Audio-Visual Speech Recognition Based on AAM Parameter and Phoneme Analysis of Visual Feature

Yuto Komai, Yasuo Ariki, and Tetsuya Takiguchi

Graduate School of System Informatics, Kobe University
Rokkodaicho 1–1, Nada-ku, Kobe, Hyogo, 657–8501 Japan
komai@me.cs.scitec.kobe-u.ac.jp, {ariki,takigu}@kobe-u.ac.jp

Abstract. As one of the techniques for robust speech recognition under noisy environment, audio-visual speech recognition using lip dynamic visual information together with audio information is attracting attention and the research is advanced in recent years. Since visual information plays a great role in audio-visual speech recognition, what to select as the visual feature becomes a significant point. This paper proposes, for spoken word recognition, to utilize **c** combined parameter(combined parameter) as the visual feature extracted by Active Appearance Model applied to a face image including the lip area. Combined parameter contains information of the coordinate value and the intensity value as the visual feature. The recognition rate was improved by the proposed feature compared to the conventional features such as DCT and the principal component score. Finally, we integrated the phoneme score from audio information and the viseme score from visual information with high accuracy.

1 Introduction

Recently, various speech recognition technologies have been put to practical use by the development of speech recognition technologies. However, in current speech recognition technologies, there is a problem that the recognition performance remarkably decreases under noisy environment, and it becomes a significant problem in aiming at the practical use of speech recognition.

Then, as one of the techniques for robust speech recognition under noisy environment, audio-visual speech recognition using lip dynamic visual information together with audio information is attracting attention and the research is advanced in recent years.

In audio-visual speech recognition, there are mainly three integration methods; early integration[1] that connects the audio feature vector with the visual feature vector, late integration[2] that weights the likelihood of the result obtained by a separate process for audio and visual signals, and synthetic integration[3] that calculates product of output probability in each state and so on. The research to lip-reading only in the visual feature is actively advanced

Y.-S. Ho (Ed.): PSIVT 2011, Part I, LNCS 7087, pp. 97–108, 2011.

because the visual feature, of course the audio feature, greatly influences the recognition rate in these processing. As the visual feature, various techniques such as width and height of lip[4], optical flow[5] and DCT[6] are employed.

In our research, the lip area is automatically extracted by Active Appearance Models[7][8] (AAM) regardless of speaker's position in the dynamic scene. Moreover, the combined parameter of AAM(**c** parameter) is employed as the feature parameter for utterance recognition. It is thought that shape information included in this parameter can express the lip contour movement, and texture information can express intensity changes such as tooth. Therefore, in this paper, we propose a method that constructs visual HMM using **c** parameter and integrates it with audio HMM. AdaBoost method[9] is employed that uses the Haar-like feature as a face area extraction method, and the late integration that does not take care of audio-visual asynchrony is employed as an integrated method of audio and visual information.

2 System Flow

Fig. 1 shows the block diagram of a processing flow. First, the face area is detected by AdaBoost method that uses the Haar-like feature on the input movie. This is because the extraction accuracy of the feature points by AAM search greatly depends on the initial search area. Therefore, the extraction accuracy of the feature points improves by giving the face area detected by AdaBoost as an initial search area of AAM.

Next, AAM is applied to the detected face area. This process contains two kinds of AAMs. One is the whole face AAM constructed with the training image set in which the feature points are given manually beforehand. The other is the lip area AAM constructed with feature points of the lip area.The purpose of utilizing two AAMs is to extract the feature points accurately on the lip area by applying the whole face AAM roughly at first and then applying the lip area AAM precisely on the extracted lip area. If **c** parameter extracted from the whole face AAM is used as a recognition parameter, the recognition rate might decrease by the information other than the lip area. Therefore, we use two kinds of AAMs to extract a more accurate parameter of the lip area.

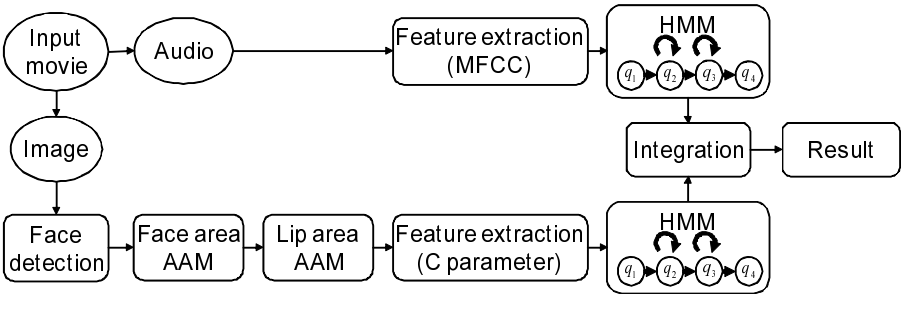

Fig. 1. System Flow

When the lip area AAM is applied to the input image, **c** parameter that generates the most similar lip area image with the input image is extracted as the visual feature. In training, audio and visual HMMs are independently constructed by using the visual feature and audio feature extracted from the same movie. Finally, the recognition result is output by integrating likelihoods from visual HMM and audio HMM.

3 Feature Extraction

3.1 Active Appearance Models

AAM is a technique to express the face model by the low-dimensional parameter. The subspace is constructed by applying PCA to shape and texture of face feature points.

The shape vector **s** that is composed of the feature points on the face image and mean shape $\bar{\mathbf{s}}$ is computed from the training image set. Inner texture of **s** is normalized to mean shape. The shape vector **s** and the texture vector **g** are given in $\mathbf{s} = (x_1, y_1, \cdots, x_n, y_n)^T$, $\mathbf{g} = (g_1, \cdots, g_m)^T$. where x_i, y_i $(1 \leq i \leq n)$ are the coodinates of the feature points. g_j $(1 \leq j \leq m)$ is the intensity value at each pixel within the area srrounded by $\bar{\mathbf{s}}$, and mean intensity value $\bar{\mathbf{g}}$ can be computed from the training image set. Vectors **s** and **g** are expressed by using eigenvector matrices $\mathbf{P_s}$ and $\mathbf{P_g}$, obtained by applying PCA to deviation from $\bar{\mathbf{s}}$ and $\bar{\mathbf{g}}$, as shown in Eq. (1) and Eq. (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 them. Moreover, $\mathbf{b_s}$ and $\mathbf{b_g}$ are combined and reduced as shown in Eq. (3) by applying PCA because there is a correlation in shape and texture parameters.

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

where $\mathbf{W_s}$ is the matrix that normalizes the difference of the unit between the shape vector and the texture vector. **Q** is an eigenvector matrix, and **c**, called combined parameter, is a parameter that controls both shape and texture. **s** and **g** are expressed as shown in Eq. (4) and Eq. (5) by **c**.

$$\mathbf{s(c)} = \bar{\mathbf{s}} + \mathbf{P_s W_s}^{-1} \mathbf{Q_s c} \tag{4}$$

$$\mathbf{g(c)} = \bar{\mathbf{g}} + \mathbf{P_g Q_g c} \tag{5}$$

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

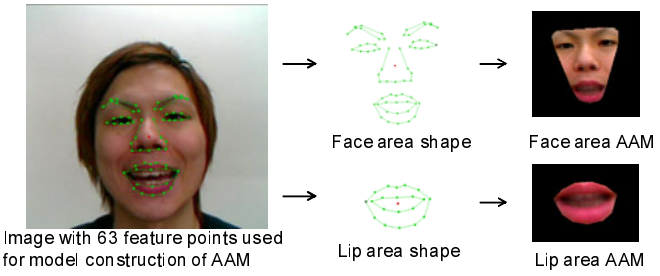

Fig. 2. Construction of two kinds of AAMs

3.2 Model Construction

Two kinds of AAMs are constructed as described in Chapter 2. The whole face AAM is constructed by using the shape information and the inside texture information from the training image set with the feature points manually given to the whole face as shown in Fig. 2. The lip area AAM is constructed with the shape information and the inside texture information extracted automatically from the feature points only on the lip area extracted by the whole face AAM.

3.3 Combined Parameter

Since the images with the mouth opening and closing are included in the training data set of AAM, the various movements of the lip can be expressed by changing **c** parameter as shown in Fig. 3. Since **c** parameter has information on detailed shape and the intensity value of the lip, we propose to utilize **c** parameter as the visual feature. As an extraction method of **c** parameter, error **e** between the image $\mathbf{g}(\mathbf{c})$ generated by AAM (this is called a model image) and the input image is formulated as shown in Eq. (6).

$$\mathbf{e(c,\ p)} \ = \ \|\mathbf{g(c)} \ - \ \mathbf{I_i}(\mathbf{W}(\mathbf{p}))\|^2 \tag{6}$$

where $\mathbf{I_i}(\mathbf{W}(\mathbf{p}))$ is the image obtained by Affine transform to the input image $\mathbf{I_i}$. **p** is an Affine parameter of scaling, rotation and translation and **W** is a function that executes the Affine transform. The number of dimension of **c** is set to 10. 78 training images are prepared. Since the video rame rate is about 1/3 of audio frame rate in our data set, there is a possibility that the visual recognition rate decreases compared to the audio recognition rate. Therefore, it is interpolated by the cubic spline function between visual frames. **c** parameters obtained thus, its Δ and $\Delta\Delta$ coefficients with 30 dimensions in total are finally used as the visual feature.

3.4 Additional Feature

In order to compare with **c** parameter, 2D DCT and pixel values on the lip area are extracted. The lip area is located by the whole face AAM, and the area is

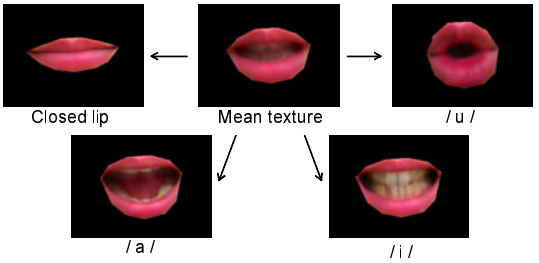

Fig. 3. Example of model images generated by changing c parameter (in a counter-clockwise fashion from the top middle, mean texture, the closed lip, utterance /a/, /i/ and /u/.)

normalized to the square with the fixed ratio of width to hight and converted into the gray scale. The feature is extracted on this area. A square size is 32 32 pixel. PCA is applied to this 1024 dimensional vector of pixel values for the dimension reduction. The number of dimension is set to 10 according to the cumulative contribution ratio 90% . PCA score, its Δ and $\Delta\Delta$ coefficients with 30 dimensions in total are used as the feature of PCA score. In a case of 2D DCT, after DCT operation, 16 low-frequency components are selected because the information concentrates on the low-frequency region in DCT. DCT, its Δ and $\Delta\Delta$ coefficients with 48 dimensions in total are used as the feature of DCT.

4 Recognition Method

As a recognition method, both word type HMM and subword type HMM are used. MFCC with 12 dimensions and logarithm power, their Δ and $\Delta\Delta$ coefficients with 39 dimensions in total are used as the audio feature. A final likelihood is calculated by the late integration of audio and visual information as shown in Eq. (7)[2].

$$L_{A+V} = \alpha L_A + (1-\alpha) L_V , \quad 0 \leq \alpha \leq 1 \tag{7}$$

where L_{A+V} is a likelihood after integration, L_A and L_V are likelihoods of audio and visual features respectively. α is the combination weight.

5 Experiment

5.1 Experimental Condition

We used ATR phoneme balance words (216 words)×10 sets and single set of 100 words (different from 216 words) chosen at random from ATR phoneme balance sentences as an utterance words. Logicool Qcam Orbit MP was used as a filming equipment and SONY ECM-PC50 was also used as a microphone. Resolution was 960×720 pixel, and the frame rate was 30fps.

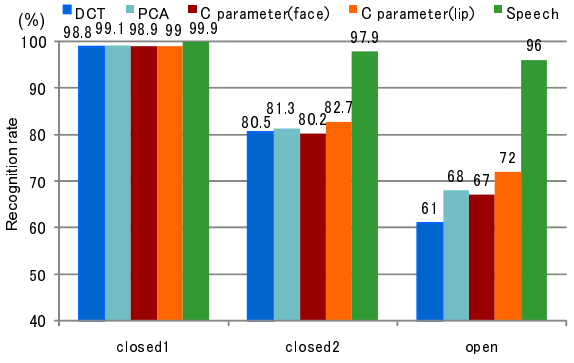

Fig. 4. Recognition results by various audio and visual features in different conditions

One specific speaker uttered in a clear tone with the frontal face. The distance from the speaker to the camera was about 40cm. The noise was added onto the speech so that SNR became 5dB, 0dB and -5dB. The leave-one-out method was applied to 216 words×10 sets, and the recognition rate was the average over the 10 sets. We call this experiment as one under the language closed condition because the same 216 words are used for training and recognition. In addition, 216 words×10 sets were used for training, and 100 words×1 set were recognized. We call this experiment as one under the language open condition, because 100 words are recognized different form 216 words used for training. Word type HMMs were constructed with 5 states and 4 mixtures and used in the language closed condition. As subword type HMMs, monophone HMMs were constructed and used in both the language closed and open conditions. The number of mixture was experimentally chosen for the best one in the language open condition.

5.2 Recognition Result by Using Respective Feature

Fig. 4 shows the result of the utterance recognition carried out separately using the visual feature and audio feature respectively. Closed1 in Fig. 4 indicates the recognition rate by word type HMM, closed2 is by subword type HMM in the language closed condition, and open is in the language open condition. C parameter(face) and C parameter(lip) indicate the recognition results by **c** parameter extracted from the whole face AAM and the lip AAM respectively.

Comparing these results in terms of the features, a high recognition rate was obtained by the conventional features and **c** parameter in closed1. Moreover, it was confirmed that the lip area **c** parameter was more effective than the conventional features in closed2 and open.

Comparing these results in terms of the conditions, the recognition rate decreased in closed2 and open compared with closed1 for the visual feature while it was high in any condition for audio feature. The difference of the conditions between closed1 and closed2 was the HMM type; word type HMM or subword

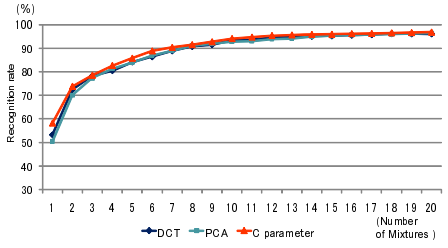

Fig. 5. Recognition rates as a function of the number of mixtures(closed2)

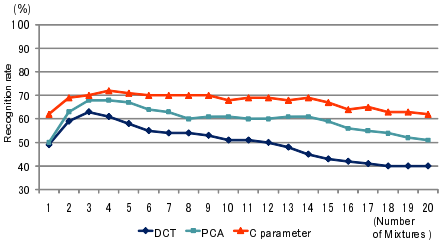

Fig. 6. Recognition rates as a function of the number of mixtures(open)

type HMM. The recognition rate by the subword type HMM was lower than that by the word type HMM because connected training of the phoneme was necessary for the subword type HMM. In the open condition, the recognition rate was lower than that in closed2. Fig. 5 and 6 show the recognition rates by the visual HMMs as a function of the number of mixtures. In the figure, as the number of mixtures increases, the recognition rate is improved in closed2. Since the increase of the number of mixtures leads to the complex model and the training words and test words are same in closed2, it seems that the model is over-fitted to the training data. On the other hand, the recognition rate tends to be lower as the number of mixtures increases in open. Due to this reason, in closed2, the recognition rate is higher than that in open.

5.3 Integrated Result of Audio and Visual Features

In order to integrate the visual result with the audio result under noisy environment, output likelihood by visual HMM with **c** parameter and that by audio HMM were integrated by Eq. (7). Fig. 7 shows the recognition results at 5dB, 0dB and -5dB SNR of the speech data. The weight $1 - \alpha$ to visual feature was increased by 0.1 from 0.0 to 1.0.

Three types of integration of the visual HMMs were carried out with the subword type audio HMM. They were word type visual HMM(closed1), subword type visual HMM(closed2) in the language closed condition and subword type visual HMM(open) in the language open condition respectively. A horizontal axis in Fig. 7 indicates the weight to visual feature. The weight 0 corresponds to audio feature only, and 1 to visual feature only.

From Fig. 7, it can be seen that, in any conditions, the recognition rate is comparatively acceptable in clean and 5dB SNR environment. Therefore, the recognition rate is high at any values of the weight and is improved by taking the optimum value of the weight. The recognition rate by audio HMM greatly falls down in the strong noisy environment at 0dB and -5dB SNR. However, it can be improved by increasing the weight to the image. From these results, it can be confirmed that the recognition rate is improved compared with audio feature by integrating the visual feature and audio feature under noisy environment.

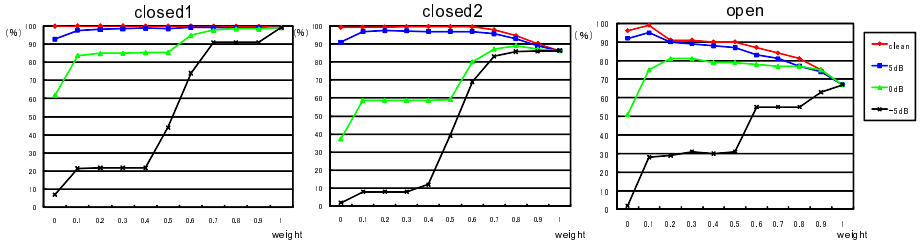

Fig. 7. Integrated result of audio and visual features

6 Phoneme Analysis of Visual Feature

6.1 Continuous Phoneme Recognition

In order to investigate the recognition accuracy of each phoneme using audio and visual features, continuous phoneme recognition was carried out for words. The language model was phoneme pair such that vowel appears after consonant and consonant appears after vowel at equal probability. The acoustic model and the visual model were the subword type audio HMM and the subword type visual HMM trained by 216words×10 sets, and the recognition words were 100 words used in the language open condition. The visual feature was **c** parameter.

Fig. 8 shows the confusion matrix of the phoneme recognition in language open condition by audio features, and Fig. 9 shows the confusion matrix of the phoneme recognition in language open condition by **c** parameter. "IN" and "LA" in the figure indicate the number of insertion errors and the number of deletion errors respectively. MoreoverCin order to evaluate the phoneme recognition accuracy, the phoneme correct and the phoneme accuracy of vowel, consonant and all phonemes were computed. The phoneme correct and the phoneme accuracy correspond to word correct and word accuracy respectively when the phoneme is regarded as a word.

Table 1 shows the result. In the table, the recognition accuracy is approximately 80% in both vowel and consonant in audio. However, the recognition accuracy of consonant is about 12% in open condition by visual feature though vowel is approximately 70%, and the accuracy of all phonemes is approximately 40%. Thus, it can be said that consonants are not recognized well by the visual feature.

6.2 Analysis of False Recognition of the Phoneme

In Fig.8 and Fig.9, both vowel and consonant recognition accuracies are high by audio feature. On the other hand, in **c** parameter, vowels are recognized well to some degree, but various errors occur more than audio feature in consonants.

The insertion error occurs a lot in "r". It is thought that the shape of the mouth becomes same in "a" and "ra" and it can not be distinguished because "r" is a consonant uttered by the movement of the tongue only. The deletion

Table 1. Phoneme correct and phoneme accuracy (%)

	Audio		Visual			
	Open		Open		Closed2	
	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct
Vowel	82.91	82.91	67.81	68.38	65.46	66.21
Consonant	72.4	75.38	11.85	21.58	37.46	45.87
All	77.65	79.26	40.74	45.74	52.86	57.05

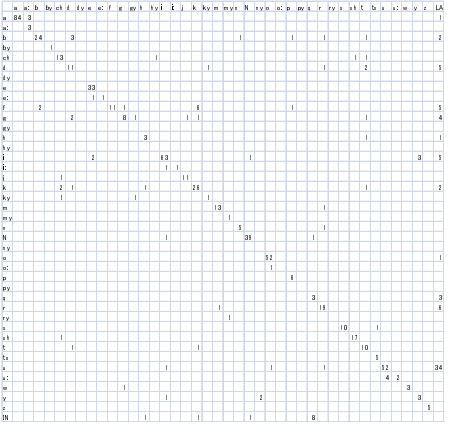

Fig. 8. Phoneme confusion matrix by audio feature (open)

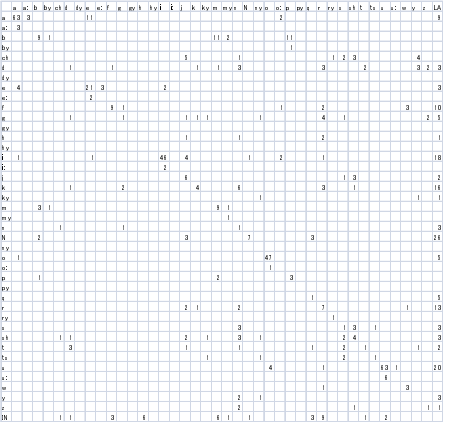

Fig. 9. Phoneme confusion matrix by visual feature (open)

error occurs a lot in "N". When "N" appears at the end of the word, the mouth becomes in a closed shape. Since the mouth is closed before and after the utterance, it is regarded as a silent section, then the deletion error occurs. Moreover, when "N" appears in the word, the shape of the mouth is kept similar to the previous vowel. Therefore, it is thought that the deletion error is increased because "N" has a large variance and sparse feature. The substitution error occurs in various phonemes. For instance, "k" is falsely recognized as the consonants such as "g", "n" and "r". It is thought that the substitution error occurs because there is no movement of the mouth in these consonants.

6.3 Experiment with Viseme

The reason why the false recognition described in 6.2 is caused is attributed to the fact that the phoneme is a minimum unit representing the sound. When the phoneme is applied to the visual feature, the phonemes with the same shape of the mouth such as "k" and "g" cannot be distinguished. Therefore, the viseme will be the best unit, instead of the phoneme, to represent the visual feature.

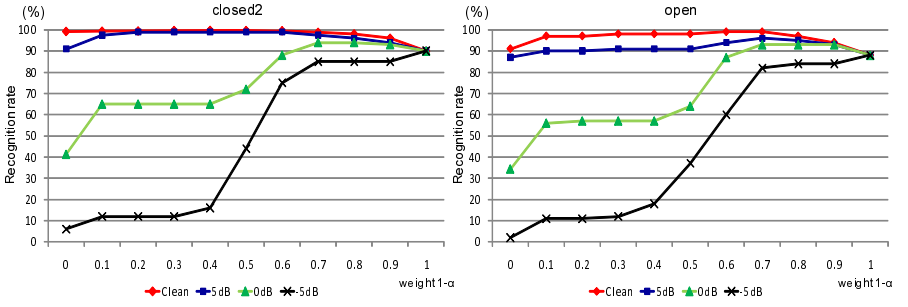

Fig. 10. Integrated result when the viseme is used for visual information and phoneme is used for audio information

Table 2. Viseme correct and viseme accuracy (%)

	Open		Closed2	
	Accuracy	Correct	Accuracy	Correct
Vowel	75.9	75.9	78.21	78.58
Consonant	47.69	57.85	63.28	68.44
All	62.54	67.35	71.59	74.08

From this viewpoint, the viseme was employed as a unit to represent the visual feature, referring to Fukuda[10], and the visual data was recognized as was done in Chapter 5 by visual HMM and the result was integrated with the audio result. The number of mixtures was set to 12 based on the best result using the viseme. There were some words that could not be distinguished like "eikyou" and "eigyou" because both became "eisyou" in viseme. For such words, the same output likelihood from the visual HMM was integrated with those from the audio HMMs with different phoneme sequence. Fig. 10 shows the integrated result in closed2 and open.

In the figure, it can be confirmed that the recognition results are better than those in Fig. 7, because the recognition rate by the visual HMM using viseme is higher than that using phoneme shown in Fig. 7. Therefore, the highest accuracy is obtained by integrating the recognition results using phoneme for audio feature and viseme for visual feature.

As the experiment, the continuous viseme recognition was carried out. Fig. 11 shows the confusion matrix, and Table 2 shows the correct and the accuracy when viseme is used.

Comparing Table 2 with Table 1, the viseme greatly improved the recognition accuracy in both vowels and consonants, compared to the phoneme case. However, it is still low by about 10 points in closed2 compared to audio. In Fig. 11, "N" has still many deletion errors as is described in 6.2 for the phoneme, and "t" has many substitution errors with various visemes. Viseme "t" includes the phoneme "t", "d" and "n". In order to discriminate these, it is important

	a	i	u	e	o	p	r	sy	w	t	s	y	vf	N	LA
a	73			10											8
i		57													21
u			68		6										27
e	3	3		26											4
o					50										5
p						54									1
r						1	11	2					3		9
sy		1						44		2		2			2
w			1						26						3
t							5	2	2	15	1	2	5		7
s								5		7	8				1
y							1	2				2			
vf							2	8		7	3		18		17
N						2								10	29
IN						10	18	1		3		1			

Fig. 11. Viseme confusion matrix using c parameter(open)

to extract the movement of the tongue because they are uttered by changing the tonge position. Moreover, if they can be discriminated, the accuracy of the viseme "vf" will be improved that has many substitution error to "t".

It is thought that there will be still room in the improvement of the visual feature. In the future, we will investigate the feature that can be extracted from the movement of the tongue described above, and the feature that can recognize "N" clearly.

7 Conclusion

We proposed to utilize **c** parameter extracted by Active Appearance Model applied to a face image for the utterance recognition. The effectiveness was confirmed by integrating **c** parameters as the visual feature with the audio feaure. The difference between the phoneme recognition accuracy by the audio feature and the visual feature was clarified by calculating the phoneme confusion matrix. In addition, the phoneme score from audio feature and the viseme score from visual feature were integrated with high accuracy.

In our approach, the utterances spoken by one specific speaker with a clear tone were recognized in the experiment. Future tasks include the recognition of utterances spoken by more people, new integration method of audio and visual feature, weight optimization technique, recognition of speech with spontaneous tone, application of AAM to images with various face directions, expansion to continuous speech recognition, and robustness to the difference of time session. Though monophone type HMM was used in this experiment because of the data amount, a further improvement of the recognition rate will be expected by increasing the data amount and using triphone type HMM.

References

1. Potamianos, G., Graf, H.P.: Discriminative Training Of HMM Stream Exponents For Audio-Visual Speech Recognition. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1998), Florham Park, NJ, pp. 3733–3736 (1998)

2. Verma, A., Faruquie, T., Neti, C., Basu, S., Senior, A.: Late Integration In Audio-Visual Continuous Speech Recognition. In: Automatic Speech Recognition and Understanding (1999)
3. Tomlinson, M.J., Russell, M.J., Brooke, N.M.: Integrating audio and visual information to provide highly robust speech recognition. In: Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP1996), pp. 821–824 (1996)
4. Kumar, K., Navratil, J., Marcheret, E., Libal, V., Ramaswamy, G., Potamianos, G.: Audio-Visual Speech Synchronization Detection Using a Bimodal Linear Prediction Model. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 53–59 (1999)
5. Iwano, K., Tamura, S., Furui, S.: Bimodal speech recognition using lip movement measured by optical-flow analysis. In: Proc. International Workshop on HSC 2001, pp. 187–190 (2001)
6. Jun, H., Hua, Z.: Research on Visual Speech Feature Extraction. In: 2009 International Conference on Computer Engineering and Technology, pp. 499–502 (2009)
7. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active Appearance Models. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 484–498. Springer, Heidelberg (1998)
8. Dornaika, F., Ahlberg, J.: Fast and reliable active appearance model search for 3-d face tracking. IEEE Transactions on Systems, Man, and Cybernetics, 1838–1853 (2004)
9. Viola, P., Jones, M.: Rapid Object Detection Using Boosted Cascade of Simple Features. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1–9 (2001)
10. Fukuda, Y., Hiki, S.: Characteristic of the mouth shape in the production of Japanese-Stroboscopic observation. In: IEICE, pp. 259–265 (1978)