# PCA-Based Feature Extraction for Fluctuation in Speaking Style of Articulation Disorders

*Hironori Matsumasa*[1]*,Tetsuya Takiguchi*[1]*, Yasuo Ariki*[1]*, Ichao LI*[2]*, Toshitaka Nakabayashi*[3]

[1)]Department of Computer and System Engineering
Kobe University, 1-1 Rokkodai, Nada, Kobe, 657-8501, JAPAN
[2)]Department of Economics
Otemon Gakuin University, 2-1-15, Nishiai, Ibaraki,Osaka, 567-8502, JAPAN
[3)]Department of Human Development
Kobe University, 3-11, Tsurukabuto, Nada, Kobe, 657-8501, JAPAN
mattu28@me.cs.scitec.kobe-u.ac.jp takigu@kobe-u.ac.jp ariki@kobe-u.ac.jp

10.21437/Interspeech.2007-374

## Abstract

We investigated the speech recognition of a person with articulation disorders resulting from athetoid cerebral palsy. Recently, the accuracy of speaker-independent speech recognition has been remarkably improved by the use of stochastic modeling of speech. However, the use of those acoustic models causes degradation of speech recognition for a person with different speech styles (e.g., articulation disorders). In this paper, we discuss our efforts to build an acoustic model for a person with articulation disorders. The articulation of the first speech tends to become unstable due to strain on muscles and that causes degradation of speech recognition. Therefore, we propose a robust feature extraction method based on PCA (Principal Component Analysis) instead of MFCC. Its effectiveness is confirmed by word recognition experiments.

**Index Terms**: articulation disorders, PCA, feature extraction

## 1. Introduction

Recently, the importance of information technology in the welfare-related fields has increased. For example, sign language recognition using image recognition technology [1], text reading systems from natural scene images [2], and the design of wearable speech synthesizers for voice disorders [3][4] have been studied.

As for speech recognition technology, the opportunities in various environments and situations have increased (e.g., operation of a car navigation system, lecture transcription into a document style in a meeting). However, the degradation can be observed in children [5], persons with a speech impediment, and so on, and there has been very little research on orally-challenged people, such as those with speech impediments. There are 34,000 people with speech impediments associated with articulation disorders, in Japan alone, and it is hoped that speech recognition systems will one day be able to recognize their voices.

One of the causes of a speech impediments is cerebral palsy. It occurs at a ratio of about 2 per 1,000 babies with cerebral palsy. Cerebral palsy results from damage to the central nervous system, and the damage causes movement disorders. Three general times are given for the onset of the disorder: before birth, at the time of delivery, and after birth. Cerebral palsy is classified as follows: 1) spastic type 2) athetoid type 3) ataxic type 4) atonic type 5) rigid type, and, a mixture of types [6].

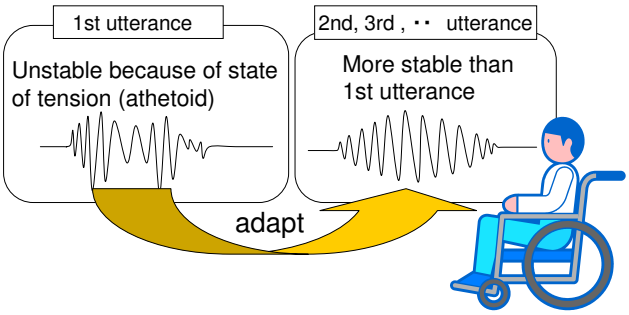


Figure 1: Corrective strategy for articulation disorders

In this paper, we focused on a person with an articulation disorder resulting from the athetoid type of cerebral palsy. Athetoid symptoms develop in about 10-15% of a cerebral palsy sufferers. In the case of a person with this type of articulation disorder, the first movements are sometimes more unstable than usual. That means, the case of movements related to speaking, the first utterance is often unstable or unclear due to the athetoid symptoms. Therefore, we recorded speech data for a person with a speech impediment who uttered a given word several times, and we investigated the influence of the unstable speaking style caused by the athetoid symptoms.

In current speech recognition technology, MFCC (Mel Frequency Cepstral Coefficient) has been widely used. The feature is derived from the mel-scale filter bank output by DCT (Discrete Cosine Transform). The low-order MFCCs account for the slowly changing spectral envelope, while the high-order ones describe the fast variations of the spectrum. Therefore a large number of MFCCs is not used for speech recognition because we are only interested in the spectral envelope, not in the fine structure.

PCA-based feature extraction has been studied [7]. We investigated applying kernel PCA to reverberant speech [8]. In this paper, we propose robust feature extraction based on PCA with more stable utterance data instead of DCT, where the main stable utterance element is projected onto low-order features, while fluctuation elements of speech style are projected onto high-order ones. Therefore, we can approximate the first utterance using more stable utterances (Fig.1). Its effectiveness is confirmed by word recognition experiments on first utterances.

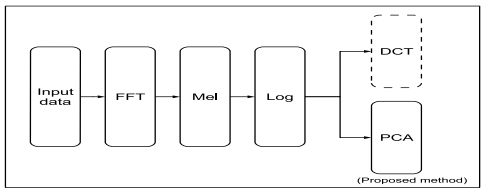


Figure 2: Feature extraction using PCA

## 2. Feature Extraction Using PCA

In this paper we investigate robust feature extraction using PCA with the more stable utterance data instead of DCT, where PCA is applied to the mel-scale filter bank output (Fig.2). We computed the filter (eigenvector matrix) using the more stable utterance. Then we applied the filtering operation to the first utterance (unstably articulated utterance) in the log-spectral domain.

Given the frame of short-time analysis $n$ and frequency $\omega$, we represent the first utterance $X_n(\omega)$ as the multiplication of the stable speech $S_n(\omega)$ and the fluctuation element of speaking style $H(\omega)$ in the linear-spectral domain:

$$X_n(\omega) = S_n(\omega) \cdot H(\omega) \tag{1}$$

The multiplication can be converted to addition in the log-spectral domain as follows:

$$\log X_n(\omega) = \log S_n(\omega) + \log H(\omega) \tag{2}$$

Next, we consider the following filtering based on PCA in order to extract the feature of stable speech only,

$$\hat{\boldsymbol{S}} = \boldsymbol{V^t X_{log}}. \tag{3}$$

For the filter (eigenvector matrix), $V$ is derived by the eigenvalue decomposition of the centered covariance matrix of a stable speech data set, in which the filter consists of the eigenvectors corresponding to the $L$ dominant eigenvalues,

$$\boldsymbol{V} = (\boldsymbol{v_1}, \cdots, \boldsymbol{v_L}). \tag{4}$$

Here $v$ is orthonormal basis ($M$ dimension). From the orthogonality, we use

$$\boldsymbol{V^t V} = \boldsymbol{I}, \tag{5}$$

where $I$ is the $L$ dimension identity matrix. Given the mean in the filtered space $m$ and in the unfiltered space $\tilde{m}$ and a set of filtered data $y$ obtained in equation (3) $\mathcal{Y}$ ($y$ corresponding to $\hat{\boldsymbol{S}}$ in equation (3)),

$$\begin{aligned} \tilde{\boldsymbol{m}} &= \frac{1}{n} \sum_{\boldsymbol{y} \in \mathcal{Y}} \boldsymbol{y} \\ &= \boldsymbol{V^t m}. \end{aligned} \tag{6}$$

Given the stable speech $s$ set $\mathcal{S}$, variance in the filtered space is represented as follows.

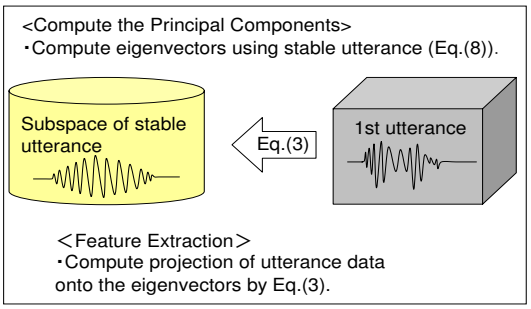


Figure 3: The calculation procedure of PCA

$$\begin{aligned} \tilde{\sigma}^2(\boldsymbol{V}) &= \frac{1}{n} \sum_{\boldsymbol{y} \in \mathcal{Y}} (\boldsymbol{y} - \tilde{\boldsymbol{m}})^t (\boldsymbol{y} - \tilde{\boldsymbol{m}}) \\ &= \frac{1}{n} \sum_{\boldsymbol{s} \in \mathcal{S}} tr(\boldsymbol{V}^t(\boldsymbol{s} - \boldsymbol{m})(\boldsymbol{V}^t(\boldsymbol{s} - \boldsymbol{m}))^t) \\ &= tr(\boldsymbol{V}^t \frac{1}{n} \sum_{\boldsymbol{s} \in \mathcal{S}} ((\boldsymbol{s} - \boldsymbol{m})(\boldsymbol{s} - \boldsymbol{m})^t) \boldsymbol{V}) \\ &= tr(\boldsymbol{V}^t \sum \boldsymbol{V}) \end{aligned} \tag{7}$$

Here $\sum$ represents the covariance matrix in the original (unfiltered) space. In PCA, the new basis is estimated by the maximization of the covariance in the projected (filtered) space, and that optimal solution results in the eigenvalue decomposition shown in equation (8).

$$\boldsymbol{V}^t \sum \boldsymbol{V} = \boldsymbol{\Lambda}. \tag{8}$$

Matrix $V$ converts $\sum$ into diagonalization.

We use $L$ eigenvectors corresponding to the biggest $L$ eigenvalues. Due to the orthogonality, the component of the convolution fluctuation element of speaking style belonging to the subspace $[\boldsymbol{v_{L+1}}, \ldots, \boldsymbol{v_M}]$ is canceled by this filtering operation. The procedure of the feature extraction is summarized in Fig.3.

## 3. Recognition Experiment

### 3.1. Experimental Conditions

The new feature extraction method was evaluated on word recognition tasks for one person with an articulation disorder. We recorded 210 words included in the ATR Japanese speech database repeating each word five times (Fig.4). The utterance signal was sampled at 16 kHz and windowed with a 25-msec Hamming window every 10 msec. Then we clipped each utterance by hand. Figure 5 shows an example of a sound wave of a person with an articulation disorder. Figure 6 shows a sound wave of a physically unimpaired person doing the same task. We used HTK [9] for all the experiments.

### 3.2. Recognition of Speaker-Independent Model

At the beginning, we attempted to recognize utterances using a speaker-independent model for an unimpaired person (this model is included in Julius [10]). The acoustic model consists of a monophone HMM set with 24 dimensional MFCC features (12-order MFCCs and their delta) and 16 mixture components for each state. Each HMM has three states and three

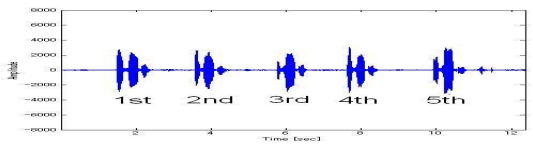


Figure 4: Example of recorded speech data

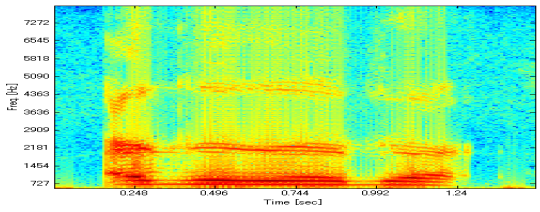


Figure 5: Example of a spectrogram spoken by a person with an articulation disorder //a k e g a t a

self-loops. Figure 7 shows the recognition rates using a speaker-independent model. In a person with an articulation disorder, a recognition rate of only 3.5% was obtained, but in a physically unimpaired person, a recognition rate of 89.7% was obtained for the same task. It is clear that the speaking style of a person with an articulation disorder differs considerably from that of a physically unimpaired person.

We carried out speaker adaptation for improvement of recognition rates using a speaker-independent model. We used MLLR (Maximum Likelihood Linear Regression) and MAP (Maximum A Posteriori) estimation for the method of speaker adaptation [11]. The second utterance was used for adaptation. We carried out recognition experiments on adaptation data sets of 20, 40, 60, 100, and 210 words.

Figure 8 shows the recognition rates using the speaker adaptation model. When the volume of adaptation data was small, little improvement in recognition rate was observed. When 210 words were used, however, the recognition rate rose significantly. A large amount of adaptation data is necessary to realize an improvement in recognition rate when using an acoustic model of an unimpaired person.

### 3.3. Recognition of Speaker Dependent Model

It was difficult to recognize utterance using an acoustic model trained by utterance of a physically unimpaired person. Therefore, we trained the acoustic model using the utterance of a person with an articulation disorder. The acoustic model consists of a HMM set with 54 context-independent phonemes with 24 dimensional MFCC features (12-order MFCCs and their delta) and 6 mixture components for each state. Each HMM has three states and three self-loops. When we attempted to recognize the 1st utterance, we used 2nd-5th utterance for training. We iterated this process for each utterance. Figure 9 shows the recognition rates using the speaker-dependent model.

As can be seen from this figure, the use of a speaker-dependent model for utterances of a person with an articulation disorder improves the recognition rates from 3.5% to 87.2%. Table 1 shows the recognition rates of each utterance in a person with an articulation disorder.

In a person with an articulation disorders, the recognition rates of the 1st utterance is 77.1%. It is lower than the others. The first utterance is the first intentional movement. It is conjectured that he experiences a more strained state during the first utterance compared to subsequent utterances. So, athetoid symptoms occur and articulation becomes difficult. It is believed that this difficulty causes fluctuations in speaking style and degradation of the recognition rates.

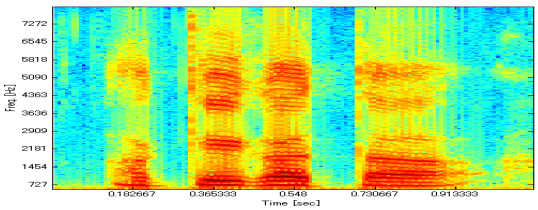


Figure 6: Example of a spectrogram spoken by a physically unimpaired person //a k e g a t a

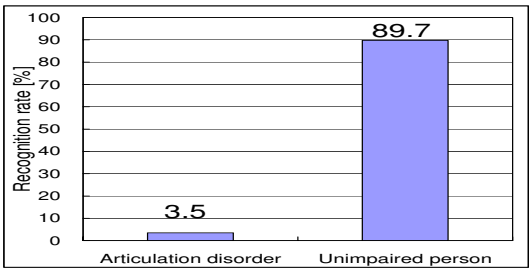


Figure 7: Recognition results for a speaker-independent model for an unimpaired person

Table 1: Recognition rate for each utterance (articulation disorder)

| 1st | 2nd | 3rd | 4th | 5th |
|---|---|---|---|---|
| 77.1 | 89.1 | 91.4 | 91.0 | 87.6 |

### 3.4. Recognition of Speaker Dependent Model using proposed method

In the new feature extraction, PCA was applied to 24 mel-scale filter bank output, and then the delta coefficients were also computed. We experimented on the number of principal components, using 11, 13, 15, 17, and 19 dimensions. Figure 10 shows the recognition rates for the 1st utterance. Figure 11 shows the recognition rates of the average of each number of dimensions. Figure 12 shows the recognition rates of 17 principal components. As can be seen from these Figures, the use of PCA instead of DCT improves the recognition rates for the 1st utterance from 79.1% to 85.2%. These results clearly show that the use of PCA instead of DCT achieves good performance when dealing with a 1st utterance. In addition, the recognition rates of the other utterances were equal to MFCC in recognition.

## 4. Summary

This paper has described a robust feature extraction technique using PCA instead of DCT, where PCA is applied to the mel-scale filter bank output. It can be expected that PCA will project the main stable utterance elements onto low-order features, while elements associated with fluctuations in speaking

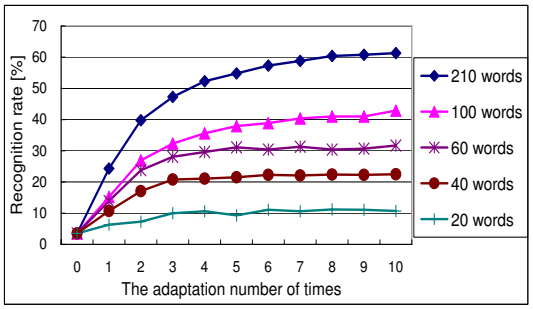


Figure 8: Adaptation result by MLLR and MAP estimation

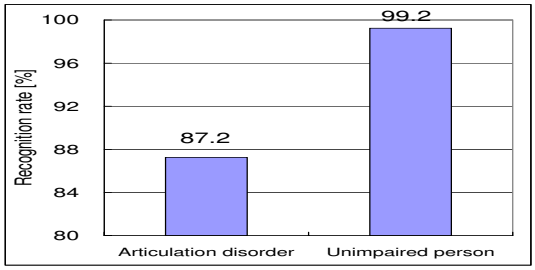


Figure 9: Recognition results for the speaker-dependent model

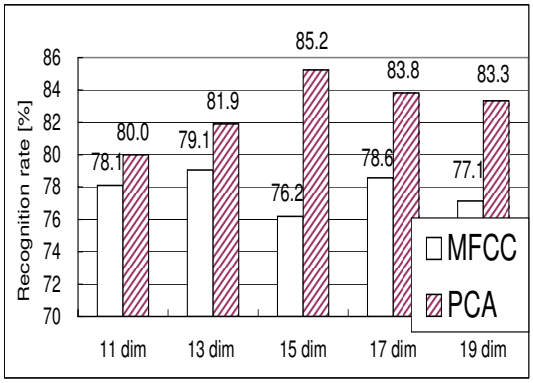


Figure 10: Recognition rate for the 1st utterance by the proposed method

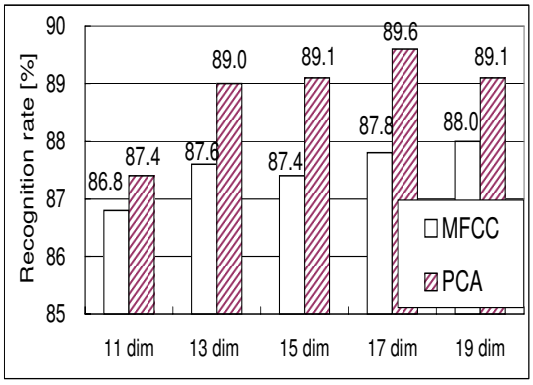


Figure 11: Recognition rate for each dimension by the proposed method

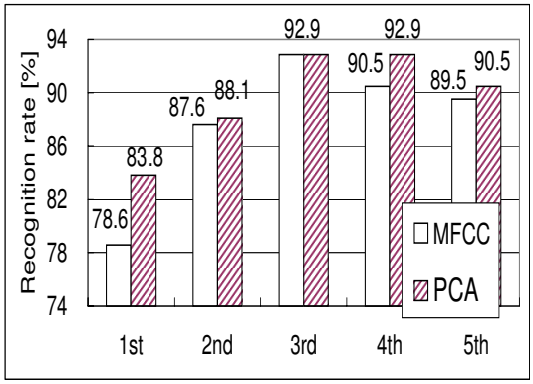


Figure 12: Recognition rate for each utterance by the proposed method (17 dimensions)

style will be projected onto high-order features. The proposed method resulted in an improvement of 6.1% in the recognition rate compared to the conventional method, MFCC. In this study, there was only one object person, so in future experiments, we will increase the number of object persons and examine the effectiveness of proposed method.

## 5. References


[1] J. Lin and W. Ying and T.S. Huang, "Capturing human hand motion in image sequences," IEEE Motion and Video Computing Workshop, pp. 99–104, 2002.

[2] M.K. Bashar, T. Matsumoto, Y. Takeuchi, H. Kudo and N. Ohnishi, "Unsupervised Texture Segmentation via Wavelet-based Locally Orderless Images (WLOIs) and SOM," 6th IASTED International Conference COMPUTER GRAPHICS AND IMAGING, 2003.

[3] T. Ohsuga and Y. Horiuchi and A. Ichikawa, "Estimating Syntactic Structure from Prosody in Japanese Speech," IEICE Transactions on Information and Systems, 86(3), pp. 558–564, 2003.

[4] K. Nakamura and T. Toda and H. Saruwatari and K. Shikano, "Speaking Aid System for Total Laryngectomees Using Voice Conversion of Body Transmitted Artificial Speech," INTERSPEECH-2006, pp. 1395–1398, 2006.

[5] D. Giuliani and M. Gerosa, "Investigating recognition of children's speech," ICASSP2003, pp. 137–140, 2003.

[6] S.T. Canale and W.C. Campbell, "Campbell's Operative Orthopaedics," Mosby-Year Book, 2002.

[7] S-M. Lee and S-H. Fang and J-W. Hung and L-S. Lee, "Improved MFCC Feature Extraction by PCA-Optimized Filter Bank for Speech Recognition," Automatic Speech Recognition and Understanding, ASRU, pp. 49–52, 2001.

[8] T. Takiguchi and Y. Ariki, "Robust Feature Extraction Using Kernel PCA," ICASSP2006, pp.509–512, 2006.

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

[10] "http://julius.sourceforge.jp/index.htm,"

[11] V.V. Digalakis and L.G. Neumeyer, "Speaker Adaptation Using Combined Transformation and Bayesian Methods", IEEE Trans. Speech and Audio Proc, 4(4), pp. 294–300, 1996.