ISCA Archive
http://www.isca-speech.org/archive

Auditory-Visual Speech Processing
(AVSP) 2013
Annecy, France
August 29 - September 1, 2013

Improvement of Lipreading Performance Using Discriminative Feature and Speaker Adaptation

Takumi Seko, Naoya Ukai, Satoshi Tamura, Satoru Hayamizu

Department of Information Science, Gifu University, Gifu, Japan
{seko,ukai}@asr.info.gifu-u.ac.jp, tamura@info.gifu-u.ac.jp, hayamizu@gifu-u.ac.jp

Abstract

In this paper, we apply a general and discriminative feature "GIF" (Genetic Algorithm based Informative feature) to lipreading (visual speech recognition), and improve the lipreading performance using speaker adaptation. The feature extraction method consists of two transforms, which convert an input vector into GIF for recognition. In the speaker adaptation, MAP (Maximum A Posteriori) adaptation is used to adapt a recognition model to a target speaker. Recognition experiments of continuous digit utterances were conducted using an audio-visual corpus CENSREC-1-AV [1] including more than 268,000 lip images. At first, we compared the GIF-based method with the baseline method employing conventional eigenlip features, using two kinds of images: pictures in the database around speakers' mouth, and extracted images only containing lips. Secondly, we evaluated the effectiveness of speaker adaptation for lipreading. The result of comparison shows that the GIF-based approach achieved slightly better than the baseline method. And it is found using the mouth-around images is more suitable than lip-only images. Furthermore, the result of speaker adaptation shows that speaker adaptation significantly improved recognition accuracy in the GIF-based method; after the adaptation, the recognition rate drastically increased from approximately 30% to 70%.

Index Terms: discriminative feature, lipreading, speaker adaptation, lip extraction, CENSREC

1. Introduction

In recent years, speech recognition technology has been widely developed with the spread of car navigation systems or smart phones. However, in real environments where these devices are often used, background acoustic noises are overlapped into speech signals. Due to the corrupted audio signals, speech recognition may not go well even if several noise reduction techniques are performed in a recognition engine. In order to overcome the degradation, audio-visual speech recognition employing visual data, e.g. lip or mouth images, has been investigated since visual information is not affected by acoustic noises; today, lip images can be easily obtained since a camera is embedded on a hardware where speech recognition is used. In the audio-visual speech recognition, a lipreading technology that estimates what word is pronounced only using visual data plays a big role in enhancing the robustness of speech recognition. This paper focuses on the lipreading technology.

In the lipreading research field, visual feature is one of the major research topics. Many features for lipreading have been proposed, for example, eigenlip [2], optical-flow-based features [3, 4], Active Appearance Model (AAM) features [5, 6], and Trajectory features [7]. Nevertheless, the lipreading recognition accuracy is not significant even in controlled conditions [8]. Another issue lies that lipreading recognition accuracy greatly depends on speakers; for some speakers a lipreading system can be performed to some extent, while its performance is quite low for some people. Therefore, this paper also focuses on speaker adaptation technique, e.g. MLLR (Maximum Likelihood Regression) [9] or MAP (Maximum A Posteriori) [10], that are used widely in the field of speech recognition [11]. Several audio-visual speech recognition systems employs such the model adaptation scheme, e.g. a method using MLLR in addition to a stream-weight optimization based on a likelihood-ratio maximization criterion [12, 13].

This study aims at firstly evaluating performance of lipreading using a new feature: a discriminative feature called GA-based Informative Feature (GIF). Lipreading recognition experiments were conducted comparing GIF with conventional lipreading features such as eigenlip features. As the second objective of this study, the usefulness of speaker adaptation in lipreading is clarified. We conducted recognition experiments using the adaptation in open and closed conditions.

This paper is organized as follows: In Section 2, GIF is introduced. In Section 3, MAP estimation is described. Section 4 describes a lipreading method and a database used in this paper, followed by lipreading recognition experiments and results in Section 5. Finally Section 6 concludes this paper.

2. GA-Based Informative Feature

In this section, the feature GIF (GA-based Informative Feature) used in lipreading in this study is briefly introduced. This feature can be utilized in various pattern recognition tasks and related works [14, 15, 16].

At first, an N-dimensional input vector $\boldsymbol{x}$ is converted into a C-dimensional intermediate vector $\boldsymbol{y}$ as:

$$\boldsymbol{y} = A\left(\boldsymbol{x}^\top\ 1\right)^\top \tag{1}$$

In Eq.(1), A is a $C \times (N+1)$ transformation matrix, where C is the number of classes that should be classified In speech recognition C is equivalent to the number of phonemes. This process is called "Stage 1". In the next process "Stage 2", the vector $\boldsymbol{y}$ is further converted into an M-dimensional output feature vector (GIF) $\boldsymbol{z}$ as:

$$\boldsymbol{z} = B\,\boldsymbol{y} \tag{2}$$

where B is an $M \times C$ transformation matrix, in this paper, $\boldsymbol{y}$ is compressed into a 10-dimensional feature vector $\boldsymbol{z}$. The "discriminative" matrix A is built so that each row vector of A corresponds to coefficients of a linear classifier. The matrix B is used for dimension reduction and orthogonalization. These matrices can be computed by applying Genetic Algorithm (GA). Details of computing these matrices are appeared in [14].

Once the first projection A and the second projection B are determined, a feature vector $\boldsymbol{z}$ can be computed by applying Eqs.(1) and (2). Before applying the second projection, a bias vector $\boldsymbol{\mu}$ is calculated beforehand as:

$$\boldsymbol{\mu} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}_t = \frac{1}{T}\sum_{t=1}^{T}A\left(\boldsymbol{x}_t{}^\top\ 1\right)^\top \tag{3}$$

where $X = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T)$ is a sequence of input vectors. Subsequently, each intermediate vector is normalized by suppressing the bias vector:

$$\hat{\boldsymbol{y}_t} = \boldsymbol{y}_t - \boldsymbol{\mu} \tag{4}$$

3. Maximum A Posteriori Estimation

This paper uses MAP estimation that is one of the adaptive techniques since it is often used in speech recognition. In this section, MAP estimation is introduced.

This technique estimates a model parameter set $\hat{\theta}$ by maximizing a posterior probability $P(\theta|X)$ where X indicates adaptive data. So the parameter set $\hat{\theta}$ can be estimated as:

$$\hat{\theta} = \arg\max_{\theta} P(\theta|X) \tag{5}$$

Let us denote a prior distribution of θ by $P(\theta)$. Then the posterior probability $P(\theta|X)$ can be deformed by the Bayes' theorem:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{6}$$

where $P(X|\theta)$ is a likelihood function that is an occurrence probability of X when the underlying population parameter is θ. $P(X)$ is assumed to be invariable since $P(X)$ is independent of θ. Thus, MAP estimation can be realized through estimating $\hat{\theta}$ which maximizes a product of $P(X|\theta)$ and $P(\theta)$ on the basis of adaptive data X:

$$\hat{\theta} = \arg\max_{\theta} P(X|\theta)P(\theta) \tag{7}$$

MAP estimation is more robust than maximum likelihood estimation when not so much observational data are available. If we have no knowledge about θ, the MAP estimation corresponds to the maximum likelihood estimation. $P(\theta)$ has a major effect on the MAP estimation in case of the small amount of input data X, in contrast, $P(X|\theta)$ has a major effect when the large amount of input data X can be used.

4. Lipreading Method

In this section, a lip reading method including feature extraction and training/recognition scheme is introduced. A database used in this paper is also described.

4.1. Database

In the following experiments, we used image data in a database CENSREC-1-AV [1]. CENSREC-1-AV is an evaluation corpus originally for audio-visual speech recognition in noise environment, which is also available for lipreading experiments. This corpus includes speech waveforms and image data of Japanese continuous 1 to 7-digit utterances recorded in office environments. In addition, this corpus has two kinds of image data: color and infrared images. These images include only a surrounding area of subject's mouth, as shown in Fig. 1 (a). In this paper, only color image data is employed. This database consists of two data sets: a training data set for building models and feature transforms, and a test data set for evaluating a lipreading method. The visual specification of CENSREC-1-AV is summarized in Table 1.

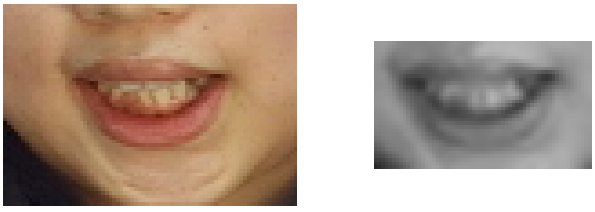
For the following reasons, we prepared another image data obtained by lip extraction from the original image data set. A sample image after the lip extraction is illustrated in Fig. 1 (b).

- To focus on movements of subject's mouth and emphasize lip information that would be necessary for lipreading.
- To save computational time of feature extraction by reducing the dimension of input vector, since the larger input vector is, the longer feature extraction takes.

In the lip extraction, an AdaBoost method using Haar-like features [17] was adopted to determine the extraction window. A lip area could not be detected in some

Table 1: *A summary of visual data in CENSREC-1-AV.*

frame rate		29.97Hz (NTSC)
utterance content		Japanese continuous 1- to 7-digit utterances
image size / depth		width 81 × height 55 pixel 24bit color
file format		Windows Bitmap Image (.bmp)
speakers	training set	female: 20, male:22
	test set	female 26, male:25
utterances	training set	3,234 utterances (77 utterances/speaker)
	test set	1,963 utterances (38-39 utterances/speaker)
images	training set	female:127,392, male:140,818, total:268,210
	test set	female:86,515, male:78,762, total:165,277



(a) original image (b) lip image

Figure 1: *Samples of (a) original image in CENSREC-1-AV and (b) extracted lip image from the original image.*

pictures, and in such the cases, the previous window was successively used. Note that all the lip images obtained have the same image size (59 × 35).

4.2. Feature extraction

A flow of feature extraction of conventional and proposed features are shown on Fig. 2. In order to reduce computational resources, the two kinds of images described above (original images / lip images) were resized to one third (26 × 18 / 19 × 11), respectively. From both images, the gray scale images were respectively obtained. The above processing except clipping is the same as a baseline system in CENSREC-1-AV [18]. After the preprocessing, a 468-dimensional (for original images) or a 209-dimensional (for lip images) input vector was then computed from intensity values in an image, enumerating all the pixels from left-top to right-bottom.

Eigenlip features were extracted to evaluate the effectiveness of GIF in lipreading. In common with above manipulation, we referred to the baseline system [18]; using the eigenvectors obtained by applying Principal Component Analysis (PCA) to several training vectors chosen from input vectors in the training data set, each input

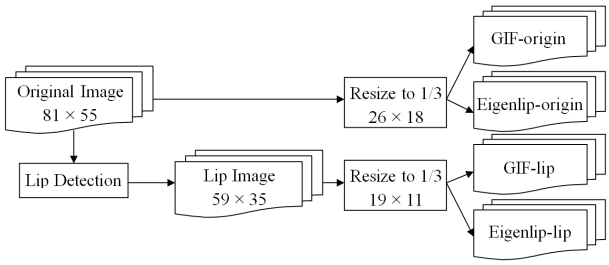


Figure 2: *Feature extraction for conventional (eigenlip) and proposed (GIF) features.*

Table 2: *Phonemes and corresponding visemes.*

phoneme	viseme	phoneme	viseme	phoneme	viseme
a	**a**	j	**sy**	t	**t**
a:		my		d	
i	**i**	ky		n	
i:		by		ts	**s**
u	**u**	gy		z	
u:		ny		s	
e	**e**	hy		y	**y**
e:		ry		k	**vf**
o	**o**	py		g	
o:		ch		h	
p	**p**	dy		N	*N*
b		sh		q	—
m		wf	*w*	sil	**sil**
r	**r**	f			

vector was converted into a feature vector consisting of component scores. GIF vectors were obtained applying transformation matrices computed from the training vectors. Classes for GIF (detailed in Section 2) correspond to visemes (visual phonemes) shown in Table 2 [19], where bold visemes are appeared in Japanese digit utterances, whereas italic ones are not.

4.3. Model training and recognition

We adopted a recognition model as Hidden Markov Model (HMM). We employed conventional model training and recognition techniques, which are widely used in speech recognition. The model training was based on the baseline [18]; at first, a time-aligned transcription was obtained using acoustic features and its models, applying the forced alignment technique. Then a visual HMM was built conducting the Baum-Welch training and using the transcription as well as visual features extracted from training images. In the recognition process, a lipreading recognition result was obtained by the Viterbi algorithm, using the visual HMM and evaluating visual features.

5. Recognition Experiments

In order to evaluate the effectiveness of the proposed method, we conducted two recognition experiments to investigate the usefulness of new visual feature and speaker adaptation.

5.1. Experiments on visual feature comparison

In the first experiment, we investigate the usefulness of GIF in lipreading by comparing with a conventional feature. Experimental condition is described, afterwards, recognition results including discussions are described.

5.1.1. Experimental condition

In order to compute GIF transformation matrices, 2,640 vectors were selected from the training data set; 220 vectors for each class were used. On the other hand, 4,620 vectors were randomly chosen from the training data for eigenlip. For both GIF and eigenlip features, 10-dimensional basic feature vectors were extracted, respectively. In addition to extracted features (static features), first-order and second-order derivatives (dynamic features) were computed and added in the same way as speech recognition. As a result, 30-dimensional feature vectors were finally obtained. Model training and recognition were conducted using HTK (Hidden Markov Model Toolkit) [20]. Models were built for each word corresponding to digit, as well as silence. Each digit HMM consisted of 16 states having 8 Gaussian mixtures, whereas a silence HMM had 3 states and there were 16 mixtures in each state.

In this paper, recognition accuracy (Acc) was used to evaluate a feature:

$$\mathrm{Acc} = \frac{H - I}{N} \tag{8}$$

where H is the number of correctly recognized digits, I is that of insert errors, N is the total number of digits in the label.

When recognition, we tested two conditions: a closed condition where the training data were recognized, and an open condition where the test data were used. An insertion penalty was optimized manually to achieve the best lipreading recognition performance.

5.1.2. Experimental result and discussion

Recognition accuracies for GIF and eigenlip feature in four conditions are shown in Table 3. And recognition results for each subject in the open condition using the original images are shown in Fig. 3; the vertical axis is recognition accuracy[%], and the horizontal axis indicates subjects in the descending order of eigenlip results.

In both closed and open conditions, Table 3 reveals that GIF has the better recognition performance than eigenlip features except one condition. This means the effectiveness of the proposed feature GIF. In addition, recognition accuracies of original images were higher than those of lip images. From the point of view of computational time, it might be said that lip extraction was reasonable since the lip image size was smaller than the original size. However, according to results that the original was better than the lip images, extracting lip images might drop some information about spoken events. Because the image size (computational time) and the performance are trade-off, it is necessary to investigate appropriate settings about them.

Table 3: *Lipreading recognition accuracies* [%].

	image	GIF	Eigenlip
closed	original	63.18	61.14
	lip	53.83	52.66
open	original	39.09	40.97
	lip	32.08	30.03

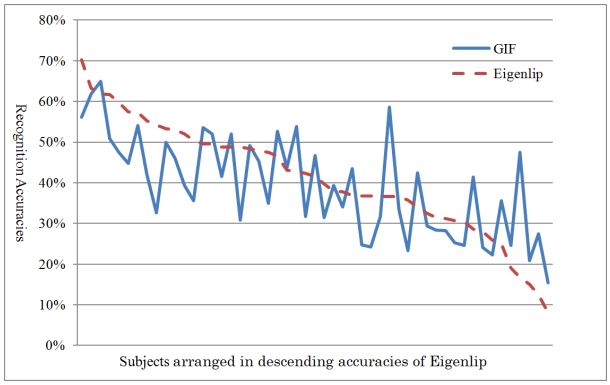


Figure 3: *Recognition accuracy for each subject in the open condition using the original images.*

In Fig. 3, it is obvious that the difference between the highest and lowest subjects in GIF is smaller than that in eigenlip. GIF can achieve more stable performance, and this becomes an advantage for GIF. Nevertheless, there are many subjects whose accuracies of eigenlip were higher than the accuracies of GIF. According to these facts, it may not be appropriate to decide which feature is much superior to another. In order to clarify this, we conducted another experiments appeared in the following.

5.2. Experiments on speaker adaptation

In the second experiment, we investigate the improvement and effectiveness of speaker adaptation. And through the results, we discuss further comparison of lipreading performance before and after adaptation, as

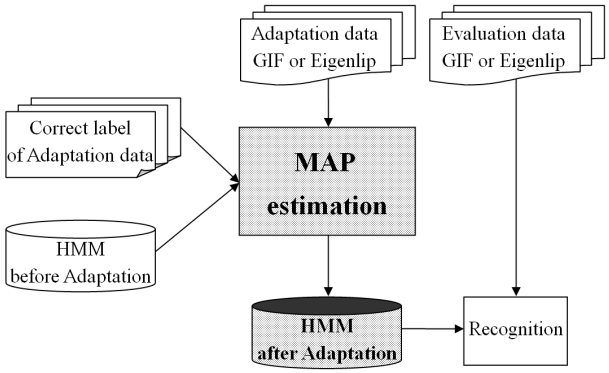


Figure 4: *Speaker adaptation.*

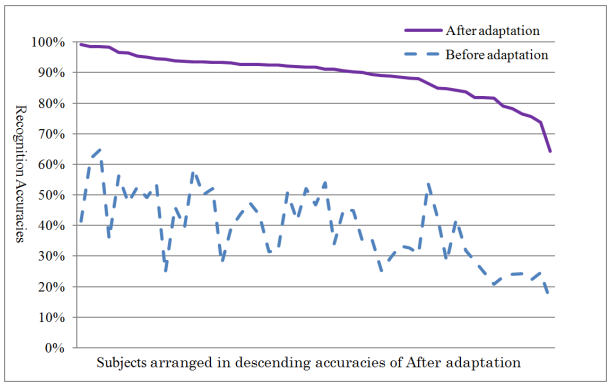


Figure 5: *Recognition accuracy before and after adaptation, using GIF in the closed condition* (I-1).

well as GIF with eigenlip features.

5.2.1. Speaker Adaptation

Fig. 4 illustrates a flow of speaker adaptation. After model training, MAP estimation was applied using a part of evaluating data. In the speaker adaptation by MAP, supervised adaptation was employed; transcribed labels of the adaptation data were used, so as to certainly improve the recognition accuracy. We assumed the situation where a user utilizes a recognition system on her/his own smart phone. Afterwards, lipreading was finally performed using the adapted model and test data.

5.2.2. Experimental condition

A part of test set in Table 1 was used as adaptation data. Two kinds of experiments shown below were conducted.

I. All of the test set were used for adaptation, and the same data were recognized.

II. A half of test data set was used for adaptation, while the other data were recognized. This experiment was conducted only for nine subjects whose accuracies of GIF were much lower than those of eigenlip in the previous experiment.

Note that only the original images were targeted in these experiments.

5.2.3. Experimental result and discussion

Fig. 5 and 6 show recognition results before and after adaptation, using GIF in the closed condition (I-1), and those using eigenlip features (I-2), respectively. In these figures, subjects were appeared in the descending order according to the accuracies after adaptation. At first, it is obvious that in all the results recognition accuracies after adaptation were greatly improved from those before adaptation; for example in Fig. 5, the range of accuracies before adaptation was approximately 20-60%, while the accuracies reached roughly 70-100% after adaptation. These facts reveal the effectiveness of speaker adaptation in lipreading. Secondly, comparing Fig. 5 with Fig. 6, recognition accuracies of GIF was much significantly improved than those of eigenlip features. Although we could not decide which feature is more appropriate in the last discussion, it is found that the GIF-based method achieved better performance than the eigenlip-based method for almost the all subjects after adaptation. This indicates GIF is now superior to eigenlip in lipreading.

Fig. 7 summarizes results in the open condition for the nine subjects (II). Focusing on Fig. 7, it is also found that recognition accuracies of GIF were greatly improved after speaker adaptation, even the GIF-based method had insufficient performances compared to the eigenlip-based method before adaptation.

Note that the adaptation was successful for several speakers whereas there were little improvements for some subjects. Checking several images in the corpus and recognition results, we found that great improvement was obtained for speakers who move their mouth significantly when uttering. In contrast, for subjects who had small mouth movements, the improvement was limited.

6. Conclusion

In this paper, we improve lipreading performance by using GIF and speaker adaptation. The results comparing GIF and conventional eigenlip features show that recognition accuracies of GIF is the same or better than those of eigenlip features. The results of speaker adaptation show, at first, accuracies of both GIF and conventional eigenlip features were significantly improved. Secondly, accuracies of GIF were higher than those of eigenlip features among most subjects.

Our future work is described as follows. Although we employed the supervised adaptation, an unsupervised adaptation technique is expected in real applications,

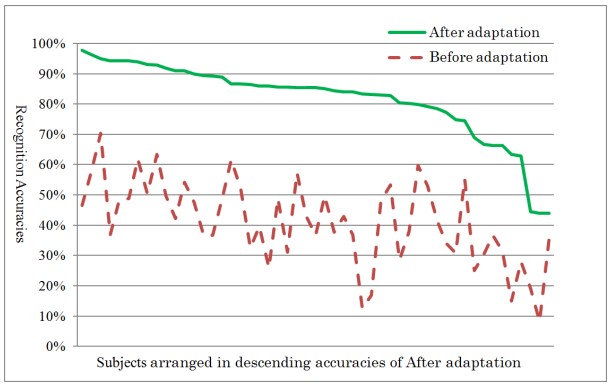


Figure 6: *Recognition accuracy before and after adaptation, using eigenlip in the closed condition* (I-2).

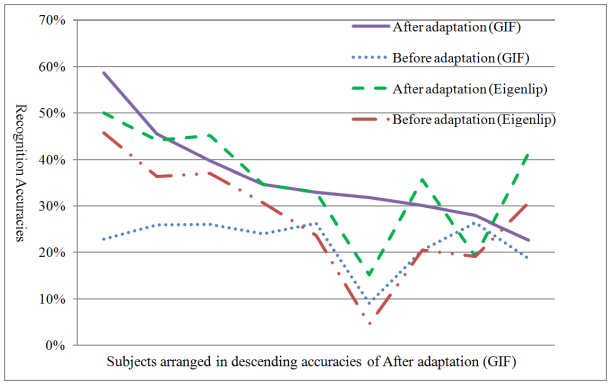


Figure 7: *Recognition accuracy before and after adaptation, using both features in the open condition* (Ⅱ).

which uses several utterances are recorded to adapt the model prior to recognition. So we will develop an unsupervised adaptation. Other adaptation techniques, for example multi-stream adaptation using audio information [21], should be also investigated to achieve further improvements.

7. Acknowledgment

The part of this work was supported by JSPS KAKENHI Grant (Grant-in-Aid for Young Scientists (B)) No.25730109.

8. References

[1] S.Tamura et al., “CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition”, Proc. AVSP2010, pp.85-88 (2010).

[2] C.Bregler et al., ““eigenlips” for robust speech recognition”, Proc. ICASSP’94, vol.2, pp.669-672 (1994).

[3] K.Mase et al., “Automatic lipreading by optical-flow analysis”, Trans. Systems and Computers in Japan, vol.22, no.6, pp-67-76 (1991).

[4] K.Iwano et al., “Bimodal speech recognition using lip movement measured by optical-flow analysis”, Proc. HSC2001, pp.187-190 (2001).

[5] T.Cootes et al., “Active appearance models”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol.23, no.6, pp.681-685 (2001).

[6] C.Neti et al., “Audio-Visual Speech Recognition”, Final Workshop 2000 Report, Center for Language and Speech Processing (2000).

[7] T.Saitoh et al., “Lip Reading Based on Trajectory Feature”, IEICE Trans. Information and Systems, vol.J90-D, no.4, pp.1105-1114 (2007).

[8] Y.Lan et al., “Comparing visual features for lipreading”, Proc. AVSP2009, pp.102-106 (2009).

[9] C.J.Leggetter et al., “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models”, Computer Speech and Language, no.9, pp.171-185 (1995).

[10] JL.Gauvain et al., “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains”, IEEE Trans. Speech and Audio Processing, vol.2, no.2, pp.291-298 (1994).

[11] K.Shinoda, “Speaker adaptation techniques for automatic speech recognition”, Proc. APSIPA ASC 2011 (2011).

[12] S.Tamura et al., “Evaluation of Multi-Modal Speech Recognition Using a Stream-Weight Optimization Based on a Likelihood-Ratio Maximization Criterion”, ASJ 2004 Spring Meeting, vol.1, pp.123-124 (2004).

[13] S.Tamura et al., “Audio-visual interaction in model adaptation for multi-modal speech recognition”, Proc. APSIPA ASC 2011 (2011).

[14] S.Tamura et al., “GIF-SP: GA-based informative feature for noisy speech recognition”, Proc. APSIPA ASC 2012 (2012).

[15] K.Sawada et al., “Statistical voice conversion using GA-based informative feature”, Proc. APSIPA ASC 2012 (2012).

[16] N.Ukai et al., “GIF-LR:GA-based Informative Feature for Lipreading”, Proc. APSIPA ASC 2012 (2012)

[17] P.Viola et al., “Rapid object detection using a boosted cascade of simple features”, Proc. CVPR2001, vol.1, pp.511-518 (2001).

[18] “CENSREC-1-AV manual – How to copy the corpus and how to obtain baseline results”, in CENSREC-1-AV (2011).

[19] Y.Fukuda et al., “Characteristics of the mouth shape in the production of Japanese – Stroboscopic observation”, Journal of Acoustical Society of Japan, vol. 3, no.2, pp.75-91 (1982).

[20] HTK Speech Recognition Toolkit, http://htk.eng.cam.ac.uk/

[21] M.Oonishi et al., “Model Adaptation using Audio-visual Interaction for Multi-modal Speech Recognition”, IEICE. SP, vol 111, no 97, pp.17-22 (2011).