interspeech 2008 incorporating SST 08

Integration of Metamodel and Acoustic Model for Speech Recognition

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

[1)]Graduate School of 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

Abstract

We investigated the speech recognition of a person with articulation disorders resulting from athetoid cerebral palsy. The articulation of the first speech tends to become unstable due to strain on speech-related muscles, and that causes degradation of speech recognition. Therefore, we proposed a robust feature extraction method based on PCA (Principal Component Analysis) instead of MFCC [1]. In this paper, we discuss our effort to integrate a Metamodel [2] and Acoustic model approach. Metamodel has a technique for incorporating a model of a speaker's confusion matrix into the ASR process in such a way as to increase recognition accuracy. Its effectiveness has been confirmed by word recognition experiments.

Index Terms: articulation disorders, PCA, feature extraction, model integration, dysarthric speech

1. Introduction

Recently, the importance of information technology in the welfare-related fields has increased. For example, sign language recognition using image recognition technology [3], text reading systems from natural scene images [4], and the design of wearable speech synthesizers for those with voice disorders [5][6] 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 during meetings, etc.). However, degradation can be observed in the case of children [7], 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 speech impediments is cerebral palsy. About 2 babies in 1,000 are born 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 [8].

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 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, the MFCC (Mel Frequency Cepstral Coefficient) has been widely used, where the feature is derived from the mel-scale filter bank output by DCT (Discrete Cosine Transform). In [9], PCA-based feature extraction has been studied. Also, we investigated applying kernel PCA to reverberant speech [10], and we proposed robust feature extraction based on PCA with more stable utterance data instead of DCT [1], 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, the PCA-based filter will be able to extract stable utterance features only (Fig.1). Our proposed method improved the recognition accuracy, but the performance was not sufficient compared to that of persons with no disability.

The speech associated with articulation disorders may decrease intelligibility due to substitutions, deletions and insertions of phonemes. In [2, 3], a metamodel was introduced to increase recognition accuracy in speakers with low intelligibility by using the phoneme confusion matrix. However, the metamodel itself was effective only when there is limited adaptation data available for a speaker.

In this paper, we introduce a technique that combines acoustic models and metamodels that generate another word hypothesis for an utterance, where acoustic models are built by using a large amount of training data uttered by a person with an articulation disorder. Its effectiveness is confirmed by word recognition experiments on speech data of a person with an articulation disorder, by comparing our previous method (PCA-based feature extraction) and metamodels only.

2. Metamodel [2, 3]

In [2, 3], a metamodel was introduced to generate word hypotheses for an utterance using an unconstrained phoneme recognizer. For word recognition, we attempt to find the word $\acute{w}$ for which the probability $Pr(w|\boldsymbol{X})$ is the largest among all

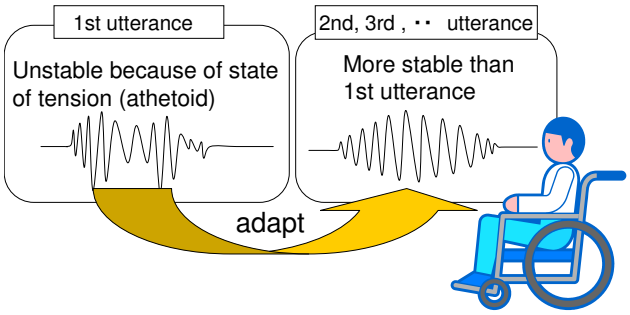

Figure 1: Corrective strategy for articulation disorders

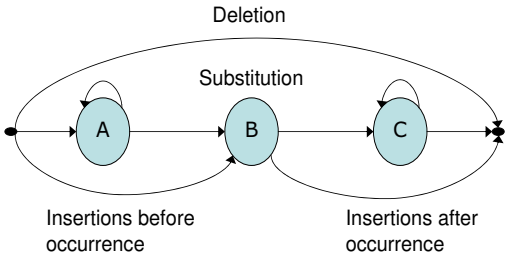

Figure 2: Architecture of a phoneme metamodel (discrete HMM)

words $w \in \mathcal{W}$. Here $\boldsymbol{X}$ is the acoustic signal input. Since the phoneme sequences form a proper subset $p \in \boldsymbol{P}$, we may rewrite $Pr(w|\boldsymbol{X})$ as

$$Pr(w|\boldsymbol{X}) = \sum_{p \in \mathcal{P}} Pr(w|p)Pr(p|\boldsymbol{X}). \tag{1}$$

Instead of choosing this subset of $\boldsymbol{P}$, we may choose the sequence $\boldsymbol{p}^*$ obtained from a phoneme classification task:

$$\boldsymbol{p}^* = \arg\max_{p \in \mathcal{P}} Pr(p|\boldsymbol{X}). \tag{2}$$

Thus, a different approximation of $Pr(w|\boldsymbol{X})$ gives

$$Pr(w|\boldsymbol{X}) \simeq Pr(w|\boldsymbol{p}^*)Pr(\boldsymbol{p}^*|\boldsymbol{X}). \tag{3}$$

The knowledge of $\boldsymbol{p}^*$ from a phoneme-recognition experiment may be used to obtain an alternative decoding of the utterance by finding the word sequence that maximizes each side:

$$\acute{w} = \arg\max_{w \in \mathcal{W}} Pr(w|\boldsymbol{p}^*). \tag{4}$$

In [2, 3], a discrete HMM is used for a phoneme metamodel, as shown in Fig.2. Each state of a metamodel has a discrete probability distribution over the symbols for the set of phonemes. The central state B of a metamodel for a certain phoneme model takes account of correct decodings or substitutions. States A and C model (possibly multiple) insertions before and after the phoneme. The metamodels are trained in the same manner as acoustic HMMs using embedded Baum-Welch re-estimation.

3. Proposed method

3.1. Feature extraction using PCA

We proposed robust feature extraction using PCA with the more stable utterance data instead of DCT (Fig. 3), where PCA is

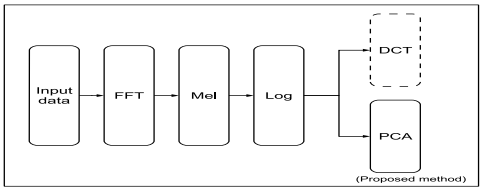

Figure 3: Feature extraction using PCA

applied to the mel-scale filter bank output [1]. 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 ω, 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:

$$X_n(\omega) = S_n(\omega) \cdot H(\omega). \tag{5}$$

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

$$\log X_n(\omega) = \log S_n(\omega) + \log H(\omega). \tag{6}$$

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

$$\hat{\boldsymbol{S}} = \boldsymbol{V}^t \boldsymbol{X}_{\boldsymbol{log}}. \tag{7}$$

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.

3.2. Integration of metamodel and acoustic model

The articulation of speech uttered by persons with speech disorders tends to become unstable due to strain on their speech-related muscles. Therefore, the fluctuation of speaking style may invoke phone fluctuations such as substitutions, deletions and insertions. Metamodels are used for speaker adaptation of a person with articulation disorders in [2]. In this paper, we integrate metamodels and acoustic models for suppressing fluctuation. Figure 4 shows a schematic of the recognition system. The metamodel-acoustic integration enables fluctuation suppression not only in feature extraction but also in recognition. The integration in recognition is represented as follows:

$$\begin{aligned} & L_{Aco+Meta}^{\acute{w}_{N-best}} \\ & = (1-\alpha) \cdot L_{Aco}^{\acute{w}_{N-best}} + \alpha \cdot L_{Meta}^{\acute{w}_{N-best}} \\ & = (1-\alpha) \cdot Pr(\boldsymbol{A}|\acute{w}_{N-best}) + \alpha \cdot Pr(\boldsymbol{p}^*|\boldsymbol{\acute{w}_{N-best}}) \end{aligned} \tag{8}$$

Here L_{Aco} and L_{Meta} represent acoustic likelihood and metamodel likelihood, respectively, and α is weight. As shown in Fig. 5, we perform metamodel recognition for only N-best words $\acute{w}_{N-best}$ obtained by word recognition. Then, we integrate the likelihoods according to (8).

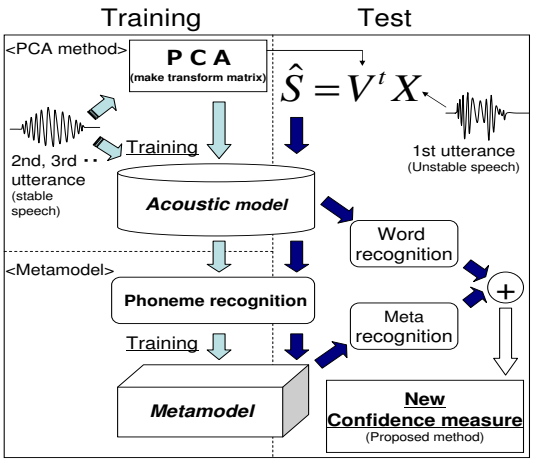

Figure 4: Integration of Metamodel and Acoustic model

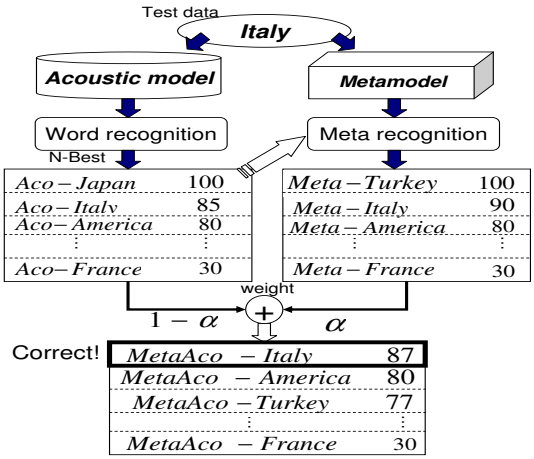

Figure 5: Example of integrated recognition

4. Recognition experiment

4.1. Experimental conditions

The new model integration 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. 6). 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. When we recognize the 1st utterance, the 2nd-5th utterances are used for training. We iterated this process for each utterance. Figure 7 shows an example of a spectrogram spoken by a person with an articulation disorder. Figure 8 shows a spectrogram spoken by a physically unimpaired person doing the same task. We used HTK [11] for all the experiments.

4.2. Recognition using only acoustic models for articulation disorder

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

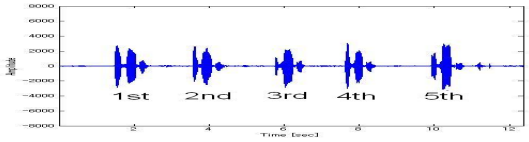

Figure 6: Example of recorded speech data

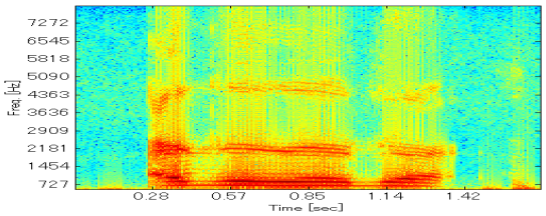

Figure 7: Example of a spectrogram spoken by a person with an articulation disorder //a k e g a t a

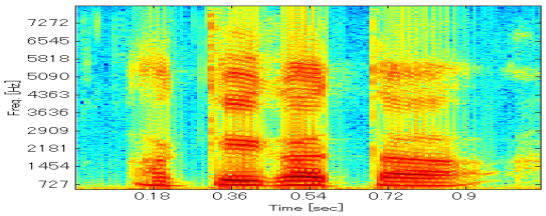

Figure 8: Example of a spectrogram spoken by a physically unimpaired person //a k e g a t a

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.

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

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

In a person with an articulation disorder, the recognition rate of the 1st utterance is 77.1%. As shown in Table 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.

4.3. Results using PCA-based feature extraction

For the 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 9 shows the recognition rates for the 1st utterance. As can be seen from Fig. 9, the use of PCA instead of DCT improves the recognition rates for the 1st utterance from 79.1% to 85.2% (13-order

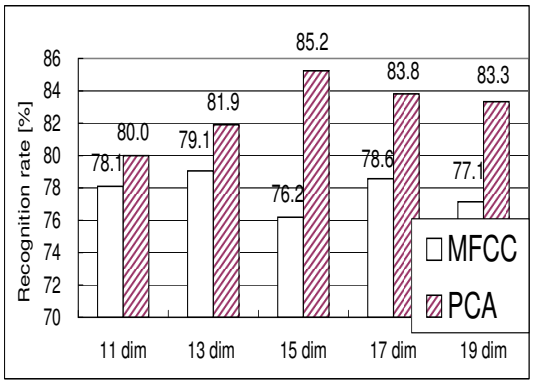

Figure 9: Recognition rates for the 1st utterance by PCA method

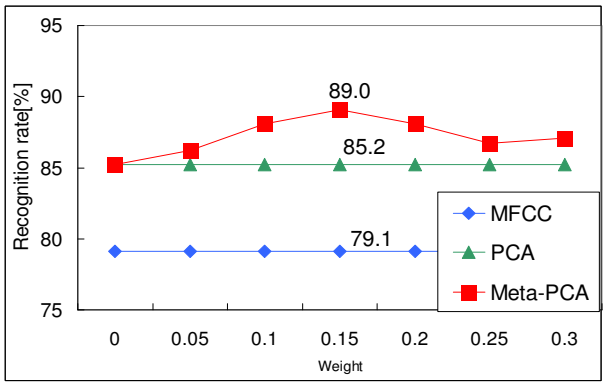

Figure 10: Recognition rates for the 1st utterance by the proposed method

MFCC and 15-principal components). These results clearly show that the use of PCA achieves better performance than DCT when dealing with a 1st utterance. In addition, the recognition rates of the other utterances were equal to MFCC in recognition.

4.4. Results using integration of metamodel and acoustic model

We integrated metamodels and acoustic models by using three-best words. The weight was changed from 0 to 0.3. Figure 10 shows the recognition rates for the 1st utterance. As can be seen from Fig. 10, the use of metamodels and acoustic models improves the recognition rate from 79.1% to 89.0% with the weight 0.15. These results clearly show that the use of integration achieves good performance. It can be expected that integration will decrease substitutions, deletions and insertions caused by phone fluctuations that are not taken account in the feature extraction.

5. Summary

The articulation of speech uttered by persons with speech disorders tends to become unstable due to strain on their speech-related muscles. This paper has described a robust PCA-based feature extraction and integration of metamodels and acoustic models. In the feature extraction, 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 style will be projected onto high-order features.

The fluctuation of speaking style may invoke phone fluctuations such as substitutions, deletions and insertions. The integration of metamodels and acoustic models enables fluctuation suppression in the recognition process, where a phoneme metamodel takes account of correct decodings or substitutions.

The proposed method resulted in an improvement of 9.9% (from 79.1% to 89%) in the recognition rate compared to the conventional method (MFCC-based acoustic models only). In this study, there was only one subject person, so in future experiments, we will increase the number of subjects and further examine the effectiveness of the proposed method.

6. References

[1] H. Matsumasa and T. Takiguchi and Y. Ariki and I. LI and T. Nakabayashi, “PCA-Based Feature Extraction for Fluctuation in Speaking Style of Articulation Disorders,” INTERSPEECH-2007, pp. 1150–1153, 2007.

[2] O. C. Morales and S. Cox, “Modelling Confusion Matrices to Improve Speech Recognition Accuracy, with an Application to Dysarthric Speech,” INTERSPEECH-2007, pp. 1565–1568, 2007.

[3] S. Cox and S. Dasmahapatra, “High-Level Approaches to Confidence Estimation in Speech Recognition,” IEEE Trans. on SAP, vol. 10, No. 7, pp. 460-471, 2002.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] D. Giuliani and M. Gerosa, “Investigating recognition of children’s speech,” ICASSP2003, pp. 137–140, 2003.

[8] S. T. Canale and W. C. Campbell, “Campbell’s Operative Orthopaedics,” Mosby-Year Book, 2002.

[9] 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.

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

[11] S. Young et. al., “The HTK Book,” Entropic Labs and Cambridge University, 1995-2002.