Advances in Computer Science and Engineering
Volume 8, Number 2, 2012, Pages 123-137

Published by Pushpa Publishing House, Allahabad, INDIA

AUDIO-VISUAL SPEECH RECOGNITION USING AAM-BASED VISUAL FEATURES

Yuto Komai, Tetsuya Takiguchi and Yasuo Ariki

Graduate School of System Informatics
Kobe University
Japan
e-mail: komai@me.cs.scitec.kobe-u.ac.jp
takigu@kobe-u.ac.jp
ariki@kobe-u.ac.jp

Abstract

As one of the techniques for robust speech recognition under noisy environments, audio-visual speech recognition (AVSR) using lip dynamic scene information together with audio information is attracting attention, and the research has made strides in recent years. However, in visual speech recognition (VSR), when a face turns sideways, the shape of the lip as viewed by the camera changes and the recognition accuracy degrades significantly. Therefore, many of the conventional VSR methods are limited to situations in which the face is viewed from the front. This paper proposes a VSR method to convert faces viewed from various directions into faces that are viewed from the front using Active Appearance Models (AAM). In the experiment, even when the face direction changes about 30 degrees relative to a frontal view, the recognition accuracy improves significantly.

Keywords and phrases: audio-visual, speech recognition, face direction.

Received June 14, 2012

1. Introduction

In recent years, audio speech recognition (ASR) software for PCs and mobile phones has become widely used and is attracting attention as a hands-free technology replacing the input from a keyboard. However, in current ASR technologies, the recognition performance degrades under noisy environments, which is a significant problem in regard to making practical use of it in speech recognition.

Human beings use a variety of information comprehensively when understanding the content of an utterance. For example, when it is hard to hear the voice, the listener pays attention to the speaker's lip movement and tries to understand what is being said. Conversely, in the case where the lip movement does not match with the speech, he may misunderstand what is being said. This is called the *McGurk effect*, and it indicates that phonological perception is not decided only by audio information but also by visual information, such as lip movement. Thus, it is important for speech recognition to integrate lip information and audio information.

A technology to recognize speech content from lip motion is called *visual speech recognition* (VSR). VSR is not influenced by noise, whereas ASR is sensitive to noise, and its recognition rate degrades significantly under noisy environments. Therefore, as one of the techniques for robust speech recognition under noisy environments, audio-visual speech recognition (AVSR), using VSR together with ASR, is attracting attention [1].

However, in VSR, when a face turns sideways, the shape of the lip, viewed from a camera fixed in front of the user, changes, and the recognition accuracy degrades significantly. Thus, many of the conventional VSR approaches are limited to situations in which the face is viewed from the front. Therefore, there is a great need to be able to recognize visual speech from arbitrary face directions.

VSR locates the lip ROI (Region of Interest) and extracts the lip features. For detection of lip ROI, traditional image processing techniques, such as color segmentation [2] and edge detection [3], were employed, along with statistical modeling techniques, such as Snakes [4], Active Shape Models (ASM) [5] and Active Appearance Models (AAM) [6]. For the visual features, appearance-based features, such as PCA [7] and DCT [8], and shape-based features, such as the width and height of the lip [9], were employed. Furthermore, a combination of both appearance and shape features, such as AAM parameters [10] has been employed recently.

In regard to research of VSR from various face directions, there is a method that trains the transformation matrices from the profile view to the frontal view and transforms the faces from side to front [11]. However, this technique requires transformation matrices in each direction. Thus, it is difficult to recognize visual speech with arbitrary face directions. In this paper, we propose a method to extract the lip area automatically in various face directions and to recognize visual speech by converting the sideways lip figure into a frontal one using Active Appearance Models (AAM). The experiment results show that the proposed method provides better performance in comparison to the conventional approaches.

2. Overview of Visual Speech Recognition

Figure 1 shows the processing flow. First, the face area is detected based on AdaBoost, using the Haar-like features on the input image. This is because the extraction of the feature points using AAM greatly depends on the initial search area. Therefore, the extraction accuracy of the feature points is improved by applying the detected face area to AAM as an initial search area. After detecting the face area, AAM is applied to the detected face area, and the facial feature points are extracted. Then AAM generates the model parameters most similar to the input image. The speaker's face direction is estimated from the generated parameters using the method described in Subsection 4.2. After estimating the face direction, using the method described in Subsection 4.3, a face in any direction is converted to a frontal face (we call this operation "*normalization*"). Finally, the lip features are extracted, and the visual speech is recognized using Hidden Markov Models (HMMs).

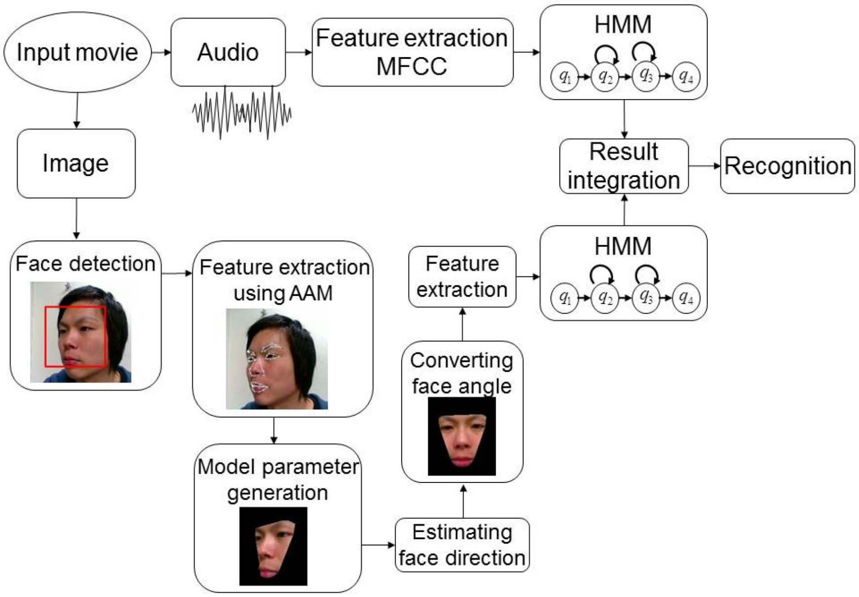


Figure 1. System flow.

The lip feature employed is an AAM model parameter [10] that includes shape information and texture information. In this paper, AAM is applied to the whole face area in order to estimate the face direction accurately, but the AAM model parameters also contain information other than the lip and its movement when whole face AAM is applied. Therefore, after normalization of face direction, some dimensions that include the lip information predominantly in the AAM parameters are extracted and recognized. These dimensions are extracted, from among all the dimension combinations, as the best combinations with the highest recognition accuracy of the visual speech.

In this paper, the audio signal is converted to MFCCs (mel-frequency cepstral coefficients) that are commonly used in a standard speech recognition system. In training, audio and visual HMMs are independently constructed using each feature vector extracted from the same movie.

3. Active Appearance Models

AAM is a technique used to express a facial model using lowdimensional parameters [6, 10]. The subspace is constructed by applying PCA to the shape and texture of face feature points.

3.1. AAM construction

The shape vector **s**, the feature points on the face images, and mean shape $\bar{\mathbf{s}}$ are computed from the training image set. The inner texture of **s** is normalized to mean shape. The shape vector **s** and the texture vector **g** are given: $\mathbf{s} = (x_1, y_1, ..., x_n, y_n)^T$, $\mathbf{g} = (g_1, ..., g_m)^T$, where x_i, y_i $(i \le n)$ are the coordinates of the feature points. g_j $(j \le m)$ is the intensity value at each pixel in $\bar{\mathbf{s}}$, and mean intensity value $\bar{\mathbf{g}}$ can be computed from the training image set.

s and **g** are expressed by using eigenvector matrices $\mathbf{P_s}$ and $\mathbf{P_g}$, obtained by applying PCA to deflection from $\bar{\mathbf{s}}$ and $\bar{\mathbf{g}}$, as shown in Equation (1) and Equation (2):

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{P_s b_s}, \tag{1}$$

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P_g b_g}, \tag{2}$$

$\mathbf{b_s}$ and $\mathbf{b_g}$ are called the *shape parameter* and the *texture parameter*, respectively, and shape vector **s** and texture vector **g** are converted to each of them, respectively. Moreover, $\mathbf{b_s}$ and $\mathbf{b_g}$ are combined and reduced as shown in Equation (3) and Equation (4) by applying PCA because there is a correlation in shape and texture:

$$\mathbf{b} = \begin{pmatrix} \mathbf{W_s b_s} \\ \mathbf{b_g} \end{pmatrix} = \begin{pmatrix} \mathbf{W_s P_s}^T (\mathbf{s} - \bar{\mathbf{s}}) \\ \mathbf{P_g}^T (\mathbf{g} - \bar{\mathbf{g}}) \end{pmatrix} = \mathbf{Qc}, \tag{3}$$

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q_s} \\ \mathbf{Q_g} \end{pmatrix}, \tag{4}$$

where $\mathbf{W_s}$ is the matrix that normalizes the difference of the unit of the shape vector and the texture vector. **Q** is an eigenvector matrix. **c** is a vector of combined shape and texture parameters. This parameter controls both shape and texture as follows:

$$\mathbf{s}(\mathbf{c}) = \bar{\mathbf{s}} + \mathbf{P_s W_s^{-1} Q_s c}, \tag{5}$$

$$\mathbf{g}(\mathbf{c}) = \bar{\mathbf{g}} + \mathbf{P_g Q_g c}. \tag{6}$$

Thus, it becomes possible to treat shape and texture together by controlling only parameter vector **c**.

3.2. Combined parameter

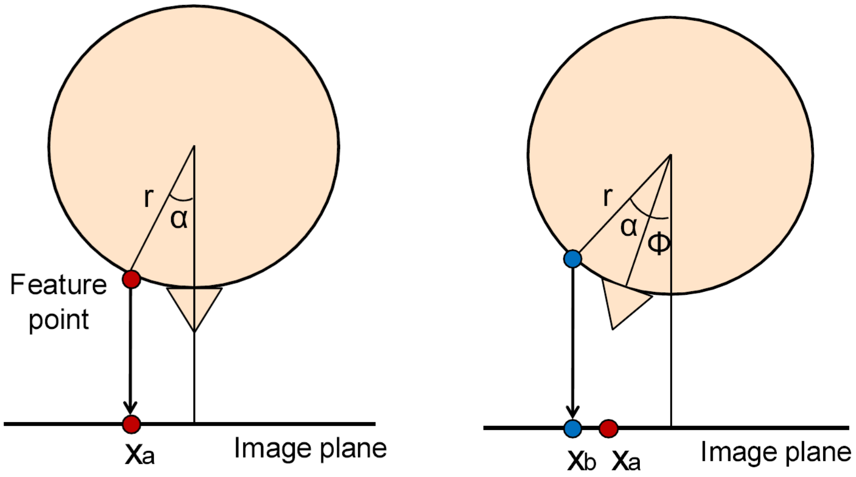
Since the images showing the mouth opening and closing are included in the training data of AAM, the various movements of the lip can be expressed by changing the **c** parameter. Since the **c** parameter has information on detailed shape and the intensity value of the lip, we utilize the **c** parameter as the visual feature. As an extraction method of the **c** parameter, an error **e** between the image $\mathbf{g}(\mathbf{c})$ generated using AAM (this is called a *model image*) and the input image is calculated as follows:

$$\mathbf{e}(\mathbf{c}, \mathbf{p}) = \| \mathbf{g}(\mathbf{c}) - \mathbf{I_i}(\mathbf{W}(\mathbf{p})) \|^2, \tag{7}$$

where $\mathbf{I_i}(\mathbf{W}(\mathbf{p}))$ is the image obtained using the affine transform to the input image $\mathbf{I_i}$. **p** represents the affine parameters of scaling, rotation and translation, and **W** represents a function that executes the affine transform. The optimal **c** parameter is obtained using the steepest descent algorithm, which minimizes the error.

4. Normalization of the Face Direction

A normalization method of the face direction was introduced in [12], and the expanded approach is proposed in this paper, where a multiple regression model is used to estimate the visual feature instead of a single regression model. Each regression model in our method depends on a phoneme class.



(a) Face direction $= 0$ (b) Face direction $= \phi$

Figure 2. Schematic of a face viewed from the head top.

4.1. Regression model

Figure 2 shows a schematic of a face viewed from the top of the head. The face is regarded as a sphere with radius r. A vertical line is drawn to the image plane from the center of the head. Then, the facial feature point at the angle α from the vertical line is projected onto the coordinates Xa of the image plane as shown in Figure 2(a). Furthermore, the facial feature point is projected onto the image plane Xb when the face rotates by the angle ϕ as shown in Figure 2(b). Δx, the distance between two feature coordinate points, is expressed as shown in Equation (8):

$$\begin{aligned}\Delta x &= x_b - x_a \\ &= r\sin(\phi + \alpha) - r\sin\alpha \\ &= r\sin\phi\cos\alpha + r\cos\phi\sin\alpha - r\sin\alpha.\end{aligned} \tag{8}$$

A regression model can be derived considering r and α as constants, as shown in Equation (9):

$$\mathbf{c} = \mathbf{c_0} + \mathbf{c_1}\cos\phi + \mathbf{c_2}\sin\phi, \tag{9}$$

where $\mathbf{c_0}$, $\mathbf{c_1}$ and $\mathbf{c_2}$ are the regression coefficient vectors estimated from the training data using the minimum mean squared error algorithm.

4.2. Estimation of the face direction

When AAM is applied to a new input image with no information of face direction, parameter $\mathbf{c}'$ is generated. Then, the direction ϕ can be estimated as shown in Equation (10) using Equation (9):

$$\begin{pmatrix} \cos\phi \\ \sin\phi \end{pmatrix} = \mathbf{B}^{+}(\mathbf{c}' - \mathbf{c_0}), \tag{10}$$

where $\mathbf{B}^{+}$ is the pseudo inverse matrix of $(\mathbf{c_1} \quad \mathbf{c_2})$:

$$\mathbf{B}^{+} = (\mathbf{B}^{T}\mathbf{B})^{-1}\mathbf{B}^{T}. \tag{11}$$

Therefore, the direction ϕ is estimated as shown in Equation (12) using $\cos\phi$ and $\sin\phi$ in Equation (10):

$$\phi = \tan^{-1}\left(\frac{\sin\phi}{\cos\phi}\right). \tag{12}$$

4.3. Converting of the directional face to frontal face

When AAM is applied to the input image, parameter $\mathbf{c}'$ is generated using AAM, and face direction ϕ is obtained using Equation (12). Then, the residual vector $\mathbf{c_{res}}$ is estimated as shown in Equation (13):

$$\mathbf{c_{res}} = \mathbf{c}' - (\mathbf{c_0} + \mathbf{c_1}\cos\phi + \mathbf{c_2}\sin\phi). \tag{13}$$

The directional face is expressed as shown in Equation (14) using Equation (13):

$$\mathbf{c_{new}} = \mathbf{c_0} + \mathbf{c_1}\cos\theta + \mathbf{c_2}\sin\theta + \mathbf{c_{res}}. \tag{14}$$

If $\theta = 0$, then a face direction is converted to the front. Figure 3 shows the result of the conversion from directional face image to frontal face image.

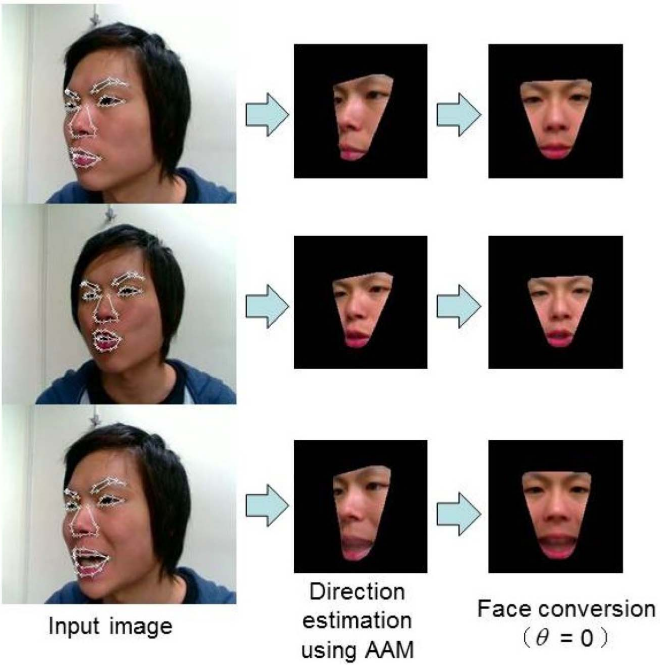


Figure 3. Examples of conversion from directional face image to frontal face image.

4.4. Multiple regression model

In this paper, a multiple regression model of Equation (9) is estimated in order to decrease the variation mismatching that occurs between frontal face image and directional face image. The *i*th regression model is represented as follows:

$$\mathbf{c}^{\mathbf{i}} = \mathbf{c}_{\mathbf{0}}^{\mathbf{i}} + \mathbf{c}_{\mathbf{1}}^{\mathbf{i}} \cos\phi + \mathbf{c}_{\mathbf{2}}^{\mathbf{i}} \sin\phi. \tag{15}$$

Each regression model is estimated using only training data for a phoneme. In this paper, six regression models are estimated using the training data for the Japanese vowels: $/a/$, $/i/$, $/u/$, $/e/$, $/o/$ and the nasal $/N/$, respectively.

In the process of conversion to a frontal face, first, $\mathbf{c}_{\mathbf{in}}$ is obtained by applying AAM to the test image. Next, the face direction θ is estimated using

$\mathbf{c_{in}}$ according to Equation (12). The optimal regression model is selected so that the minimum distance between $\mathbf{c^i}(\theta)$ and $\mathbf{c_{in}}$ is achieved as follows:

$$\hat{\mathbf{i}} = \arg\min_{\mathbf{i}} \| \mathbf{c^i}(\theta) - \mathbf{c_{in}} \|. \tag{16}$$

Then, as described in Subsection 4.3, the face direction is converted to a frontal view.

5. Audio-visual Integration

Audio and visual HMMs are trained separately. In testing, a final likelihood is calculated using the late integration of likelihoods from audio HMMs and visual HMMs as follows:

$$L_{A+V} = (1 - \alpha) L_A + \alpha L_V, \quad 0 \le \alpha \le 1, \tag{17}$$

where L_A and L_V are likelihoods of audio and visual features, respectively. α is the combination weight.

6. Experiment

6.1. Experimental condition

Two subjects spoke ATR phoneme-balanced words (216 words) $\times 10$ sets for a frontal face, the same 216 words $\times 1$ set for a 15-degree face and 30-degree face, respectively. Resolution was 320×240 pixels, and the frame rate was 30 fps.

The leave-one-out method was applied to 216 words $\times 10$ sets, where 216 words $\times 9$ sets for the frontal face image were used for training HMMs, the remaining one set for the frontal face image and the 216 words for the directional faces were used for test, and the recognition rate was the average over the 10 sets. Monophone HMMs were constructed with 5 states and 16 mixtures.

The number of AAM training images was 108, and the number of feature points on each image was 63. As a result of feature extraction (described in Section 3), the AAM parameter for two subjects was reduced to 5 dimensions and 9 dimensions, respectively, for 95% of the cumulative proportion. Including the AAM parameter, its delta and delta-delta parameters, were then used as the visual features. 12-dimensional MFCC parameters, along with their delta and delta-delta parameters, were used as the audio features.

A visual feature using the minimum cross-pose variance (MCPV) has been proposed in [13], where MCPV highlights the DCT feature component most robust to changes in head pose. MCPV has been commonly used for lipreading, and this visual feature was also used in order to compare the proposed **c** visual feature in this experiments.

6.2. Experimental results

Table 1 shows the results of pose estimation for a 0-degree face, 15-degree face, and 30-degree face. The average direction of pose estimation for the 0-degree face, 15-degree face, and 30-degree face is 0.97 degrees, 14.27 degrees, and 32.63 degrees, respectively. As shown in these results, good performance was obtained for pose estimation.

Table 1. Results [deg] of pose estimation for 0 degrees, 15 degrees and 30 degrees (The upper and lower bounds of the 95% confidence interval are also shown)

Face direction	Average	Upper	Lower
0 degrees	0.97	2.9	–0.95
15 degrees	14.27	17.73	10.82
30 degrees	32.63	37.77	27.5

Table 2 shows the recognition rates for only visual features without normalization of face direction. “Front” indicates the recognition rate of the frontal face image. “15 degrees” and “30 degrees” indicate the recognition rate of the 15-degree face and 30-degree face images, respectively. As shown in Table 2, although a high recognition rate is obtained for “front”, the recognition rates for the directional faces degrade seriously. This is because the shape of the lip viewed by the camera changes for the directional faces. Therefore, the recognition rates are affected seriously.

Table 2. Visual recognition rates [%] without normalization of face direction

	Front	15 degrees	30 degrees
c parameter	80.67	13.39	1.30
DCT (MCPV mask)	74.23	29.01	2.67

Table 3. Visual recognition rates [%] with normalization of face direction

	Front	15 degrees	30 degrees
c parameter (single regression)	78.67	54.32	42.35
c parameter (multiple regression)	79.56	54.72	49.37
DCT (MCPV mask)	74.84	52.34	47.61

Table 3 shows the recognition rates for only visual features with normalization of face direction. The recognition rate of “15 degrees” improved by about 41.3 points and the recognition rate of “30 degrees” improved by about 48.1 points compared with Table 2. Thus, it was confirmed that the proposed method is effective for directional face images, and the performance of the multiple regression approach is better than that of the single regression approach. However, the recognition rate of “30 degrees” is lower by about 5.35 points compared to “15 degrees”. One of the reasons for this is that the extraction accuracy of the feature points using AAM degrades when the face direction angle becomes large. Moreover, when converting directional faces, there is a possibility that the lip information is collapsed slightly, causing the recognition rate to degrade.

Table 3 also shows the comparison with the performance of the AAM-based features with that of the conventional DCT-based features using MCPV. 20 components of the 2D DCT feature were selected using MCPV, and including the MCPV feature, its delta and delta-delta features, were then used as the visual features. As shown in Table 3, it was confirmed that the AAM-based features using multiple regression are more effective than the conventional MCPV-based features.

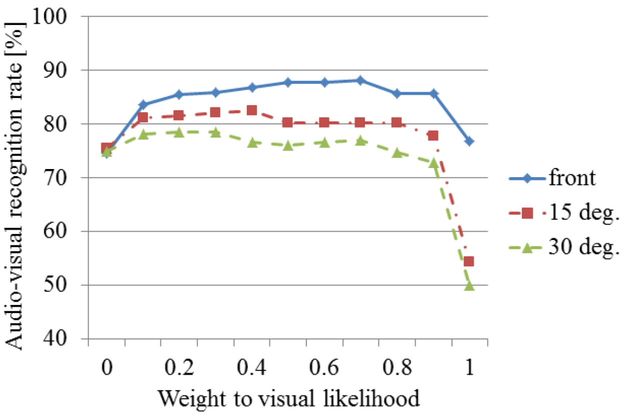


Figure 4. Audio-visual recognition results at SNR of 20 dB.

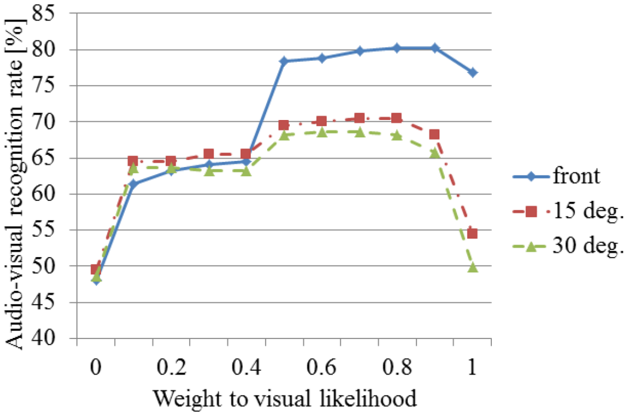


Figure 5. Audio-visual recognition results at SNR of 10 dB.

In order to integrate the visual result with the audio result under noisy environments, the likelihoods from visual HMMs and audio HMMs were integrated according to Equation (17). Figure 4, Figure 5 and Figure 6 show the audio-visual recognition results at SNRs of 20 dB, 10 dB and 0 dB, respectively. The combination weight was increased by 0.1 from 0.0 to 1.0, where the weight 0 corresponds to the audio feature only, and 1.0 to the visual feature only. As shown in both figures, the recognition rate is improved by taking the optimum value of the weight. Although the recognition rate using only audio HMMs greatly decreased in the strong noisy environment at SNR of 0 dB, it could be improved by increasing the weight to the image.

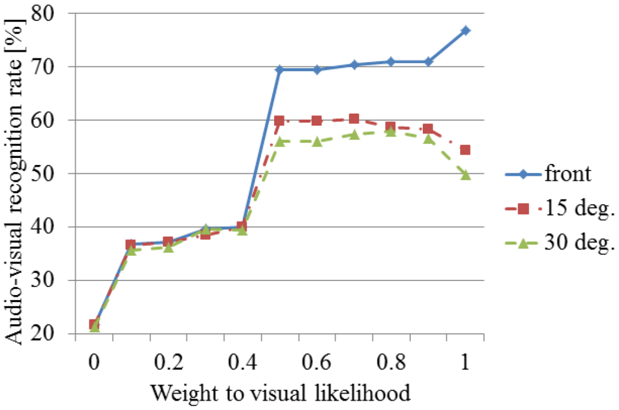


Figure 6. Audio-visual recognition results at SNR of 0 dB.

7. Conclusion

We proposed the method to recognize visual speech by converting directional face images into frontal face images. The experimental results showed that the recognition rate of the directional face images was improved in comparison to those without face direction conversion. Also, it could be confirmed that the recognition rate is improved in comparison to that having only audio features by integrating the visual features and audio features under noisy environments. Future work will include the recognition of utterances spoken by more people, expansion to continuous speech recognition, and recognition of speech with spontaneous tone.

References

[1] G. Potamianos, C. Neti, J. Luettin and I. Matthews, Audio-visual automatic speech recognition: an overview, Issues in Visual and Audio-visual Speech Processing, G. Bailly, E. Vatikiotis-Bateson and P. Perrier, eds., MIT Press, 2004.

[2] Wang Lirong, Xu Jing and Zhao Yanyan, Research of visual features detection and tracking methods about audio-visual bimodal speech recognition, Proc. IFITA, 2010, pp. 204-207.

[3] Ye-Peng Guan, Automatic extraction of lip based on wavelet edge detection, Proc. SYNASC, 2006, pp. 125-132.

[4] Matthew Ramage and Euan Lindsay, Wrapping snakes for improved lip segmentation, Proc. ICASSP, 2009, pp. 1205-1208.

[5] K. L. Sum, W. H. Lau, S. H. Leung, W. C. Alan, Liew and K. W. Tse, A new optimization procedure for extracting the point-based lip contour using active shape model, Proc. ICASSP, 2001, pp. 1485-1488.

[6] T. F. Cootes, Active appearance model, Proc. European Conference on Computer Vision, 1998, no. 2, pp. 484-498.

[7] M. J. Tomlinson, M. J. Russell and N. M. Brooke, Integrating audio and visual information to provide highly robust speech recognition, Proc. ICASSP, 1996, pp. 821-824.

[8] He Jun and Zhang Hua, Research on visual speech feature extraction, Proc. ICCET, 2009, pp. 499-502.

[9] Takami Yoshida and Kazuhiro Nakadai, Audio-visual speech recognition system for a robot, Proc. AVSP, 2010, pp. 8-13.

[10] Yuxuan Lan, Barry-John Theobald, Richard Harvey, Eng-Jon Ong and Richard Bowden, Improving visual features for lipreading, Proc. AVSP, 2010, pp. 142-147.

[11] Patrick Lucey, Gerasimos Potamianos and Sridha Sridharan, An extended pose-invariant lipreading system, Proc. AVSP, 2007, pp. 176-180.

[12] T. F. Cootes, K. Walker and C. J. Taylor, View-based active appearance models, Image and Vision Computing 20(9-10) (2002), 657-664.

[13] A. Pass, J. Zhang and D. Stewart, An investigation into features for multi-view lipreading, Proc. ICIP, 2010, pp. 2417-2420.