GIF-LR:GA-based Informative Feature for Lipreading

Naoya Ukai, Takumi Seko, Satoshi Tamura and Satoru Hayamizu
Department of Information Science, Gifu University, Gifu, Japan
E-mail: {ukai,seko}@asr.info.gifu-u.ac.jp, tamura@info.gifu-u.ac.jp, hayamizu@gifu-u.ac.jp

***Abstract*—In this paper, we propose a general and discriminative feature "GIF" (GA-based Informative Feature), and apply the feature to lipreading (visual speech recognition). The feature extraction method consists of two transforms, that convert an input vector to GIF for recognition. The transforms can be computed using training data and Genetic Algorithm (GA). For lipreading, we extract a fundamental feature as an input vector from an image; the vector consists of intensity values at all the pixels in an input lip image, which are enumerated from left-top to right-bottom. Recognition experiments of continuous digit utterances were conducted using an audio-visual corpus including more than 268,000 lip images. The recognition results show that the GIF-based method is better than the baseline method using eigenlip features.**

I. Introduction

Speech recognition technology has been developed for last decades and widely utilized in various fields. However, the recognition accuracy is still drastically degraded in noisy or real environments, e.g. in car conditions. 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 and lip images can be easily obtained. In the audio-visual speech recognition, a lipreading technology that estimates what word or sentence is pronounced only using visual data is embedded to enhance the robustness of speech recognition. The rest of this paper focuses on the lipreading technology.

Lipreading has also been explored, in particular, visual feature is one of the major topics. Many features for lipreading have been proposed, for example, eigenlip [1], Discrete Cosine Transform (DCT) coefficients, optical-flow-based features [2], [3], and Active Appearance Model (AAM) features [4], [5]. Nevertheless, the lipreading recognition accuracy is not significant even in controlled conditions; for continuous digit or word recognition task in audio-visual speech recognition, the accuracy of speech recognition in clean condition has achieved almost 100%, whereas the accuracy of lipreading is still roughly 50% [6]. Note that in general it is difficult for lipreading to perfectly recognize utterances since visual information do not contain all the information about human speech activities. Taking the fact into account, however, the accuracy should be further improved.

In this paper, we propose a new discriminative feature called GA-based Informative Feature (GIF). In the extraction method, an input vector is converted into an output GIF vector by using two transformation matrices. The transformation matrices are obtained using Genetic Algorithm (GA) and labeled training data. We apply our proposed feature GIF to lipreading, in order to improve the lipreading accuracy. Intensity values in an image is assembled into an input vector, subsequently GIF for lipreading is computed. Hidden Markov Model (HMM), which is widely used in speech recognition, is built using training data consisting of GIF vectors. In order to evaluate our proposed feature, lipreading recognition experiments are conducted comparing GIF with conventional lipreading features such as eigenlip features.

This paper is organized as follows: In Section II, our proposed feature GIF is introduced. Section III describes a lipreading method and database used in this paper, followed by lipreading recognition experiments and results. Finally Section IV concludes this paper.

II. GA-based Informative Feature

In this section, our proposed feature GIF (GA-based Informative Feature) is introduced. This feature can be utilized in various pattern recognition tasks and related works [7], [8].

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. 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. These matrices A and B are computed and optimized by Genetic Algorithm (GA). By employing GA, we can deal with any optimization problem, using not only linear but also non-linear and non-differentiable functions.

A. Stage 1: Getting a first transformation matrix

A binary classifier which distinguishes an input vector into a certain class or its complementary class is focused on. The transformation matrix A is eventually obtained by the following process for all classes.

1) Step 1: Building a candidate classifier:

For an i-th class ($1 \leq i \leq C$), assume a linear classifier f for an input vector $\boldsymbol{x} = (x_j)$:

$$f(\boldsymbol{x}; \boldsymbol{a}_i) = \left(\sum_{j=1}^{N} a_{i,j} x_j \right) + a_{i,N+1} \tag{3}$$

where $\boldsymbol{a}_i = (a_{i,1}, \cdots, a_{i,N}, a_{i,N+1})$ includes classifier parameters corresponding to a part of A. The classifier is designed to return a positive value if $\boldsymbol{x}$ is in the class or a negative value otherwise. The classifier parameters are computed using a training set $R = \{\boldsymbol{r}_n\}$ and the standard GA:

1. Initialization
   An initial population G_0 including K individuals is created. An individual has $(N+1)$ chromosomes, each of which encodes a classifier parameter.
2. Fitness function
   For a k-th individual $\boldsymbol{v}_k$ in an h-th generation G_h, a fitness function $E(\boldsymbol{v}_k)$ is calculated as:
   $$E(\boldsymbol{v}_k) = \sum_{n=1}^{|R|} l_n \cdot sgn\big(f_i(\boldsymbol{r}_n; \boldsymbol{a})\big) \tag{4}$$
   where $\boldsymbol{a}$ is a parameter set obtained by decoding $\boldsymbol{v}_k$, and l_n is a transcribed label that corresponds to 1 if $\boldsymbol{r}_n$ belongs to the class or -1 otherwise. The minimum value of $E(\boldsymbol{v}_k)$ is set to 1.
3. Elitist selection, inheritance, mutation and crossover
   Conventional GA operations are employed to form a next generation G_{h+1} from a current generation G_h; elitist selection and inheritance are applied to copy a certain individual to G_{h+1}; for genetic diversity, mutation and crossover operations are also conducted to generate a new individual that is different from its parent(s).
4. Generation change
   The above processing (2. and 3.) is repeated from $h=0$ to $h=F-1$. A final population G_F is then generated.

2) Step 2: Completing a binary classifier:

From G_F, individuals having the (K/I) highest fitness values are extracted and added to a candidate population G_C. By repeating the step I times, the candidate population is completed. Step 1 is then applied again where G_C is used as an initial generation, then the best-fit individual $\hat{\boldsymbol{v}}$ is obtained. The transformation parameter set $\hat{\boldsymbol{a}}$ is consequently acquired by decoding the solution $\hat{\boldsymbol{v}}$. It is often pointed out that the solution is not stable since GA highly depends on an initial population and operations randomly determined. The two-step approach enables us to compute a stable GA solution.

B. Stage 2: Getting a second transformation matrix

To enhance the discriminative and recognition ability, and to reduce the dimension of feature vectors, we employ a second-stage procedure explained below:

1. For an i-th class, a mean vector $\bar{\boldsymbol{\mu}}_i$ is calculated as:
   $$\bar{\boldsymbol{\mu}}_i = \frac{1}{|R_i|} \sum_{\boldsymbol{r} \in R_i} A\big(\boldsymbol{r}^\top \ 1\big)^\top \tag{5}$$
   where R_i is a subset of the training data, in which all vectors belong to the i-th class.
2. Let us denote a linear transformation g, for a vector $\boldsymbol{y} = (y_j)$ obtained in the first stage, by the following equation:
   $$g(\boldsymbol{y}; \boldsymbol{b}_m) = \sum_{j=1}^{C} b_{m,j} y_j \tag{6}$$
   where $\boldsymbol{b}_m = (b_{m,1}, \cdots, b_{m,C})$ indicates classifier parameters and a part of B.
3. For $m = 1$, the projection g is determined so that a variance of transformed mean vectors would be maximized. The parameter set $\boldsymbol{b}_1$ is optimized by applying GA explained previously, where the fitness function is modified as:
   $$E(\boldsymbol{v}_k) = var(w_1, \cdots, w_C) \ \text{ where } \ w_i = g(\bar{\boldsymbol{\mu}}_i; \boldsymbol{b}) \tag{7}$$
   In Eq.(7), $\boldsymbol{b}$ is obtained by decoding $\boldsymbol{v}_k$.
4. For $m=2$, $\boldsymbol{b}_2$ is optimized so as to maximize a variance just as same as $\boldsymbol{b}_1$, under the constraint that an inner product between ${\boldsymbol{b}_1}^\top$ and ${\boldsymbol{b}_2}^\top$ is zero:
   $$\left\langle {\boldsymbol{b}_1}^\top, {\boldsymbol{b}_2}^\top \right\rangle = \sum_{j=1}^{C} b_{1,j} b_{2,j} = 0 \tag{8}$$
5. For any m $(2 \leq m \leq M)$, an m-th parameter set $\boldsymbol{b}_m$ is calculated in the same way, under the restriction that all inner products should be zero.

C. Feature vector computation

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:

$$\boldsymbol{\mu} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{y}_t = \frac{1}{T} \sum_{t=1}^{T} A\big({\boldsymbol{x}_t}^\top \ 1\big)^\top \tag{9}$$

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:

$$\hat{\boldsymbol{y}}_t = \boldsymbol{y}_t - \boldsymbol{\mu} \tag{10}$$

III. Recognition experiment

In order to evaluate the effectiveness of the proposed method in lipreading, we conducted lipreading recognition experiments. In this section, at first, a database, feature extraction as well as a training/recognition scheme are introduced. Secondly, experimental condition is described. Recognition results including discussions are finally appeared.

TABLE I
A SUMMARY OF IMAGE DATA IN THE CENSREC-1-AV DATABASE.

frame rate		29.97Hz (NTSC)
image size / depth		width 81 pixel × 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

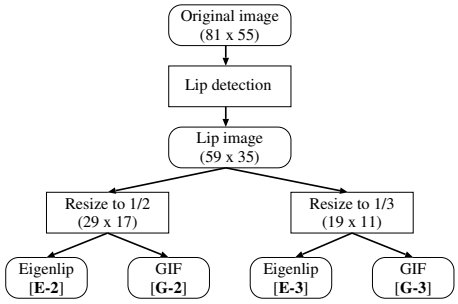

Fig. 1. Feature extraction for conventional (eigenlip) and proposed (GIF) features.

A. Database

We used image data in a corpus CENSREC-1-AV [9] that is built for audio-visual speech recognition as well as lipreading. The database includes speech waveforms and image data of Japanese continuous digit utterances recorded in office environments. Regarding image conditions, a subject uttered digits in front of a blue screen, then face pictures were captured by a optical camera. Note that images around subjects' mouth (lower part of a face) were available in CENSREC-1-AV, manually extracted with a fixed window. CENSREC-1-AV has two kinds of image data, color (optical) and infrared, and in this paper only the former is employed.

The database consists of two data sets: a training data set for building recognition models and feature transforms, and a test data set for evaluating a lipreading method. The specification is summarized in TABLE I.

B. Feature extraction

Fig.1 depicts a flow of feature extraction of conventional and proposed features. At first, lip detection was performed to extract lip images from gray-scale pictures in the database. An AdaBoost method using Haar-like features [10] was adopted to determine the extraction window. A lip area could be rarely detected in several pictures, and in such the cases, the previous window was successively used. Note that all the images obtained have the same image size (59×35). After extraction, in order to reduce computational resources in the following processes, an extracted image was resized to half (29×17) and one third (19×11) of the image, respectively. Fig.2 shows a sample image obtained from image in the CENSREC-1-AV database.

A 493-dimensional (half) or 209-dimensional (one-third) input vector was then computed from intensity values in an image, enumerating all the pixels from left-top to right-bottom. For eigenlip features **E-2** and **E-3**, we referred to a baseline system in CENSREC-1-AV [11]; several training vectors were chosen in order to conduct Principal Component Analysis (PCA) and obtain eigenvectors. Using the eigenvectors, each input vector is converted into a feature vector consisting of component scores. For the proposed features **G-2** and **G-3**, transformation matrices were computed using training vectors. An output feature is finally obtained applying the matrices. Classes in Section II correspond to visemes (visual phonemes) shown in TABLE II [12], where bold visemes are appeared in Japanese digit utterances, whereas italic ones are not.

Fig. 2. A sample obtained lip image.

TABLE II
VISEMES AND CORRESPONDING PHONEMES.

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

† **bold** is used in this paper, whereas *italic* is not.

C. Model training and recognition

Conventional HMM-based training and recognition were employed for lipreading. Model training was based on the baseline [11]; at first, time-aligned transcription was obtained using acoustic features and its models, applying the forced-alignment technique. Then visual HMM was built conducting the Baum-Welch training and using the transcription and visual features explained in the last subsection. When recognition, a lipreading recognition result was obtained by the Viterbi algorithm, using the visual HMM and testing visual features.

D. Experimental condition

In feature extraction, 4,620 vectors were randomly chosen from the training data to build eigenvectors, whereas 2,640 vectors were used to compute GIF transformation matrices. After 10-dimensional eigenlip and GIF vectors were extracted, first-order and second-order derivatives were computed and added, similar to conventional speech recognition. As a result, 30-dimensional feature vectors were finally obtained. TABLE III shows GIF parameters appeared in Section II.

HMMs were constructed for all the digits and 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. 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

TABLE III
GIF PARAMETERS USED IN THE EXPERIMENTS.

feat.	image size	K	F	I	N	C	M
G-2	29×17 (half)	2000	40	30	493	13	10
G-3	19×11 (one-third)	1000	30	30	209		

TABLE IV
LIPREADING RECOGNITION ACCURACIES OF FOUR FEATURES.

cond.	image size	eigenlip(%)	GIF(%)
closed	29×17 (half)	49.91 (**E-2**)	59.79 (**G-2**)
	19×11 (one-third)	52.66 (**E-3**)	53.83 (**G-3**)
open	29×17 (half)	27.49 (**E-2**)	35.73 (**G-2**)
	19×11 (one-third)	30.03 (**E-3**)	32.08 (**G-3**)

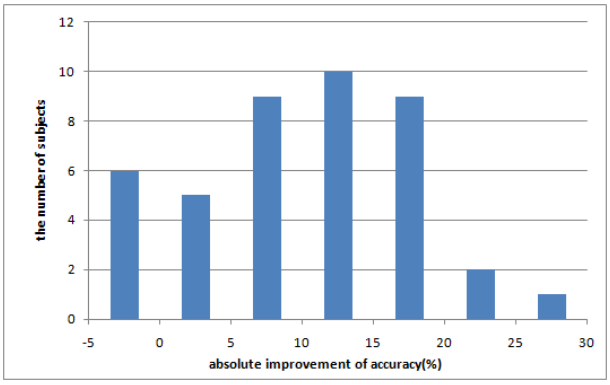

Fig. 3. A histogram of recognition improvements from **E-2** to **G-2** in the closed condition.

achieve the best lipreading recognition performance. The other experimental settings were the same as [9], [11].

In this paper, recognition accuracy (Acc) is used to evaluate a feature, calculated by Eq.(11). Here, H is the number of correctly recognized digits, I is that of insert errors, N is the total number of digits in the label.

$$Acc = \frac{H - I}{N} \tag{11}$$

E. Experimental result

TABLE IV shows lipreading recognition results. In the closed condition, the proposed features achieved 9.88% (**G-2** vs **E-2**) and 1.17% (**G-3** vs **E-3**) improvements respectively. These results indicate that our feature can achieve better lipreading performance than eigenlip features. Eigenlip features were extracted without taking viseme classes into account, on the other hand, GIF was computed after each viseme classification was applied. This is one of the reason why our feature was better than eigenlip. Regarding the other visual features, many conventional features were tested in a word recognition task, and the result shows AAM-based features achieved 54–59% recognition accuracy, followed by eigenlip (45%) and DCT (31%) [6]. The AAM-based features employed PCA which is compared in this experiment. Furthermore, such the model-based features might be damaged when modeling is failed. From the above discussion, using our feature extraction method has a possibility to improve the lipreading accuracy from the existing method. In addition, it is necessary to compare our method with state-of-the-art methods, such as lip reading using AAM[13].

It is mentioned that human lipreading ability is roughly 70% in average in a non-context task [2]. Compared to this upper limit, the lipreading accuracy of our proposed feature is considered to be reasonable.

Fig.3 indicates a histogram of absolute improvement of accuracy in the closed condition, comparing the proposed feature **G-2** and eigenlip feature **E-2**. For most subjects approximately 5–20% improvements were obtained, on the other hand, more than 20% improvements were also observed. And for three subjects the performances decreased. The reason such the large difference was occurred might be that the amount of mouth movements was different among subjects; some subjects moved their mouth only a little, or the others moved clearly. The adaptation method widely used in speech recognition [14] is one of the possible solutions.

IV. CONCLUSION

In this paper, we propose a new discriminative feature GIF, and apply the feature to lipreading. In GIF extraction, two transformations are conducted that are estimated by GA. Lipreading recognition experiments were conducted in closed and open conditions, comparing conventional eigenlip features. The results show the effectiveness of our proposed feature in lipreading.

Our future work includes improvement of input features, not the intensity values but other informative values. Model or feature adaptation, and application of the lipreading method to audio-visual speech recognition are also included in our future work.

REFERENCES

[1] C.Bregler et al., ""eigenlips" for robust speech recognition," Proc. ICASSP'94, vol.2, pp.669-672 (1994).
[2] K.Mase et al., "Automatic lipreading by optical-flow analysis," Trans. Systems and Computers in Japan, vol.22, no.6, pp-67-76 (1991).
[3] K.Iwano et al., "Bimodal speech recognition using lip movement measured by optical-flow analysis," Proc. HSC2001, pp.187-190 (2001).
[4] T.Cootes et al., "Active appearance models," IEEE Trans. Pattern Analysis and Machine Intelligence, vol.23, no.6, pp.681-685 (2001).
[5] C.Neti et al., "Audio-Visual Speech Recognition," Final Workshop 2000 Report, Center for Language and Speech Processing (2000).
[6] Y.Lan et al., "Comparing visual features for lipreading," Proc. AVSP2009, pp.102-106 (2009).
[7] S.Tamura et al., "GIF-SP: GA-based informative feature for noisy speech recognition," Proc. APSIPA ASC 2012 (2012).
[8] K.Sawada et al., "Statistical voice conversion using GA-based informative feature," Proc. APSIPA ASC 2012 (2012).
[9] S.Tamura et al., "CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition," Proc. AVSP2010, pp.85-88 (2010).
[10] P.Viola et al., "Rapid object detection using a boosted cascade of simple features," Proc. CVPR2001, vol.1, pp.511-518 (2001).
[11] "CENSREC-1-AV manual – How to copy the corpus and how to obtain baseline results," in CENSREC-1-AV (2011).
[12] 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).
[13] T.F.Cootes et al., "Active Appearance Models," Proc. ECCV'98, vol.2, pp.484-498 (1998).
[14] K.Shinoda, "Speaker adaptation techniques for automatic speech recognition," Proc. APSIPA ASC 2011 (2011).