INTEGRATION OF PHONEME-SUBSPACES USING ICA FOR SPEECH FEATURE EXTRACTION AND RECOGNITION

Hyunsin Park, Tetsuya Takiguchi, and *Yasuo Ariki*

Graduate School of Engineering, Kobe University
1-1, Rokkodai-cho, Nada-ku, Kobe, 657-8501, Japan
silentbattle@me.cs.scitec.kobe-u.ac.jp, takigu@kobe-u.ac.jp, ariki@kobe-u.ac.jp

ABSTRACT

In our previous work, the use of PCA instead of DCT shows robustness in distorted speech recognition because the main speech element is projected onto low-order features, while the noise or distortion element is projected onto high-order features [1]. This paper introduces a new feature extraction technique that collects the correlation information among phoneme subspaces and their elements are statistically mutual independent. The proposed speech feature vector is generated by projecting observed vector onto integrated space obtained by PCA and ICA. The performance evaluation shows that the proposed method provides a higher isolated word recognition accuracy than conventional methods in some reverberant conditions.

Index Terms— Speech recognition, Feature extraction, Subspace integration, PCA, ICA,

1. INTRODUCTION

In the case of distant or hands-free speech recognition, system performance decreases sharply due to ambient noises. Hence, there have been many studies carried out on robust feature extraction: RASTA speech processing [9], channel normalization [10], noise estimation [11], dereverberation [12], speech enhancement or separation based on Principal Component Analysis (PCA) [13, 14, 15], and so on.

In recent years, MFCC (Mel-Frequency Cepstrum Coefficient) is the widely used speech feature. However, since the feature space of MFCC by DCT (Discrete Cosine Transform) is not directly dependent on speech data, the observed signal with noise does not show good performance. In [2], the use of subspace method by PCA (Principal Component Analysis) shows robustness in noisy speech recognition. And in [1], the use of PCA also shows robustness in distorted speech recognition. That is because clean speech components are extracted from observed signal by projecting observed signal onto the speech subspace which retains the structure of the speech.

This paper proposes a new feature extraction technique that collects the statistically independent information among phonemes by projecting input vector onto integrated space by applying ICA (Independent component analysis) and subspace method maintaining robustness. The evaluation experiments by isolated word speech recognition for clean and reverberant speech show the effectiveness of the proposed method.

The content of this paper is as follows: In section 2, we describe the conventional feature extraction method using PCA or ICA. In section 3, we propose a new feature extraction method based on section 2 In section 4, we describe our speech recognition experiments using the proposed method and discuss about the result. Finally, conclusions are drawn in section 5.

2. CONVENTIONAL METHODS

2.1. Feature Extraction by PCA

Principal Component Analysis (PCA) is defined as an orthogonal linear transformation that transforms data to a new coordinate system. This is also usually used for dimensionality reduction and decorrelation of feature coefficients.

The P-dimensional data vector at t-th frame is denoted as $\mathbf{x}_t$ here. The covariance matrix S is derived as follow.

$$S = \frac{1}{N}\sum_{t=1}^{N}(\mathbf{x}_t - \bar{\mathbf{x}})(\mathbf{x}_t - \bar{\mathbf{x}})^T \tag{1}$$

Here $\bar{\mathbf{x}}$ is a mean vector. The eigenvectors that make the new coordinate system are computed by eigenvalue decomposition of the covariance matrix S as follow:

$$S\,\phi_k = \lambda_k \phi_k, (k = 1, 2, \cdots, P) \tag{2}$$

where ϕ_k is an eigenvector corresponding to the eigenvalue λ_k. Selecting $Q(< P)$ eigenvectors corresponding to the Q largest eigenvalues, the new feature vector $\mathbf{y}_t$ is obtained by the following equation.

$$\mathbf{y}_t = \Phi^T(\mathbf{x}_t - \bar{\mathbf{x}}) \tag{3}$$

$$\Phi = (\phi_1, \phi_2, \cdots, \phi_Q) \tag{4}$$

The eigenvalue estimated by PCA means the variance within the data. In clean speech data, the important speech

components for speech recognition have generally large variation. By selecting eigenvectors corresponding to some large eigenvalues to make the new projected space, it is possible to extract only the effective components.

Also, for convolution noise (distortion), PCA-based feature extraction is applied to the log mel-scale filter bank output [1] because we expect that PCA will project the main speech element onto low-order features, while convolution noise (distortion) elements will be projected onto high-order ones. Our recognition results show that the use of PCA instead of DCT provides better performance for distorted speech.

2.2. Feature Extraction by ICA

Independent component analysis is a method for separating a mutual independent source signals from mixed signals. ICA has broad field of application. In [4], ICA was used to speech feature extraction. It is assumed that the observed speech vector $\mathbf{x}$ by short time (ST)-DFT is supposed to be linearly coupled as $\mathbf{x} = A\mathbf{s}$, where A is mixing matrix and $\mathbf{s}$ is source vector, To extract independent components vector $s' = W\mathbf{x}$, we have to estimate W by maximizing the statistical independence of the estimated components. The statistical independence is usually represented by negentropy or kurtosis that is fourth-order cumulant. And maximization of statistical independence is implemented in gradient algorithm or fixed-point algorithm. It is shown in [4] that W obtained by applying ICA to speech data set from single microphone, worked like band-pass filter.

In this paper, we use FastICA [3] that is based on fixed-point iteration scheme using negentropy. The FastICA algorithm for finding one $\mathbf{w}$ that derives one independent component is as follow :

1. Center the data to make its mean zero.
2. Whiten the data to give $\mathbf{z}$.
3. Choose an initial vector $\mathbf{w}$ of unit norm.
4. Let $\mathbf{w} \leftarrow E\{\mathbf{z}g(\mathbf{w}^T\mathbf{z})\} - E\{g'(\mathbf{w}^T\mathbf{z})\}\mathbf{w}$, where g is the function that gives approximations of negentropy.
5. Let $\mathbf{w} \leftarrow \mathbf{w}/\|\mathbf{w}\|$.
6. If not converged, go back to step 4.

To estimate more independent components, different kinds of decorrelation schems should be used; refer to [3] for more information.

3. PROPOSED METHOD

In this paper, we consider a method of incorporating information dealing with the relation between phonemes into feature space. The most commonly used feature space, MFCC, is obtained analytically by applying DCT to Log MFB (Mel-frequency Filter Bank). This space does not contain any information associated with inter-phoneme relationships. Following is an explanation of the proposed method to construct a new feature space (transformation matrix) using PCA and ICA.

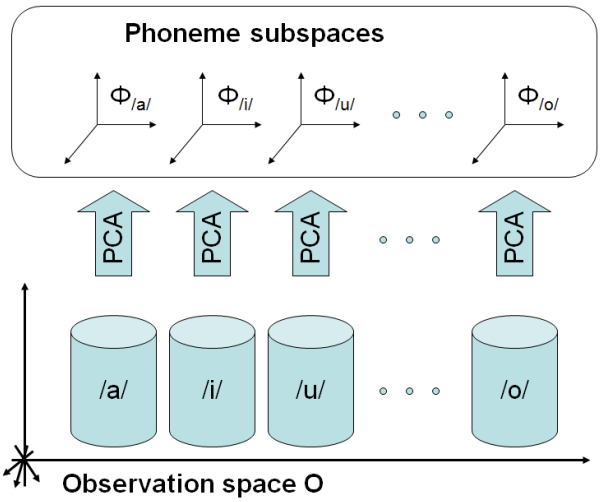

Fig. 1. *Observation space and phoneme subspaces using PCA*

Following is an explanation of the method we used to make the feature space using PCA and ICA. Our previous method [1] involved applying PCA to the set of the whole clean speech, but the proposed method in this paper involves dividing the whole speech data into data sets for each phoneme and then applying PCA to each of these phoneme data sets. Herewith, the structures of the phoneme data are obtained. Subsequently, we make the new feature space by merging these phoneme subspaces using ICA. This new feature space contains the information that is about the correlation among phonemes and statistically mutual independent.

3.1. Phoneme subspaces using PCA

In this subsection, we define phoneme subspaces and feature vector projected onto these subspaces using PCA.

As in (3) and (4), the feature vector $\mathbf{y}_t^i$ projected onto the i-th phoneme subspace Φ^i is defined as follow:

$$\mathbf{y}_t^i = \Phi^{iT}(\mathbf{x}_t - \bar{\mathbf{x}}^i) \tag{5}$$

$$\Phi^i = (\phi_1^i, \phi_2^i, \cdots, \phi_Q^i) \tag{6}$$

Here, we set all phoneme subspace dimensionality as Q to be same as a matter of convenience. And, we define V (the matrix of the whole phoneme subspace) and C (the mean vector of the whole phoneme) as follows:

$$\begin{aligned} V &= [\Phi^1, \Phi^2, \cdots, \Phi^M] \\ C &= [(\Phi^{1T}\bar{\mathbf{x}}^1)^T, (\Phi^{2T}\bar{\mathbf{x}}^2)^T, \cdots, (\Phi^{MT}\bar{\mathbf{x}}^M)^T] \end{aligned} \tag{7}$$

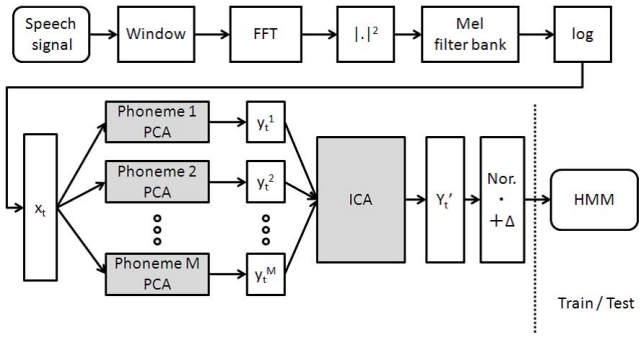

Fig. 2. *Speech feature extraction process*

M means the number of phonemes. Finally, super vector $\mathbf{y}_t$ is obtained by concatenating $\mathbf{y}_t^i$ as follow:

$$\begin{aligned}\mathbf{y}_t = \begin{bmatrix}\mathbf{y}_t^1\\ \mathbf{y}_t^2\\ \vdots\\ \mathbf{y}_t^M\end{bmatrix} &= \begin{bmatrix}\Phi^{1T}[\mathbf{x}_t - \bar{\mathbf{x}}^1]\\ \Phi^{2T}[\mathbf{x}_t - \bar{\mathbf{x}}^2]\\ \vdots\\ \Phi^{MT}[\mathbf{x}_t - \bar{\mathbf{x}}^M]\end{bmatrix}\\ &= \begin{bmatrix}\Phi^{1T}\mathbf{x}_t - \Phi^{1T}\bar{\mathbf{x}}^1\\ \Phi^{2T}\mathbf{x}_t - \Phi^{2T}\bar{\mathbf{x}}^2\\ \vdots\\ \Phi^{MT}\mathbf{x}_t - \Phi^{MT}\bar{\mathbf{x}}^M\end{bmatrix}\\ &= V^T\mathbf{x}_t - C^T\end{aligned} \tag{8}$$

Fig. 1 shows the basic concepts of phoneme subspaces. From observation space, we derive M phoneme subspaces.

3.2. Integration of Phoneme subspaces using ICA

The super vector, $\mathbf{y}_t$ defined in the preceding subsection has a very large dimension ($M \times Q$) and many elements with similar trend. To compress the dimensionality and extract correlation information among phonemes, FastICA is applied to the set of super vectors $\mathbf{y}$. Let V' as the transformation matrix that integrates phoneme subspaces obtained by ICA. Our proposed speech feature vector $\mathbf{y}$ is generated as follow:

$$\mathbf{y}_t' = V'\mathbf{y}_t = V'(V^T\mathbf{x}_t - C^T). \tag{9}$$

Fig. 2 shows the process to obtain proposed speech feature vector $\mathbf{y}_t'$ from the speech signal. $\mathbf{y}_t'$ is normalized and time derivatives are added to input HMM for training and test.

4. ISOLATED WORD RECOGNITION EXPERIMENTS

4.1. Experiment conditions

In order to confirm the efficiency of the proposed method, the speech data were extracted from the A-set of the ATR Japanese database and the room impulse response was extracted from the RWCP sound scene database [5]. The total number of speakers was four (2 males and 2 females). The training data was composed of 2,620 utterances per speaker, and 1,000 clean or reverberant utterances made by convolving impulse responses were used for testing per speaker. Speech signals were digitized into 16 bits at a sampling frequency of 12 kHz. For spectral analysis, an ST-DFT was performed on 32-ms windowed and 8-ms shifted frames. Next, a 32-channel mel-frequency filter bank (MFB) analysis was performed on the above components. The logarithms of MFB components were then computed. The experiments were conducted to compare MFCC, PCA, PCA-PCA (integrating subspaces by PCA), and PCA-ICA (integrating subspaces by ICA; proposed method) with mean normalized coefficients (16 Dim. + Δ 16 Dim.). These analyses were realized by using HTK toolkits[6]. The models of 54 context-independent phonemes were trained by using four sets of 2,620 clean words spoken by four speakers respectively to construct common HMMs. Each HMM has three states and three self-loops, and each state has four Gaussian mixture components. Table 1 shows the conditions for estimating the projection matrices V, V', and C in (7) or (9).

Table 1. *The number of frames used to calculate subspace and the dimension*

Transformation	Frames	Dimension
LogMFB	–	32
DCT (MFCC)	–	16
Φ *(PCA)*	4000	16
V (phoneme subspaces by PCA)	54×100	54×16
V' (integrated subspace by ICA)	4000	16

4.2. Results

Fig. 3 shows the results of isolated word speech recognition. The recognition rate means the average of the four speakers. As the reverberation time lengthens, the recognition rate declines. Especially the rate decrease sharply over 300ms reverberation time. It is shown that the proposed method outperforms MFCC in all reverberant conditions. However, our method shows somewhat lower performance than PCA and PCA-PCA in clean or short reverberant conditions. We think this result was given by two main reason. Firstly, the phoneme subspaces were not optimized because the dimensions of phoneme subspaces were set to be equal. It is thinkable that the integrated feature subspace was affected by undesired subspaces. Secondly, in this experiment, we used whole proposed feature vector as input to simple HMMs. Each element of proposed feature vector ranges independently of each other. So simple HMMs is not expressive for this feature vector. Therefore we are going to investigate how to find opti-

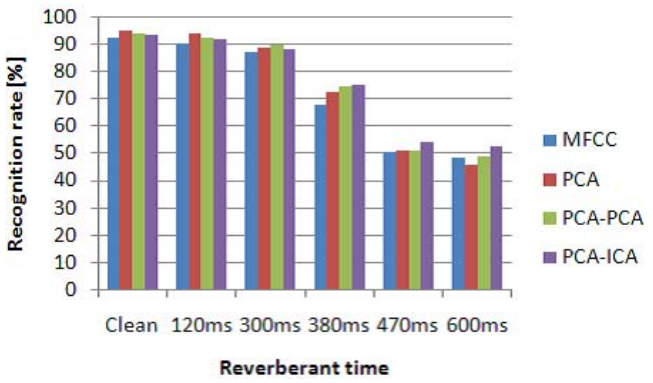

Fig. 3. *The results of isolated word speech. Ave. for 4 speakers.*

mal dimensionality and how to make acoustic model based on HMM using independent components by reference to [7] and [8]

5. CONCLUSIONS

We proposed the new speech feature extraction method which emphasizes the phonetic information from observed speech using PCA and ICA. The proposed method is to extract speech feature that contains the correlated information among phoneme subspace and their elements are statistically mutual independent. The experiment results on isolated word recognition under clean and 360-ms reverberant conditions show that the proposed method outperforms conventional method using MFCC. However our method shows somewhat lower performance than PCA and PCA-PCA. Our next step is to study how to make the optimized phoneme-subspaces and how to make acoustic model based on HMM using independent components by our methods. The proposed method can be combined with other methods, such as speech signal processing or model adaptation, to improve the recognition accuracy in real-life environments.

6. REFERENCES

[1] T. Takiguchi and Y. Ariki, "Robust Feature Extraction Using Kernel PCA," *Proc. ICASSP2006*, pp. 509-512, 2006.

[2] K. Hermus, P. Wambacq, and H. V. Hamme, "A Review of Signal Subspace Speech Enhancement and Its Application to Noise Robust Speech Recognition," *Journal on Advances in Signal Processing*, vol. 2007, Article ID 45821, 15 pages, 2007.

[3] A. Hyvärinen and E. Oja, "Independent Component Analysis: Algorithms and Applications," *Neural Networks*, vol. 13(4-5), pp. 411-430, 2000.

[4] OW. Kwon, TW. Lee, "Phoneme recognition using ICA-based feature extraction and transformation," *Signal Processing*, vol. 84(6), pp. 1005-1019, 2004.

[5] S. Nakamura, K. Hiyane, F. Asano, T. Nishimura and T. Yamada, "Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition," *Proc. LREC2000*, Vol. 2, pp. 965-968, 2000.

[6] S. Young et. al., *The HTK Book*, Entropic Labs and Cambridge University, 1995-2002.

[7] R. Vetter, N. Virag, P. Renevey and J.-M. Vesin, "Single Channel Speech Enhancement Using Principal Component Analysis and MDL Subspace Selection," *Proc. Eurospeech99*, pp. 2411-2414, 1999.

[8] C-W. Ting and J-T. Chien, "Factor Analysis of Acoustic Features for Streamed Hidden Markov Modeling," *Proc. ASRU2007*, pp. 30-35, 2007

[9] H. Hermansky and N. Morgan, "RASTA Processing of Speech," *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, pp. 578-589, 1994.

[10] C. Avendano, S. Tivrewala, and H. Hermansky, "Multiresolution channel normalization for ASR in reverberant environments," *Proc. Eurospeech1997*, pp. 1107-1110, 1997.

[11] W. Li, K. itou, K. Takeda and F. Itakura, "Two-Stage Noise Spectra Estimation and Regression Based In-Car Speech Recognition Using Single Distant Microphone," *Proc. ICASSP2005*, pp. 533-536, 2005.

[12] K. Kinoshita, T. Nakatani and M. Miyoshi, "Efficient Blind Dereveration Framework for Automatic Speech Recognition," *Proc. Interspeech2005*, pp. 3145-3148, 2005.

[13] R. Vetter, N. Virag, P. Renevey and J.-M. Vesin, "Single Channel Speech Enhancement Using Principal Component Analysis and MDL Subspace Selection," *Proc. Eurospeech99*, pp. 2411-2414, 1999.

[14] S-M. Lee, S-H. Fang, J-W. Hung and L-S. Lee, "Improved MFCC Feature Extraction by PCA-Optimized Filter Bank for Speech Recognition," *Proc. ASRU2001*, pp. 49-52, 2001.

[15] F. Asano, Y. Motomura, H. Asoh and T. Matsui, "Effect of PCA Filter in Blind Source Separation," *Proc. ICA2000*, pp. 57-62, 2000.