Multimodal Speech Recognition of a Person with Articulation Disorders Using AAM and MAF

Chikoto Miyamoto [#1], Yuto Komai [#2], Tetsuya Takiguchi [#3], Yasuo Ariki [#4], Ichao Li [*]

[#] *Graduate School of Engineering, Kobe University*
1-1 Rokkodai, Nada-ku, Kobe, 657-8501 Japan
[#1] miyamoto@me.cs.scitec.kobe-u.ac.jp
[#3] takigu@kobe-u.ac.jp
[#4] ariki@kobe-u.ac.jp

[*] *Department of Economics, Otemon Gakuin University*
2-1-15, Nishiai, Ibaraki, Osaka, 567-8502, Japan

***Abstract*—We investigated the speech recognition of a person with articulation disorders resulting from athetoid cerebral palsy. The articulation of speech tends to become unstable due to strain on speech-related muscles, and that causes degradation of speech recognition. Therefore, we use multiple acoustic frames (MAF) as an acoustic feature to solve this problem. Further, in a real environment, current speech recognition systems do not have sufficient performance due to noise influence. In addition to acoustic features, visual features are used to increase noise robustness in a real environment. However, there are recognition problems resulting from the tendency of those suffering from cerebral palsy to move their head erratically. We investigate a pose-robust audio-visual speech recognition method using an Active Appearance Model (AAM) to solve this problem for people with articulation disorders resulting from athetoid cerebral palsy. AAMs are used for face tracking to extract pose-robust facial feature points. Its effectiveness is confirmed by word recognition experiments on noisy speech of a person with articulation disorders.**

I. Introduction

Recently, the importance of information technology in 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 those with 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 during meetings, etc.). However, degradation can be observed in the case of 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.

MMSP'10, October 4-6, 2010, Saint-Malo, France.
978-1-4244-8112-5/10/$26.00

One of the causes of speech impediments is cerebral palsy. About two 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 [6].

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, his/her movements are sometimes more unstable than usual. That means, in cases where movements are related to speaking, their utterances are often unstable or unclear due to 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 [7], we proposed robust feature extraction based on PCA (Principal Component Analysis) 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, the PCA-based filter will be able to extract stable utterance features only. In this paper, we focus on the fact that recognition rate of a person with an articulation disorder decreases compared to that of a physically unimpaired person, especially in speech recognition using dynamic features only. In a person with an articulation disorder, a dynamic feature is not an adequate expression of temporal frequency changes. We use multiple acoustic frames (MAF) as an acoustic dynamic feature to solve this problem.

Further, in a real environment, current speech recognition systems do not have sufficient performance because the quality of the target speech is degraded by the influence of the noise signals that are added to the target speech signal. Audio-visual speech recognition techniques using face information in addition to acoustic information are promising directions

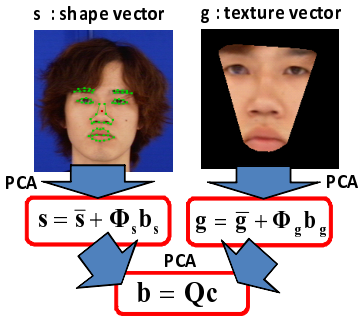

Fig. 1. The composition of AAM

for increasing the robustness of speech recognition, and many audio-visual methods have been proposed (ex. [8]). Lucey, et al. [9] proposed a system that is able to recognize speech from both frontal and profile views. Iwano, et al. [10] proposed a recognition method that uses lip information extracted from side-face images. In the case of a person with an articulation disorder, there is a recognition problem due to the tendency of his/her erratic head movement. We investigated a pose-robust audio-visual speech recognition method using an Active Appearance Model (AAM) [11] to solve this problem. AAMs are used in many applications, such as body part tracking. In particular, it is useful for pose-robust face tracking.

The rest of this paper is organized as follows. In section 2, facial feature point extraction and generation of frontal view face images using an AAM are described. In section 3, the audio-visual recognition method is explained. In section 4, our audio-visual speech recognition experiments are reported. Finally, conclusions are drawn in section 5.

II. Facial feature point extraction and generation of frontal view face images using AAM

An AAM [11] is used in many applications, such as facial part tracking. It is a statistical model that shows the correlation between shape (coordinate values of feature points) and texture (intensity of each pixel). Since the dimensions of these features are reduced by PCA, using AAM can provide fast and stable object tracking. The compositions of AAM, searching of the AAM, and generation of frontal view face images using the AAM are described as follows.

A. The Composition of AAM

Fig. 1 shows the composition of AAM. In this paper, a vector composed of coordinate values on feature points is called shape vector s. To build an AAM, the face region is extracted from images along its feature points and its shape is normalized into a mean shape $\bar{s}$. A vector composed of intensity values from the extracted face region is called texture vector g. PCA is performed on a set of shape vectors s and texture vectors g in training data. The formula (1) gives that:

$$s = \bar{s} + \Phi_s b_s, g = \bar{g} + \Phi_g b_g \tag{1}$$

where $\bar{s}$ is a mean shape vector of s, $\bar{g}$ is a mean texture vector of g. Φ_s and Φ_g are orthogonal matrixes, where each column vector is a base vector and b_s, b_g are coefficients of basis vector. After this, units of the shape and the texture are normalized by matrix Ws, and PCA is performed again. c is a parameter vector controlling both the shape and the texture. Q is an orthogonal matrix that corresponds with the vector c. This is given in the formula (2):

$$b = \begin{pmatrix} W_s b_s \\ b_g \end{pmatrix} = \begin{pmatrix} W_s \Phi_s^T (s - \bar{s}) \\ \Phi_g^T (g - \bar{g}) \end{pmatrix} = Qc \tag{2}$$

B. Searching of AAM

When a new image is given, the AAM, which is created as above, is translated, rotated and synthesized using a parameter vector c. Then, face tracking can be treated as an optimization problem in which we minimize the texture difference between a new image and a synthesized image.

$$c* = arg \min_c |g_n - g_s|^2 \tag{3}$$

where g_s is the texture of a synthesized image made from parameter c, and g_n is the texture of the new image. By getting the parameter $c*$, we can create an AAM that is the most similar to the new image and extract the facial feature points.

C. Generation of Frontal View Face Images

Generation of frontal view face images makes it possible to recognize facial expressions in any face direction and to reduce the amount of training data. Two characteristics of the model vector c, which is obtained by AAM, are used. One characteristic is that the low component of c contains information about the face direction. The other is that the face direction θ and the model parameters c are highly correlated with each other [14]. This is described as follows:

$$c = c0 + c1 * \theta \tag{4}$$

where $c0$ and $c1$ are constant vectors that are learned from the training data using the least square method. Given a face image with a parameter c', we can estimate the face direction θ' as follows.

$$\theta' = (c' - c0)/c1 \quad (c1 \neq 0) \tag{5}$$

After estimating the face direction θ', the residual vector c_{res} is also estimated:

$$c_{res} = c' - (c0 + c1 * \theta') \tag{6}$$

To generate frontal view face images, the frontal direction namely, θ = 0, is given to the formula (6):

$$c_{front} = c0 + c_{res} \tag{7}$$

Thus, we can generate the frontal view images using the above c_{front}.

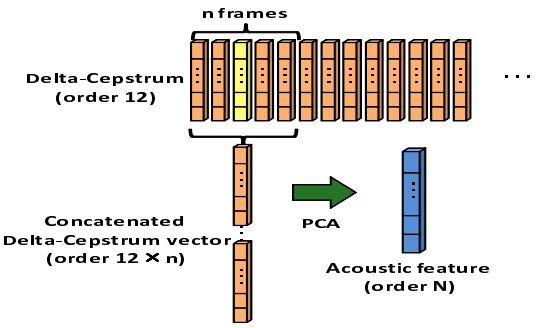

Fig. 2. Multiple acoustic frames (MAF) construction

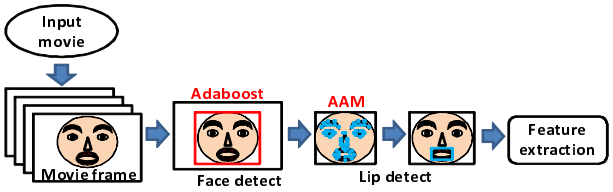

Fig. 3. Visual feature extraction

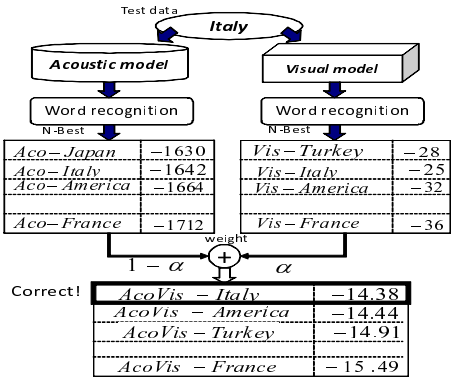

Fig. 4. Example of integrated recognition

TABLE I
RECOGNITION RATE [%] USING ONLY ACOUSTIC MODEL FOR ARTICULATION DISORDER

MFCC	ΔMFCC	MFCC+ΔMFCC
56.7	0.93	68.6

III. AUDIO-VISUAL SPEECH RECOGNITION

A. Acoustic Feature Extraction

In this paper, we focus on the fact that the recognition rate associated with the speech of a person with an articulation disorder decreases compared with that of a physically unimpaired person, especially in speech recognition using dynamic features only (Table I). Because the articulation of speech tends to become unstable due to strain on speech-related muscles, delta-cepstrum is not a good enough expression of temporal frequency changes. Therefore, we use multiple acoustic frames (MAF) as an acoustic dynamic feature to solve this problem. Fig. 2 shows the construction flow of MAF. An acoustic feature vector is constructed by concatenating MAF, and PCA is used to reduce the dimension. Finally, we use cepstrum and this feature combination as an acoustic feature.

B. Visual Feature Extraction

Fig. 3 shows the flow of our visual feature extraction. AAM cannot extract feature points exactly when initial search positions in the input image differ greatly from actual facial feature points. First, face regions are extracted from the input face image sequence by AdaBoost [12] based on Haar-like features and set to the initial search positions. Next, AAM is employed to detect the detailed facial feature points and to set the facial direction to frontal automatically without manual specification. As a result, the lip area is obtained and resized to 32× 32 pixels to be unaffected by the lip size in the screen. Finally, visual feature vectors are derived from the mosaic lip area image by DCT, and PCA is used to reduce the dimension.

C. Audio-Visual Integration

Acoustic and visual HMMs are trained separately. The audio-visual integration enables not only noise-robust recognition but also speaking fluctuation suppression. The integration in recognition is represented as follows:

$$L_{Aco+Vis}^{\acute{w}_{N-best}} = (1-\alpha)\cdot L_{Aco}^{\acute{w}_{N-best}} + \alpha \cdot L_{Vis}^{\acute{w}_{N-best}} \quad (8)$$

Here L_{Aco} and L_{Vis} represent acoustic likelihood and visual likelihood, respectively, and α is weight. As shown in Fig. 4, we perform integrated recognition for only N-best words $\acute{w}_{N-best}$ obtained using word recognition. Then, we integrate the likelihoods according to (8).

IV. RECOGNITION EXPERIMENT

A. Experimental Conditions

The proposed method was evaluated on word recognition tasks for one person with an articulation disorder. We recorded 5,240 utterances (2,620 words, repeating each word two times) and 1,080 utterances (216 words, repeating five times), included in the ATR Japanese speech database. Audio signals are sampled at 16 kHz and windowed with a 25 msec Hamming window every 10 msec. The frame rate of visual signals was 30fps. Fig. 5 shows examples of a spectrogram spoken by a person with an articulation disorder, and by a physically unimpaired person doing the same task. In acoustic experiments, 5,240 clean utterances were used for training and 1,065 clean or noisy utterances for testing. Noises, whose SNR was adjusted 0, 5, 10, 15 and 20 dB, are overlapped with testing data.

B. Recognition Using Only an Acoustic Model for Articulation Disorder

It was difficult to recognize utterances using an acoustic model trained by utterances of a physically unimpaired person. Therefore, we trained the acoustic model using the utterances 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), 12-order MFCCs only and delta-MFCCs only.

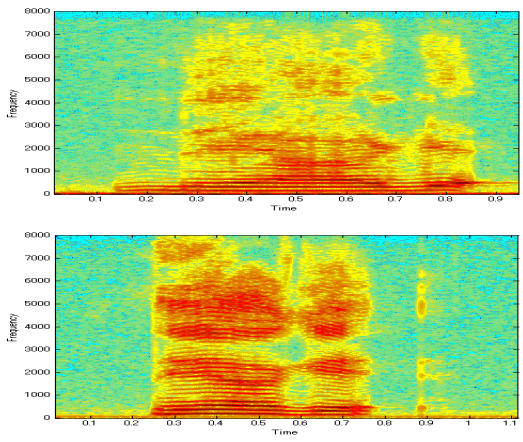

Fig. 5. Examples of a speech spectrogram; upper by a person with an articulation disorder, lower by a physically unimpaired person //n e a g e

TABLE II
RECOGNITION RATE [%] USING MAF

MFCC	ΔMFCC (3 frames)	MFCC+ΔMFCC (3 frames)
56.7	58.9	68.8

Each HMM has three states and three self-loops set with 4 mixture components for each state. As shown in Table I, the recognition rate of delta-MFCCs in a person with an articulation disorder is 0.93%, and it is lower than that using MFCCs.

C. Results Using MAF as Acoustic Feature

The number of input acoustic frames n was experimentally optimized and set to 3. The dimension number of MAF N was set to 12 to compare the recognition rate of delta-MFCCs. As can be seen from Fig. II, the use of MAF improves the recognition rates using delta-MFCCs from 0.93% to 58.9%. clearly show that the use of MAF achieves better performance than delta-MFCCs when dealing with dynamic features.

D. Results Using Audio-Visual Integration

We integrated acoustic models and visual models by using five-best words. The weight α increased from 0 to 1 at an interval of 0.5. Table III shows the recognition rates obtained by the audio-only, visual-only and audio-visual methods at various SNR conditions. In all the SNR conditions, the recognition rates were improved by using the audio-visual method compared with the rates obtained by the audio-only method. These results clearly show that the use of integration achieves good performance.

V. 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 acoustic feature extraction using MAF and a pose-robust audio-visual speech recognition method using AAM. In the acoustic feature extraction, MAF are used as an acoustic dynamic feature instead of a delta-MFCC. It can be expected that MAF will be an adequate expression of temporal frequency changes compared with a delta-MFCC.

TABLE III
COMPARISON OF AUDIO-ONLY, VISUAL-ONLY, AND AUDIO-VISUAL RECOGNITION RATE [%]

SNR	Audio-only	Visual-only	Audio-visual (optimized α)
clean	68.8	35.9	74.1 (0.15)
20 dB	68.0	35.9	73.7 (0.15)
10 dB	57.7	35.9	64.3 (0.1)
5 dB	51.6	35.9	58.9 (0.3)
0 dB	4.9	35.9	35.9 (1.0)

In a real environment, current speech recognition systems do not achieve sufficient performance. Audio-visual integration enables not only noise-robust recognition but also speaking fluctuation suppression. Further, there is a recognition problem due to the tendency of the speaker's erratic head movement. AAM enables face tracking to extract pose-robust facial feature points. The proposed method resulted in an improvement of 7.3% (from 51.6% to 58.9%) in the recognition rate (5 dB SNR) compared to the audio-only method.

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.

REFERENCES

[1] J. Lin, 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, 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, T. Toda, H. Saruwatari and K. Shikano, "Speaking Aid System for Total Laryngectomees Using Voice Conversion of Body Transmitted Artificial Speech," INTERSPEECH, 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] H. Matsumasa, T. Takiguchi, Y. Ariki, I. LI and T. Nakabayashi, "PCA-Based Feature Extraction for Fluctuation in Speaking Style of Articulation Disorders," INTERSPEECH, pp. 1150–1153, 2007.
[8] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive Multimodal Fusion by Uncertainty Compensation with Application to Audio-Visual Speech Recognition," IEEE Transactions on Audio, Speech and Language Processing, vol. 17, no. 3, pp. 423–435, 2009.
[9] P. Lucey, G. Potamianos and S. Sridharan, "A Unified Approach to Multi-Pose Audio-Visual ASR," INTERSPEECH, pp. 650–653, 2007.
[10] K. Iwano, T. Yoshinaga, S. Tamura and S. Furui, "Audio-Visual Speech Recognition Using Lip Information Extracted from Side-Face Images," EURASIP Journal on ASMP, vol.2007, ID 64506, 2007.
[11] T. F. Cootes, G. J. Edwards and C. J. Taylor, "Active appearance models," ECCV, volume 2, pp. 484–498, 1998.
[12] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.1–9, 2001.
[13] L. Wiskott, J.-M. fellous, N. Kruger and C. von der Malsburg,"Face Recognition by Elastic Bunch Graph Matching," IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), pp.775–779, 1997.
[14] T. F. Cootes, K. Walker and C. J. Taylor, "View-based active appearance models," Image and Vision Computing 20, pp. 227–232, 2002.