Phoneme Recognition Based on Fisher Weight Map to Higher-Order Local Auto-Correlation

Yasuo Ariki, Shunsuke Kato, Tetsuya Takiguchi

Department of Computer and System Engineering
Kobe University, 1-1 Rokkodai, Nada, Kobe, 657-8501, JAPAN
ariki@kobe-u.ac.jp

Abstract

In this paper, we propose a new feature extraction method based on higher-order local auto-correlation (HLAC) and Fisher weight map (FWM). Widely used MFCC features lack temporal dynamics. To solve this problem, 35 types of local auto-correlation features are computed within two-dimensional local regions. These local features are accumulated over more global regions by weighting high scores on the discriminative areas where the typical features among all phonemes are well expressed. This score map is called Fisher weight map. We verified the effectiveness of the HLAC and FWM through vowel recognition and total phoneme recognition.

Index Terms: phoneme recognition, linear discriminant analysis, Fisher weight map.

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, 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[1]. Fig.1 shows a time wave and spectrogram of a word "democrats". On the lower frequency band, several formant transitions are observed and in a high frequency band, the plosive is observed.

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 [2]. 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 propose, in this paper, 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. In the vowel recognition, the formant features were proved to be a discriminative features by investigating the resultant Fisher weight maps.

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, phoneme recognition experiments are shown.

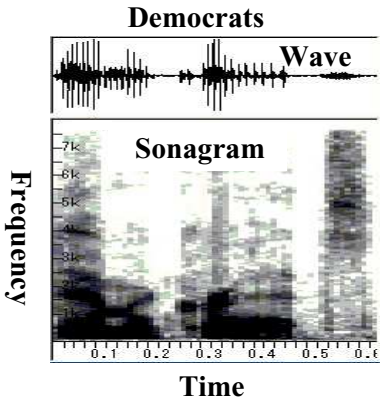


Figure 1: *Example of a spectrogram of speech signal.*

2. Extraction flow of geometrical discriminative features

Fig.2 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.

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.

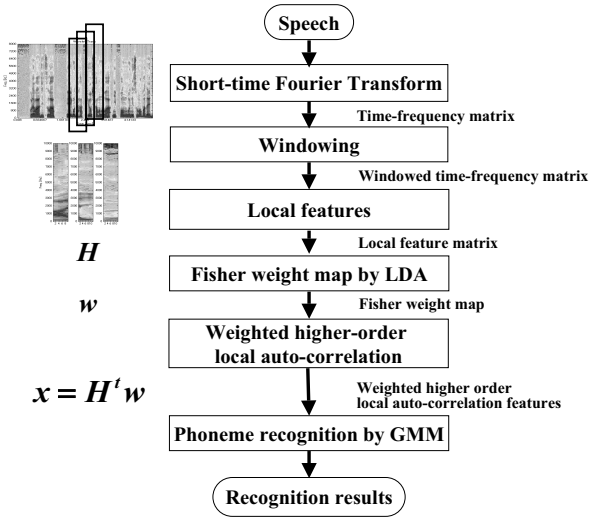


Figure 2: *Flow of new feature extraction.*

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.3. 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.4, according to Otsu[2]. In the figure, "2" and "3" indicate the square and the cube.

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;

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

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.5 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 $\mathbf{1}$ consisting of M

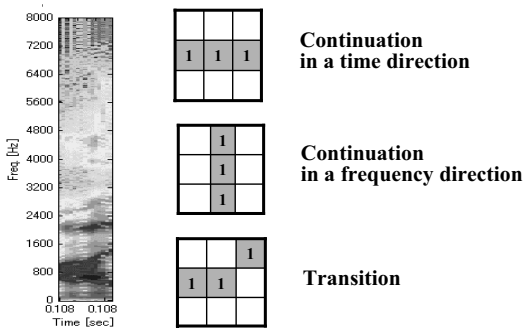


Figure 3: *Local features.*

"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.

4. Fisher weight map

In order to find the Fisher weight map, Fisher's discriminative criterion is utilized[2]. Let N be the number of training data. Then the local feature matrices for the training data is 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.

5. Phoneme recognition experiments

5.1. Experimental setup

We carried out Japanese 5 vowel recognition and total 27 phonemes recognition. Speech material was continuous speech data spoken by one male speaker and was manually segmented into phoneme sections. In the vowel recognition, 100 data segmented by hands for each vowel (in total 500 data) were used to train each vowel GMM and other 100 data were tested for each vowel. In the total phoneme recognition, 2578 data segmented by hands for all phonemes were used for total phoneme training and other 2578 phoneme data were tested.

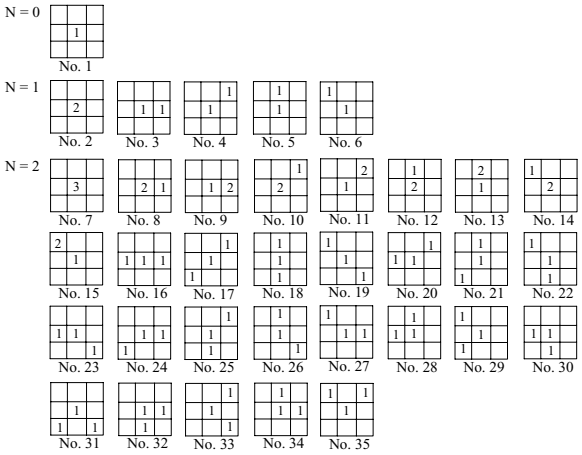


Figure 4: *35 types of local patterns.*

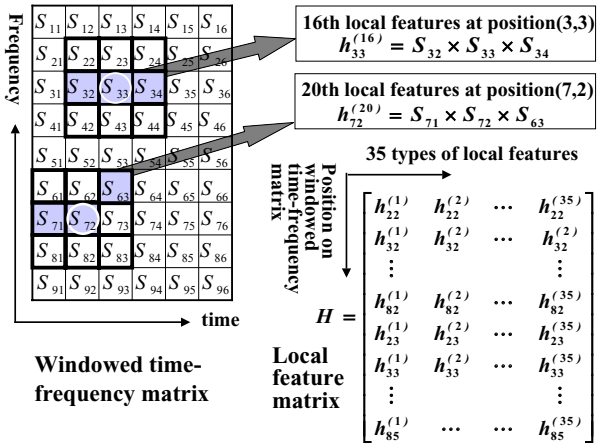


Figure 5: *Local feature matrix.*

Speech waveform was transformed into time-frequency matrix by short-time Fourier transformation with 25ms frame width and 10ms frame shift. Then a window with T frame width and S frame shift was moved on the time-frequency matrix and the windowed time-frequency matrix was generated. The number of eigen vectors W included in the Fisher weight map was optimized in the phoneme recognition. The number of Gaussian mixtures G in phoneme GMM was also optimized experimentally.

5.2. Recognition results

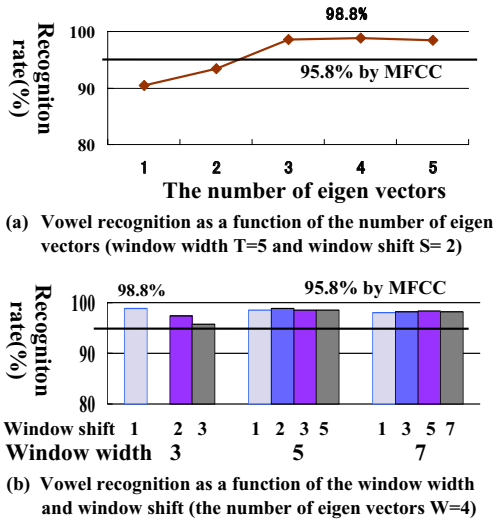
Table1 shows the recognition results for vowel and total phonemes, compared with the recognition result using MFCC with 12 coefficients (delta is not used). In the vowel recognition, the highest recognition rate 98.8% was obtained with the moving window width $T = 3$ frames, the window shift $S = 1$ frame, the number of eigen vectors $W = 3$ (35 $\times$ 3=105 dimensions) in the Fisher weight map and the number of Gaussian mixtures $G = 1$ in vowel GMMs. Compared with MFCC, the recognition rate 98.8%, 3 points higher, was achieved.

Table 1: *Phoneme recognition result.*

	Vowel	Total phonemes
Proposed method	98.8%	81.7%
MFCC	95.8%	84.6%

Fig.6 shows the dependency of vowel recognition on the number of eigen vectors W in the Fisher weight map, the moving window width T and the window shift S. From the figure, they are optimized at $W = 4$, $T = 3$ and $S = 1$. The number of Gaussian mixtures G in vowel GMM was optimized for each condition. Looking into the Fisher weight map thus obtained ($W = 5$), formant frequencies show higher score as shown in Fig.7(a) with

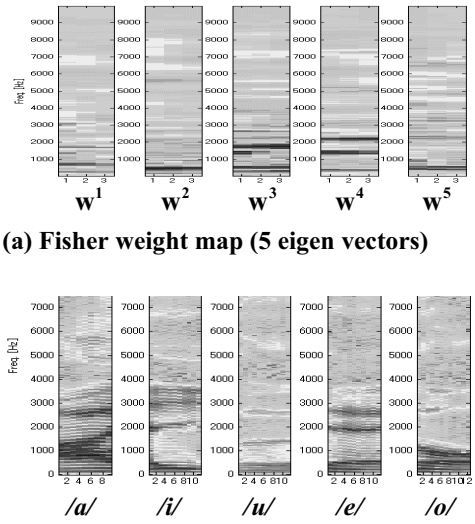
horizontal black stripes.



(a) Vowel recognition as a function of the number of eigen vectors (window width T=5 and window shift S= 2)

(b) Vowel recognition as a function of the window width and window shift (the number of eigen vectors W=4)

Figure 6: *Parameter dependency in vowel recognition.*



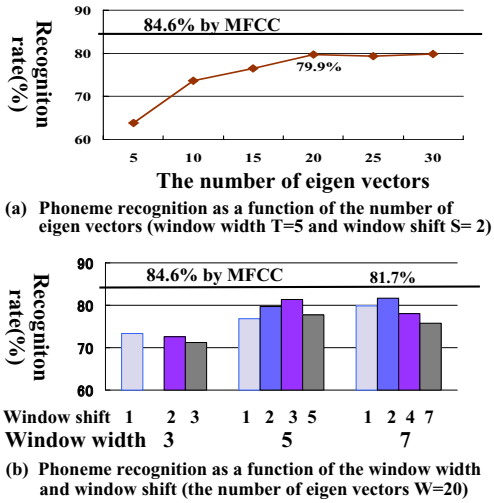
(a) Fisher weight map (5 eigen vectors)

(b) Vowels on time-frequency matrix

Figure 7: *Fisher weight map for vowels.*

In the total phoneme recognition, the highest recognition rate 81.7% was obtained with the moving window width $T = 7$ frames, the window shift $S = 2$ frame, the number of eigen vectors $W = 20$ (35 $\times$ 20=700 dimensions) in the Fisher weight map and the number of Gaussian mixtures $G = 8$ in phoneme GMMs. Compared with MFCC, the recognition rate was lower due to the recognition degradation of the special phonemes such as /h/, /m/, /r/, /t/, /w/ and /y/ with recognition rate 45.0%, 48.0%, 31.0%, 58.0%, 66.0% and 57.0% respectively.

Fig.8 shows the dependency of phoneme recognition on the number of eigen vectors W in the Fisher weight map, the moving window width T and the window shift S. From the figure, they



(a) Phoneme recognition as a function of the number of eigen vectors (window width T=5 and window shift S= 2)

(b) Phoneme recognition as a function of the window width and window shift (the number of eigen vectors W=20)

Figure 8: *Parameter dependency in phoneme recognition.*

are optimized at $W = 20$, $T = 7$ and $S = 2$. The number of Gaussian mixtures G in phoneme GMM was optimized for each condition.

6. Conclusion

We described the new feature extraction method based on higher-order local auto-correlation (HLAC) and Fisher weight map (FWM). The effectiveness was verified through vowel recognition with 3 point improvement compared with MFCC. For total phoneme recognition, at present, the recognition rate is still less than MFCC. However it will be improved by employing pair-wise linear discriminant analysis[3].

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 [4].

7. References

[1] 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.

[2] Yusuke Shinohara, Nobuyuki Otsu, "Facial Expression Recognition Using Fisher Weight Maps", FGR 2004, pp.499-504, 2004.

[3] S. Kitazawa, H. Kojima, and S. Doshita, "Multiclass Pattern Recognition Based on Pairwise Discrimination", Trans. IEICE (A), vol.J72-A, no.1, pp.41-48, 1989 (Japanese).

[4] 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.