r on time-frequency matrix composed of time t and frequency f. The $r + a_i^{(k)}$ indicates the other position, where "1" is attached, within the k-th local pattern.

By limiting local patterns within 3 frames $\times$ 3 bands area at reference position r, setting the order N to be 2 and omitting the equivalence of translation, the number of displacement set $(a_1, \cdots, a_N)$ becomes 35. Namely 35 types of local patterns are obtained at each position r on the time-frequency matrix as shown in Fig.3, according to Otsu[5].

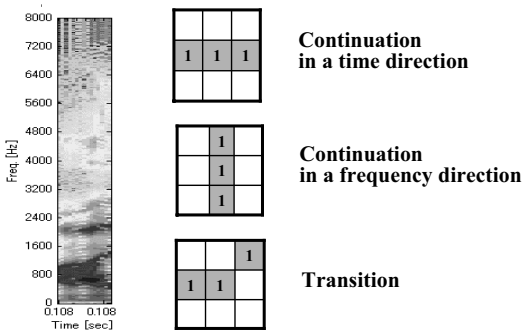

Figure 2. Local features.

3.2 Weighted higher order local auto-correlations

Higher-order local auto-correlation x_k for the k-th local pattern is obtained by summing the local features shown in Eq.1 on the time-frequency matrix. It is formalized as follows;

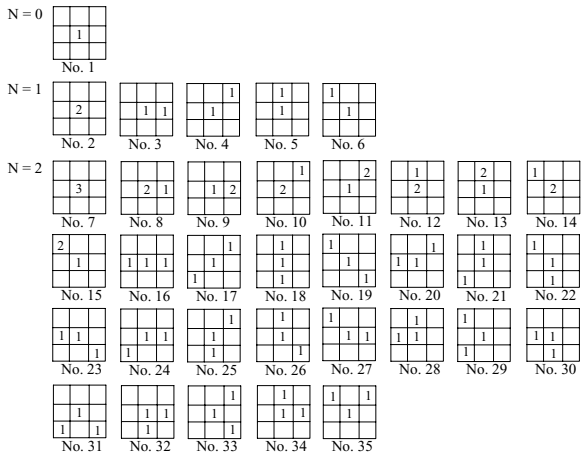

Figure 3. 35 types of local patterns.

$$\begin{aligned} x_k &= \sum_r h_r^{(k)} \\ &= \sum_r I(r) I(r + a_1^{(k)}) \cdots I(r + a_N^{(k)}) \end{aligned} \tag{2}$$

In order to express the higher-order local auto-correlation in the matrix form, all the local features shown in Eq.1 for the k-th local pattern are collected on the time-frequency matrix and presented as a following vector.

$$\mathbf{h}^{(k)} = [h_{2,2}^{(k)} \cdots h_{2,T-1}^{(k)}, \cdots h_{F-1,T-1}^{(k)}]^t \tag{3}$$

here the dimension of the vector is $M = T - 2$ (time) $\times$ $F - 2$ (frequency).

The higher-order local auto-correlation x_k for the k-th local pattern is expressed as follows using the M-dimensional vector $\mathbf{h}^{(k)}$.

$$x_k = \mathbf{h}^{(k)t} \mathbf{1} \tag{4}$$

A local feature matrix is obtained as follows by placing the M-dimensional vectors $\mathbf{h}^{(k)}$ in the horizontal direction one by one for all the 35 local patterns.

$$\mathbf{H} = [\mathbf{h}^{(1)} \cdots \mathbf{h}^{(K)}] \tag{5}$$

The higher-order local auto-correlation vector $\mathbf{x}$ is obtained by packing the x_k and is expressed as follows;

$$\mathbf{x} = [x_1 \cdots x_K]^t = \mathbf{H}^t \mathbf{1} \tag{6}$$

Fig.4 shows an example of computing the local feature matrix $\mathbf{H}$. Here, moving 35 local patterns on the windowed time-frequency matrix (9×6), the local features are computed. These local features are packed into the local feature matrix $\mathbf{H}$ (28×35). The higher-order

local auto-correlation vector $\mathbf{x}$ presents the existence of the local patterns on all over the time-frequency matrix. Therefore, it is not the discriminative vector. In order to make the higher-order local auto-correlation vector $\mathbf{x}$ have the discriminative ability, local features of the same local pattern are summed over the windowed time-frequency matrix by putting the high weight on the local features where class difference appears clearly. This is done by replacing the vector **1** consisting of M "1"s by the weighting vector $\mathbf{w}$. Then the weighted higher-order local auto-correlation vector $\mathbf{x}$ is obtained as follows;

$$\mathbf{x} = \mathbf{H}^t \mathbf{w} \tag{7}$$

Here $\mathbf{w}$ is called Fisher weight map because it is computed based on linear discriminant analysis.

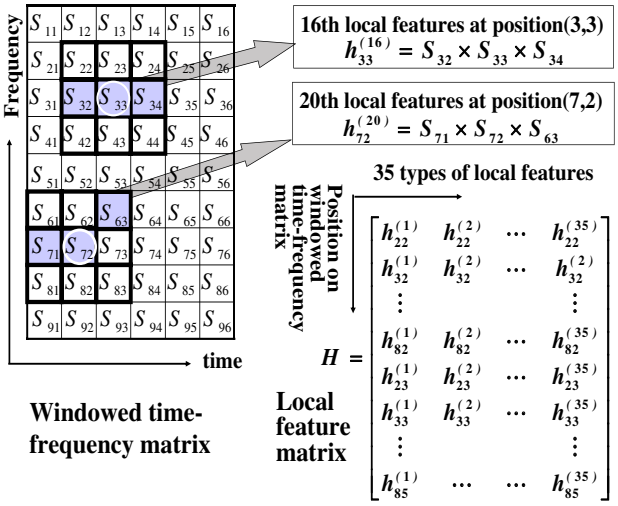

Figure 4. Local feature matrix.

4 Fisher weight map

In order to find the Fisher weight map, Fisher's discriminative criterion is utilized[5]. Let N be the number of training data. Then the local feature matrices for the training data are denoted as $\{\mathbf{H}_i \in R^{M \times K}\}_{i=1}^{N}$. The corresponding weighted higher-order local auto-correlation vectors, the within-class covariance matrix and the between-class covariance matrix are denoted as $\{\mathbf{x}_i\}_{i=1}^{N}$, $\tilde{\Sigma}_W$ and $\tilde{\Sigma}_B$ respectively. Then the Fisher discriminative criterion $J(\mathbf{w})$ is expressed as follows using those denotations.

$$J(\mathbf{w}) = \frac{tr\tilde{\Sigma}_B}{tr\tilde{\Sigma}_W} = \frac{\mathbf{w}^t \Sigma_B \mathbf{w}}{\mathbf{w}^t \Sigma_W \mathbf{w}} \tag{8}$$

where Σ_W and Σ_B is the within-class covariance matrix and the between-class matrix of the local feature matrices (training data).

The Fisher weight map is obtained as eigen vectors $\mathbf{w}$ based on the following generalized eigen value decomposition derived by maximizing the Fisher discriminative criterion under the constraint such that $\mathbf{w}^t \Sigma_W \mathbf{w} = 1$

$$\Sigma_B \mathbf{w} = \lambda \Sigma_W \mathbf{w} \tag{9}$$

Since the Fisher weight map is composed of several eigen vectors, the number of eigen vectors is optimized in the phoneme recognition process.

However, if the number of eigen vectors are set to 25, the weighted higher-order local auto-correlation vector $\mathbf{x}$ shown in EQ.7 equals to 875 (35×25) dimensional vector. It is so high that the GMM used in the phoneme recognition can not be estimated accurately and stably. To solve this problem, PCA (Principal Component Analysis) is used to reduce the dimension effectively.

5 Phoneme recognition experiments

5.1 Experimental setup

We carried out speaker dependent and independent Japanese 25 phoneme recognition. Speech material was continuous speech data spoken by six male speakers and four female speakers and was manually segmented into phoneme sections. In the speaker dependent phoneme recognition, 2578 data (about 100 data for each phoneme) segmented by hands for all phonemes were collected from individual speaker and used for phoneme training (Fisher weight map and phoneme GMMs) . Other 2578 phoneme data from individual speaker were tested. Phoneme recognition rate was computed by averaging the results from ten speakers.

On the other hand, in the speaker independent phoneme recognition, the training data from ten speakers were collected together and used for Fisher weight map and phoneme GMMs training. In the phoneme recognition, the test data from individual speaker was tested in the same way as the speaker dependent manner.

Speech waveform was transformed into time-frequency matrix by short-time Fourier transformation with 25ms frame width and 10ms frame shift. Then the frequency was converted into mel-scale by mel-fiter bank (64 dimension). A window with T frame width and S frame shift was moved on the time-frequency matrix and the windowed time-frequency matrices were generated. T and S were optimized experimentally to 5 and 1 respectively. The number of eigen vectors W included in the Fisher weight map and the number of Gaussian mixtures G in phoneme GMM were experimentally optimized in the phoneme recognition. The number of dimensions D of the weighted higher-order local auto-correlation vector $\mathbf{x}$ reduced by PCA was also experimentally optimized.

5.2 Speaker dependent phoneme recognition using single feature

Fig.5 shows the results of speaker dependent phoneme recognition using the proposed feature, compared with the recognition result using MFCC.

The highest phoneme recognition rate 79.5% was obtained by the proposed feature with the number of eigen vectors $W = 25$ (35×25=875 dimensions) in the Fisher weight map,

the number of dimensions $D = 150$ of the weighted higher-order local auto-correlation vector $\mathbf{x}$ reduced by PCA and the number of Gaussian mixtures $G = 8$ in the phoneme GMMs. Compared with MFCC and ΔMFCC, the recognition rate was improved by 5 points and 3.7 points respectively due to the direct expression of temporal features by the proposed method. When the PCA was not applied, since the dimension is so high as 875, the recognition rate was almost same as that of MFCC.

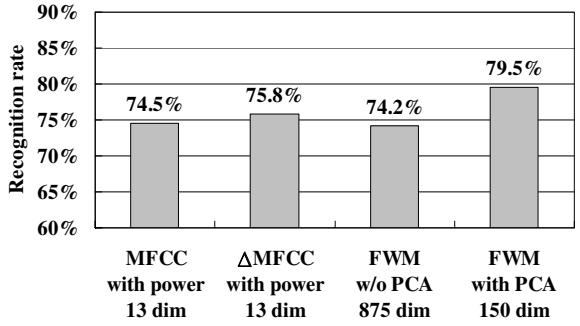

Figure 5. Results of speaker dependent phoneme recognition using single feature.

5.3 Speaker dependent phoneme recognition by feature integration

Since FWM showed the highest phoneme recogntion rate using single feature, it was combined with MFCC and ΔMFCC in the phoneme recognition. The feature combination was based on a stream weighting method which concatenated two or more feature vectors by weighting the respective feature. The weight was experimentally optimized, changing the weight ratio from 0.0:1.0 to 1.0:0.0 by 0.1 step. In this case, the dimension of FWM was decreased to 55 from 150 due to computation time.

Fig.6 shows the phoneme recognition result. FWM improved the recognition rate by 2.6 points and 6.0 points after combined with MFCC and ΔMFCC respectively compared with original FWM (79.5% in Fig.5). Comibination of two features MFCC and ΔMFCC still showed the highest score 86.7%. When three features FWM, MFCC and ΔMFCC were combined together, the recogntioin rate showed the highest score 88.3%. This indicates that the FWM has information to improve the recognition obtained by MFCC and ΔMFCC combination.

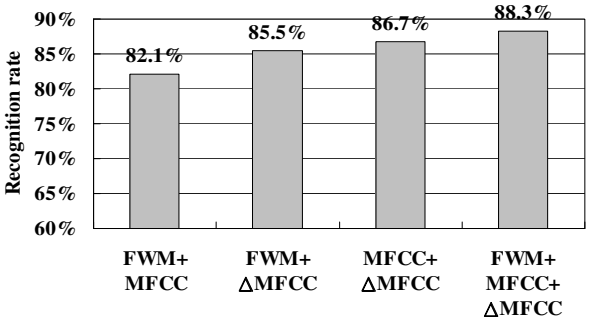

Figure 6. Results of speaker dependent phoneme recognition by feature integration.

5.4 Speaker independent phoneme recognition using single feature

Fig.7 shows the results of speaker independent phoneme recognition using the proposed feature FWM, compared with the recognition result using MFCC.

The highest phoneme recognition rate 84.2% was obtained by the proposed feature FWM with the number of eigen vectors $W = 35$ (35 $\times$ 35=1225 dimensions) in the Fisher weight map, the number of dimensions $D = 50$, instead of $D = 150$, of the weighted higher-order local auto-correlation vector $\mathbf{x}$ reduced by PCA and the number of Gaussian mixtures $G = 8$ in the phoneme GMMs. Compared with MFCC and ΔMFCC, the recognition rate was improved by 11 points and 9.2 points respectively due to accumulation of the direct expression of temporal features of 10 person by the proposed method. Compared with speaker dependent result shown in Fig.5, the result of MFCC and ΔMFCC decreased due to data variation. However the result of FWM showed 4.7 points improvement by speaker independency due to less data variation of Fisher weight map produced by 10 person.

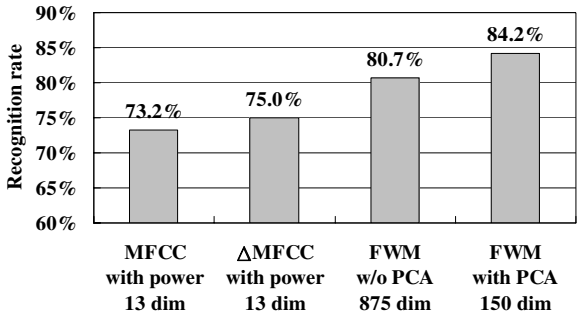

Figure 7. Results of speaker independent phoneme recognition by single feature.

5.5 Speaker independent phoneme recognition by feature integration

FWM was combined with MFCC and ΔMFCC based on a stream weighting method. The result is shown in Fig.8. FWM improved the recognition rate by 1.4 points and 2.9 points after combined with MFCC and ΔMFCC respectively compared with original speaker independent FWM (84.2% in Fig.7). When three features FWM, MFCC and ΔMFCC were combined together, the recogntioin rate showed the highest score 89.0% that was 1.9 points higher than the result of MFCC+ΔMFCC. This indicates that the FWM has information to improve the recognition rate obtained by MFCC and ΔMFCC combination.

6 Conclusion

We described the new feature extraction method based on higher-order local auto-correlation and Fisher weight map (FWM). The effectiveness was verified through speaker dependent and speaker independent phoneme recognition. From the speaker dependent phoneme recognition, the proposed FWM showed 79.5% recognition rate, by 5.0% point higher than the result by MFCC. Furhermore by combing FWM with MFCC and ΔMFCC, the

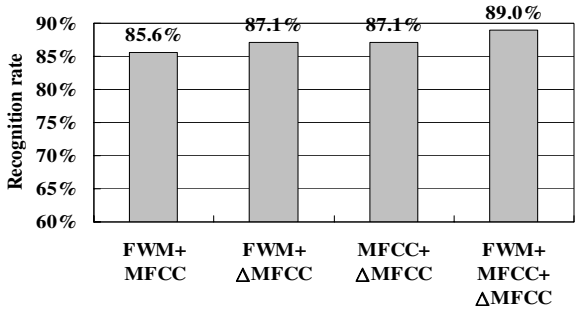

Figure 8. Results of speaker independent phoneme recognition by feature integration.

recognition rate improved to 88.3%. In the speaker independent phoneme recognition, it showed 84.2% recognition rate, by 11.0 points higher than the result by MFCC. By combining FWM with MFCC and ΔMFCC, the recognition improved to 89.0%.

As future works, we will investigate the noise robustness of the proposed method because the higher order local auto-correlation used in the method is thought to be robust for noisy speech recognition. Another plan is to extend the method into HMM expression and to apply it to the continuous phoneme recognition. The problem of the method will be lack of the normalization like CMN and composition of GMM or HMM with noise components. We will investigate these problems theoretically as studied in [7].

References

[1] K. Elinius, M. Blomberg, “Effect of emphasizing transitional or stationary parts of the speech signal in a discrete utterance recognition system,” IEEE Proc. ICASSP ’82, pp.535-538, 1982.

[2] T. Nitta, “A novel feature-extraction for speech recognition based on multiple acoustic -feature planes,” IEEE proc. ICASSP ’98, pp.29-32, 1998.

[3] Ken Schutte, James Glass, “Speech Recognition with Localized Time-Frequency Pattern Detectors”, Proceedings of ASRU 2007, pp.341-344, 2007.

[4] T. Nitta, ”Feature Extraction for Speech Recognition Based on Orthogonal Acoustic-feature Planes and LDA”, Proceedings of IEEE ICASSP’1999, pp.421-424, May 1999.

[5] N. Otsu, ”Facial Expression Recognition Using Fisher Weight Maps”, FGR 2004, pp.499-504, 2004.

[6] Y. Ariki, S. Kato, ”Phoneme Recognition Based on Fisher Weight Map to Higher-Order Local Auto-Correlation”, Interspeech2006, pp.377-380, Sept. 2006.

[7] Cooke, M. P., Green, P. D., Josifovski, L. B., and Vizinho, A., ”Robust automatic speech recognition with missing and uncertain acoustic data”, Speech Communication, 34, pp.267-285, 2001.