Audio-visual speech recognition using deep bottleneck features and high-performance lipreading

Satoshi TAMURA*, Hiroshi NINOMIYA†, Norihide KITAOKA‡, Shin OSUGA§,
Yurie IRIBE¶, Kazuya TAKEDA† and Satoru HAYAMIZU*
* Gifu University, Japan E-mail: {tamura@info., hayamizu@}gifu-u.ac.jp
† Nagoya University, Japan E-mail: {ninomiya.hiroshi@g.sp.m., takeda@}is.nagoya-u.ac.jp
‡ Tokushima University, Japan E-mail: kitaoka@is.tokushima-u.ac.jp
§ Aisin Seiki Co., Ltd., Japan E-mail: sohsuga@elec.aisin.co.jp
¶ Aichi Prefectural University, Japan E-mail: iribe@ist.aichi-pu.ac.jp

***Abstract*—This paper develops an Audio-Visual Speech Recognition (AVSR) method, by (1) exploring high-performance visual features, (2) applying audio and visual deep bottleneck features to improve AVSR performance, and (3) investigating effectiveness of voice activity detection in a visual modality. In our approach, many kinds of visual features are incorporated, subsequently converted into bottleneck features by deep learning technology. By using proposed features, we successfully achieved 73.66% lipreading accuracy in speaker-independent open condition, and about 90% AVSR accuracy on average in noisy environments. In addition, we extracted speech segments from visual features, resulting 77.80% lipreading accuracy. It is found VAD is useful in both audio and visual modalities, for better lipreading and AVSR.**

I. Introduction

Automatic Speech Recognition (ASR) has been widely spread, and today, many devices have speech interfaces using ASR technology. However, a crucial problem still remains that recognition performance severely degrades in noisy or real environments. As one of methods to compensate the degradation, Audio-Visual Speech Recognition (AVSR), namely bimodal or multi-modal speech recognition, has been studied for a couple of decades. Since a lip image sequence is not basically affected by acoustic noise, visual information is expected to help a recognizer so as to achieve better performance.

Meanwhile, Deep Learning (DL) has attracted a lot of attentions of researchers in many pattern recognition fields including computer vision and speech recognition. There are two basic strategies to apply DL to ASR systems: a hybrid approach [1] and a tandem approach [2]. In the former approach, Deep Neural Networks (DNNs) are built to estimate posteriori probabilities on Hidden Markov Model (HMM) states for test data. This strategy is called DNN-HMM. On the other hand, in the latter approach, DNNs are used to generate new features from input ones. Here HMMs having Gaussian Mixture Models (GMMs), named GMM-HMM, are usually adopted for recognition. For conventional ASR, many studies have been done, showing that both strategies are effective to improve ASR accuracy [3], [4], [5].

There are several works using DL technology in AVSR. For instance, a bimodal deep audoencoder was proposed to obtain multi-modal feature vectors [6]. A deep brief network was also utilized performing middle-level feature combination [7]. In terms of recognition model, multi-stream HMMs, that is often employed in AVSR, were built using features obtained by deep denoising autoencoder [8]. We also developed an AVSR method using Deep BottleNeck Features (DBNFs) based on the tandem approach, and tested our method in noisy environments [9]. As a result, we could improve an AVSR performance method using audio and visual DBNFs compared not only to audio-only ASR but also to a conventional audio-visual baseline system.

In order to further improve the performance, however, visual speech recognition (lipreading) must be still investigated. In noisy conditions, AVSR performance depends on visual recognition ability, however, visual-only recognition accuracy is quite insufficient: 39.3% word accuracy for a digit recognition task when only using visual DBNFs derived from Principal Component Analysis (PCA) features [9]. It is also reported that the performance is roughly 27-59% in speaker-independent condition [10]. Such the performance is roughly equivalent to those obtained in SNR 5-15dB acoustically noisy environments for conventional ASR [11]. Therefore, finding effective visual features is one of key issues to improve not only lipreading but AVSR performance.

Many researchers have proposed and investigated various features for lipreading or audio-visual ASR; PCA also known as "eigenlip" [12], 2D Discrete Cosine Transform (DCT) and Linear Discriminant Analysis (LDA) e.g. [13], have been often employed. Because lip movements are much effective to identify visual units (visemes) and to detect visual activities, optical flow is sometimes used [14], [15]. All the above features are appearance-based, on the other hand, some shape-based features are also considered. For instance, width and height of one's lip are basic shape-based parameters, and lip contour information is sometimes utilized. An active appearance model or any other face model is often chosen to extract shape parameters for lipreading, e.g. [16], [17].

In this paper, we aim at improving AVSR performance by investigating the following aspects: (1) combining basic visual features and subsequently applying our DBNF method to obtain high-performance features for visual speech recognition, (2) using the new visual features and audio DBNFs to achieve a better AVSR system, and (3) performing visual Voice Activity Detection (VAD) to avoid recognition errors in silence periods for lipreading. Novelties of this paper thus lies in effectiveness of incorporating several basic features and applying DBNF techniques to the combined features, further improvement of AVSR from our previous work [9] by using our new visual DBNFs in addition to audio DBNFs, and importance of VAD not only for an audio modality but a visual modality.

The rest of this paper is organized as follows. Section II briefly describes DL-based AVSR. Several kinds of basic visual features for visual DBNF are introduced in Section III. Section IV shows database, experimental setup, result, and discussion. Finally Section V concludes this paper.

II. Audio-visual speech recognition with deep learning

In our AVSR scheme, we employ multi-stream HMMs that can control contributions of audio and visual modalities according to recognition environments. We also compute audio and visual features using the tandem approach. Both methods are briefly introduced in this section. In the following description, let us denote audio DBNF and visual DBNF by **DBAF** (Deep Bottleneck Audio Feature) and **DBVF** (Deep Bottleneck Visual Feature), respectively.

A. Multi-stream HMM

In most successful AVSR systems, firstly audio and visual features are separately extracted from audio signals and facial region of interests in visual image sequences, respectively. Both features are secondly concatenated into audio-visual features. Then, multi-stream HMMs are applied to audio-visual features. A conventional multi-stream HMM in AVSR has two streams, an audio stream and a visual stream, in addition to corresponding stream weight factors, λ_a and λ_v. For an audio-visual feature $\boldsymbol{f}_{avt}$ at time t, a log likelihood $b_{av}(\boldsymbol{f}_{avt})$ is computed by Eq.(1):

$$b_{av}(\boldsymbol{f}_{avt}) = \lambda_a b_a(\boldsymbol{f}_{at}) + \lambda_v b_v(\boldsymbol{f}_{vt}) \quad (1)$$

where $b_a(\boldsymbol{f}_{at})$ and $b_v(\boldsymbol{f}_{vt})$ are audio and visual log likelihoods for an audio feature $\boldsymbol{f}_{at}$ and a visual feature $\boldsymbol{f}_{vt}$ respectively, and $\boldsymbol{f}_{avt} = (\boldsymbol{f}_{at}{}^\top \boldsymbol{f}_{vt}{}^\top)^\top$. In most schemes, stream weight factors are subject to:

$$\lambda_a + \lambda_v = 1 \ , \ 0 \leq \lambda_a, \lambda_v \leq 1 \quad (2)$$

Stream weights should be determined according to noise environments and visual conditions, using some criteria, e.g. [18], [19], or empirically predefined.

B. Deep Bottleneck feature

Today DNN has been rapidly employed contributing to great success in many kinds of pattern recognition tasks. In this paper, we employ a DNN as a feature extractor. Figure 1 depicts a DNN used in this work. An input layer corresponds to an input vector. A current feature vector $\boldsymbol{f}_t$ in addition to previous and incoming several feature vectors $\boldsymbol{f}_{t-T}, \cdots, \boldsymbol{f}_{t-1}, \boldsymbol{f}_{t+1}, \cdots, \boldsymbol{f}_{t+T}$ are concatenated to one vector as the input vector. An output layer is designed to match the input feature, otherwise, assigned to classification results. In our case, the output layer corresponds to all HMM states appeared in a recognition model. In our tandem approach there is a bottleneck hidden layer, having few units compared to the other hidden layers. A feature vector is then composed from outputs obtained from all the units in the bottleneck layer. DNN training consists of two stages: pre-training and fine-tuning; unsupervised pre-training is conducted in a layer-wise manner [20], before all the parameters are fine-tuned [21].

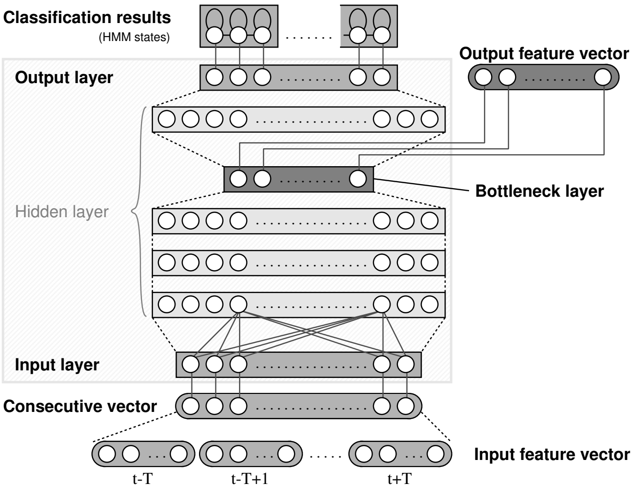


Fig. 1. A DNN for bottleneck feature extraction.

In this work, an audio DNN and a visual DNN are respectively built. For audio feature extraction, we firstly prepared conventional Mel-Frequency Cepstral Coefficients (MFCCs). An audio GMM-HMM is secondly trained using training features. Frame-level state alignments for the training data are thirdly obtained. An audio DNN for DBAF is then built using audio features each which has MFCC vectors in consecutive frames. A visual DNN for DBVF is also obtained as well, except that basic visual features are used instead of MFCCs. To obtain high-performance DBVFs, it is crucially important to employ good basic visual features.

III. Visual features

In this section, we introduce four appearance-based features (PCA, DCT, LDA and GIF) and one shape-based feature (COORD) as well as concatenated features for DBVF. In the following description, let us denote an N-dimensional input image vector at frame t by $\boldsymbol{v}_t = (v_{x,y})$ having intensity values of every pixels $v_{x,y}$ in an image, and the dimension of output feature vectors by M. In some discriminative schemes, the number of classes we should classify is indicated as C.

A. PCA

PCA is one of most common methods in the pattern recognition domain. A covariance matrix of training feature vectors is decomposed to orthogonal vectors (eigenvectors) with corresponding variances (eigenvalues). A transformation matrix A_p is then obtained by choosing M eigenvectors that have larger eigenvalues. Now we compute an M-dimensional feature vector $\boldsymbol{f}_t^{(PCA)}$ from an input feature in Eq.(3):

$$\boldsymbol{f}_t^{(PCA)} = A_p \cdot \boldsymbol{v}_t \tag{3}$$

B. DCT

DCT is also well known in various signal processing and pattern recognition fields, since DCT provides efficiently compressed representations. Like JPEG that is a famous image format, 2D DCT is conducted to an image. After resizing an image to $S \times S$, a DCT coefficient $d_{i,j}$ is computed in the following Eq.(4).

$$d_{i,j} = c_i c_j \sum_{x=1}^{S} \sum_{y=1}^{S} v_{x,y} \cos\left\{\frac{\pi(2x-1)i}{2S}\right\} \cos\left\{\frac{\pi(2y-1)j}{2S}\right\} \tag{4}$$

where

$$c_i = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } i = 0 \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

A feature vector $\boldsymbol{f}_t^{(DCT)}$ is hereby generated by picking up low-dimensional components in a zigzag manner.

C. LDA

LDA is also a famous method in pattern recognition, which provides discriminative transformation. To conduct LDA, training data and corresponding transcription labels are prepared beforehand. According to the labels, at first we calculate a covariance matrix S_i for an i-th class, as well as a global covariance matrix S_G. Secondly, within- and between-class scatter matrices (S_W and S_B respectively) are obtained as Eq.(6):

$$S_W = \sum_{i=1}^{C} S_i \ , \ S_B = S_G - S_W \tag{6}$$

Thirdly, PCA is applied to $S_W^{-1} S_B$ so as to get a transformation matrix A_l. Finally, a feature vector is appeared in Eq.(7):

$$\boldsymbol{f}_t^{(LDA)} = A_l \cdot \boldsymbol{v}_t \tag{7}$$

D. GIF

Some of authors have proposed a feature extraction method, called GA-based Informative Feature (GIF), in which transformation matrices are generated using a genetic algorithm [22], [23]. Similar to LDA, GIF requires a training data set and its label. In GIF, an input vector is converted to a C-dimensional intermediate vector as:

$$\boldsymbol{y}_t = G_1 \cdot (\boldsymbol{v}_t{}^\top \ 1)^\top \tag{8}$$

In Eq.(8), G_1 is a $C \times (N+1)$ matrix, such that a j-th row vector performs a binary classifier; a positive value should be observed if the input vector belongs to the j-th class, otherwise a negative value must appear. Next, a feature vector $\boldsymbol{z}_t$ is computed in Eq.(9):

$$\boldsymbol{z}_t = G_2 \cdot \boldsymbol{y}_t \tag{9}$$

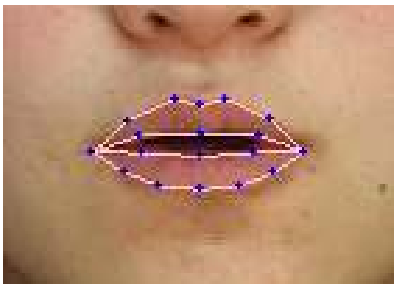


Fig. 2. An example of lip images (mouth detection results) with lip feature points.

where G_2 is an $M \times C$ matrix ($M < C$), performing orthogonalization and dimension reduction. These matrices are determined and optimized using a genetic algorithm and training data. Note that based on preliminary experiments, the first transformation is only applied in this paper; a visual feature vector $\boldsymbol{f}_t^{(GIF)}$ is now obtained as Eq.(10):

$$\boldsymbol{f}_t^{(GIF)} = G_1 \cdot (\boldsymbol{v}_t{}^\top \ 1)^\top \tag{10}$$

E. COORD – Shaped-based feature

To extract a mouth region from a frontal-face image and to employ shape-based features for lipreading, automatic mouth detection and lip feature point estimation are conducted in this paper. Our method includes face tracking and model fitting techniques [24]; here, the scheme is briefly introduced. In the face tracking, a Constrained Local Model (CLM) [25], [26], that is a 3D face model having eyes, nose and mouth, is firstly fitted to a 2D image. Next, 3D face pose and location are estimated, especially mouth information is utilized for the following process. Of lip contours, 18 feature points are detected using a linear regression function based on Haar-like features, while mouth model fitting is performed in which a 3D statistical mouth model is associated with a 2D image applying a CLM-based scheme. Figure 2 illustrates an example of mouth detection and lip feature point extraction results.

After obtaining the mouth feature points, a center-of-gravity point (x_t^C, y_t^C) is computed as Eq.(11):

$$x_t^C = \frac{1}{18} \sum_{i=1}^{18} x_t^i \ , \ y_t^C = \frac{1}{18} \sum_{i=1}^{18} y_t^i \tag{11}$$

where (x_t^i, y_t^i) is an i-th feature point (i=1,2,$\cdots$,18). Relative coordinates of all the feature points are simply concatenated as a 36-dimensional shape-based vector:

$$\boldsymbol{s}_t = \left({x'}_t^1, {y'}_t^1, {x'}_t^2, {y'}_t^2, \cdots, {x'}_t^{18}, {y'}_t^{18} \right)^\top \tag{12}$$

where

$${x'}_t^i = x_t^i - x_t^C \ , \ {y'}_t^i = y_t^i - y_t^C \tag{13}$$

We further tried to apply either of PCA, LDA and GIF to $\boldsymbol{s}_t$, in order to achieve better performance. As a result, GIF is adopted to obtain a feature vector $\boldsymbol{f}_t^{(COORD)}$.

Fig. 3. An example of frontal-face images used in this paper.

F. Concatenated features

In this work, two visual features are prepared, having the above basic visual ones: PCA, DCT, LDA, GIF and COORD. At first, the former four appearance-based features were concatenated into a new feature vector $\boldsymbol{f}_t^{(PDLG)}$ as Eq.(14).

$$\boldsymbol{f}_t^{(PDLG)} = \left(\boldsymbol{f}_t^{(PCA)^\top} \boldsymbol{f}_t^{(DCT)^\top} \boldsymbol{f}_t^{(LDA)^\top} \boldsymbol{f}_t^{(GIF)^\top} \right)^\top \tag{14}$$

Similarly, all the vectors were combined to compose a feature vector $\boldsymbol{f}_t^{(PDLGC)}$ as Eq.(15).

$$\boldsymbol{f}_t^{(PDLGC)} = \left(\boldsymbol{f}_t^{(PDLG)^\top} \boldsymbol{f}_t^{(COORD)^\top} \right)^\top \tag{15}$$

IV. Experiment

In order to evaluate visual features and AVSR performance, we conducted two recognition experiments: (1) visual speech recognition (lipreading) using either of visual features described in Section III or DBVF, and (2) audio-visual speech recognition using enhanced DBVFs introduced in this paper, in addition to DBAFs. Furthermore, we examined another aspect: (3) lipreading excluding non-speech segments, in order to investigate importance of VAD for the visual modality.

A. Database

A Japanese audio-visual corpus CENSREC-1-AV was used [27]. CENSREC-1-AV is designed to evaluate audio-visual speech recognition but is still available and suitable for lipreading, providing audio-visual data as well as a baseline system. In total, 93 subjects spoke connected-digit utterances, making a 42-subject 3,234-utterance training set and a 51-subject 1,963-utterance test set.

Mouth images were included in CENSREC-1-AV, however, we employed their original frontal-face images (720×480) in order to apply the mouth detection and lip feature point extraction methods mentioned in Section III-E. A sample of original frontal-face images is shown in Figure 3. We manually annotated feature points in hundreds of training images to build our feature point extraction model. The size of lip images was fixed as 140×100, of which central point just corresponded to (x_t^C, y_t^C). Note that any preprocessing such as scaling and rotation normalization was not conducted.

TABLE I
Training and test data sets.

	Training set †	Test set ‡
Audio	clean, cityroad (5 SNRs), expressway (5 SNRs), music (5 SNRs)	clean, cityroad (6 SNRs), expressway (6 SNRs), music (6 SNRs), music+cityroad (6 SNRs), music+expressway (6 SNRs)
Visual	clean	clean

† 5 SNRs = 20dB, 15dB, 10dB, 5dB and 0dB

‡ 6 SNRs = 20dB, 15dB, 10dB, 5dB, 0dB and -5dB

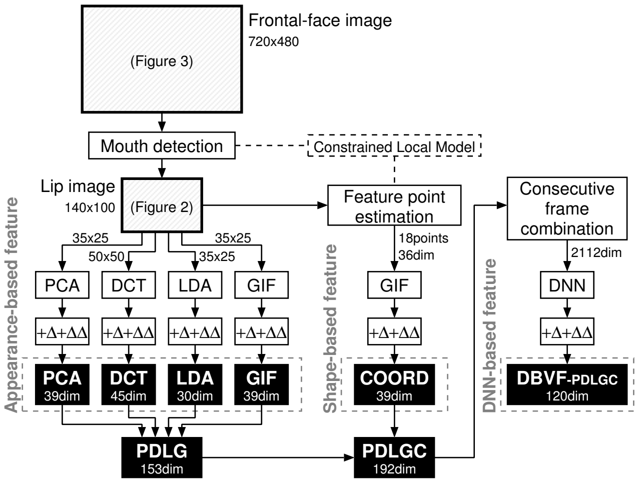


Fig. 4. Visual feature extraction.

B. Features

As mentioned, MFCCs were used to obtain DBAFs. To train an audio DNN and a multi-stream GMM-HMM, and to evaluate AVSR in different noise conditions, not only clean data but noisy data were prepared. Interior car noises recorded on city roads and expressways were added to clean speech data at several SNR levels (20dB, 15dB, 10dB, 5dB, 0dB and -5dB). Assuming a situation using a car-stereo system, we also prepared musical waveforms as another noise. We generated six music-overlapped speech data having different SNR levels as well. In addition, two kinds of noisy speech data, in which not only musical sounds but city-road or expressway noises existed, were similarly added to clean data. As a result, clean speech data and 30 kinds of noisy speech data were prepared. A training data set consisted of clean speech data as well as city-road, expressway, and music-overlapped noisy speeches, excluding -5dB data. Consequently, the training data set had 16 kinds of speech data. A test data set included all the speech data. Both data sets are summarized in Table I.

Visual feature extraction is depicted in Figure 4 and its conditions are shown in Table II. At first, we extracted five visual features (PCA, DCT, LDA, GIF and COORD) respectively. Viseme-based transcriptions were prepared for LDA, GIF and COORD. We adopted 13 visemes (a, i, u, e, o, p, r, sy, t, s, y, vf, and sil) that appear in Japanese digit pronunciation [28], [29], thus we set C =13. Two concatenated feature vectors PDLG and PDLGC were then

TABLE II
EXPERIMENTAL SETUP OF VISUAL FEATURE EXTRACTION.

Feature	Dimension		Remarks
	Static	$+\Delta, \Delta\Delta$	
PCA	13	39	image=35×25, c.c.r.=90%
DCT	15	45	image=50×50
LDA	10	30	image=35×25, c.c.r.>99%
GIF	13	39	image=35×25
COORD	13	39	
PDLG	(51)	153	
PDLGC	(64)	192	
DBVF-PDLGC	40	120	DNN config is in Table III.

c.c.r.=Cumulative Contribution Ratio.

TABLE III
EXPERIMENTAL SETUP OF DNN.

# of units	Input	Hidden	Bottleneck	Output
DBAF	429	2,048	40	179
DBVF-PDLGC	2,112	2,048	40	179

	Pre-training	Fine-tuning
# of epochs	10	50
Minibatch size	256	256
Learning ratio	0.004	0.006
Momentum	0.9	0.0

obtained. Finally a visual DNN was built to compute DBVFs. In order to evaluate DBVFs from PDLGC proposed in this paper, conventional DBVFs from PCA in our previous work [9] were also prepared. To distinguish both DBVFs, we call the former DBVF (proposed one) **DBVF-PDLGC**, and the latter DBVF (conventional one) **DBVF-PCA**. Note that when computing DBVF-PCA, pictures in CENSREC-1-AV were used. All the visual features had first- and second-order time derivatives (Δ and $\Delta\Delta$) in addition to static parameters.

C. Baseline and proposed methods

Model training and recognition were basically the same as CENSREC-1-AV. A left-to-right GMM-HMM was prepared for each word (digit) and silence. A digit HMM consisted of 16 states, while a silence HMM had 3 states. Each state in a digit HMM contained 20 Gaussian components, while there were 36 components on each state in a silence HMM. Because there were 11 digit HMMs (one, two, $\cdots$, nine, zero and oh), the total number of HMM states was 179. The following training and recognition were conducted using HMM Tool Kit (HTK) [30].

For comparison, a baseline audio-only ASR was prepared which is provided in CENSREC-1-AV; GMM-HMMs were trained using 39-dimensional MFCC features. Unimodal speech recognition systems using GMM-HMMs and DBNFs were also used as baseline methods. Table III shows DNN setup. In order to obtain accurate time-aligned transcriptions that were used for visual model training, audio HMMs were trained prior to visual HMMs applying embedded training and using MFCCs. The time-aligned transcription labels were then obtained using the audio HMMs and the training speech data. Next, visual HMMs were built applying bootstrap training and using basic visual features PCA with the labels. After building HMMs, audio and visual DNNs were trained. As input features, MFCC features were used for an audio DNN, while either of PCA or PDLGC features were chosen for a visual DNN. We adopted five previous and five incoming features in addition to a current feature vector, thus we set T=5. An output layer corresponded to audio or visual HMM states, as mentioned, using state-level frame alignments. There were 40 units in a bottleneck layer in all the cases, therefore, 40-dimensional DBAF and DBVF were obtained. Here, audio and visual HMMs were rebuilt using DBAFs and DBVFs, respectively. Finally, multi-stream HMMs for AVSR were generated from the audio and visual HMMs.

Recognition for test data was conducted, performing speaker-independent open-condition evaluation. Stream weight factors in AVSR were empirically optimized in this work.

TABLE IV
DIGIT RECOGNITION ACCURACY USING EVERY VISUAL FEATURES.

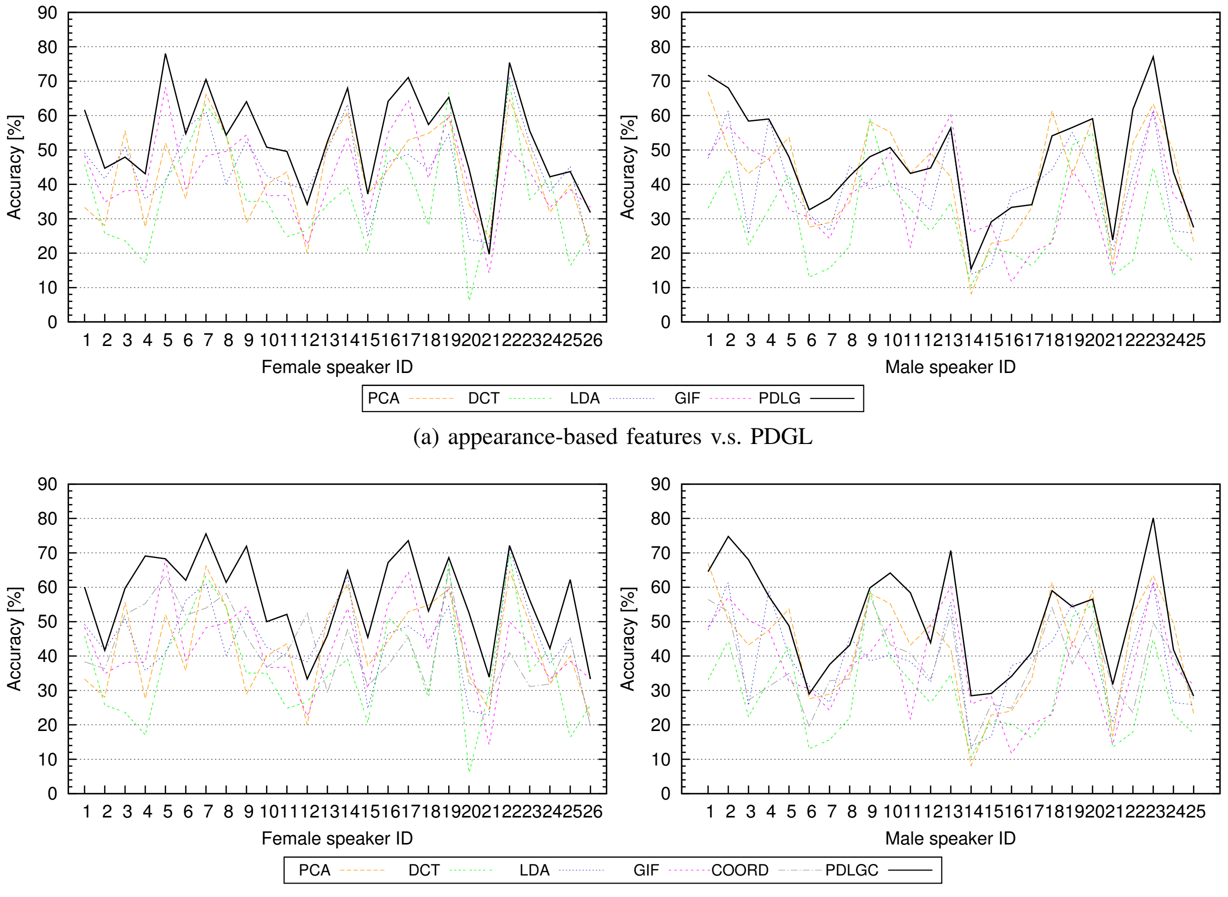
Feature	Insertion penalty	
	w/o	w/
PCA	13.67	42.52
DCT	11.76	33.06
LDA	31.99	41.70
GIF	13.02	39.76
COORD	14.82	39.78
PDLG	38.41	50.05
PDLGC	41.65	53.65
DBVF-PDLGC	69.44	73.66

D. Experiment (1) - Comparison of visual features

At first, we compared and evaluated visual features by carrying out visual speech recognition experiments. Table IV indicates lipreading performance using every visual features. In Table IV, results with and without insertion penalty optimization are indicated. Recognition accuracy in our previous work [9], using DBVF-PCA, was 39.3% without insertion penalty adjustment.

Among appearance-based features with optimizing the penalty factor, PCA achieved the best performance followed by LDA and GIF, but the differences were not so large. Recognition performance of shape-based feature was almost the same as GIF. The combined feature PDLG having the four appearance features could improve recognition accuracy to 50.05%. Furthermore, another combined feature PDLGC including PDLG and COORD achieved much better performance 53.65%. It is also observed when using PDLG or PDLGC, the performances without manual hyper-parameter optimization became better compared to basic visual features. These indicate effectiveness of combining different kinds of visual features. Finally, a DNN-based feature DBVF-PDLGC which was derived from PDLGC could accomplish more than 70% recognition accuracy. Compared to previous researches including our past work, we believe our approach has significantly succeeded.

We analyzed lipreading performance of appearance- and shape-based features as well as their combinations for each person. Figure 5 represents recognition accuracy for every testing subjects (26 females and 25 males); results in Figure 5



(a) appearance-based features v.s. PDGL

(b) appearance-based and shape-based features v.s. PDGLC

Fig. 5. Lipreading accuracy for each speaker in the test set.

TABLE V
AVERAGE DIGIT RECOGNITION ACCURACY FOR AUDIO-ONLY, VISUAL-ONLY AND AUDIO-VISUAL ASR SYSTEMS OVER ALL THE CONDITIONS.

Feature	Accuracy [%]
Audio-only (DBAF)	61.73
Visual-only (DBVF-PDLGC)	66.97
AVSR	89.87

correspond to Table IV. It is observed that which appearance-based or shape-based feature was the best strongly depended on a subject. On the other hand, combined features PDLG and PDLGC were successful in most cases. This indicates using different kinds of visual features simultaneously can deal with speaker differences, causing stable and better recognition performance.

E. Experiment (2) - AVSR using DBNFs

Second, we conducted AVSR experiments using DBAFs and DBVFs derived from PDLGC. Figure 6 represents average recognition accuracy at each SNR level, for audio-only speech recognition using MFCC and DBAF, lipreading using DBVF-PDLGC and audio-visual speech recognition. Table V summarizes average performance over all the 31 conditions for each method. Note that our previous results in [9] were equivalent to 39.7% for ASR using MFCC, 39.3% for lipreading using DBVF-PCA, and 81.1% for AVSR using DBAF and DBVF-PCA. We did not apply insertion penalty optimization, and stream weights were set as λ_a =0.6 and λ_v =0.4.

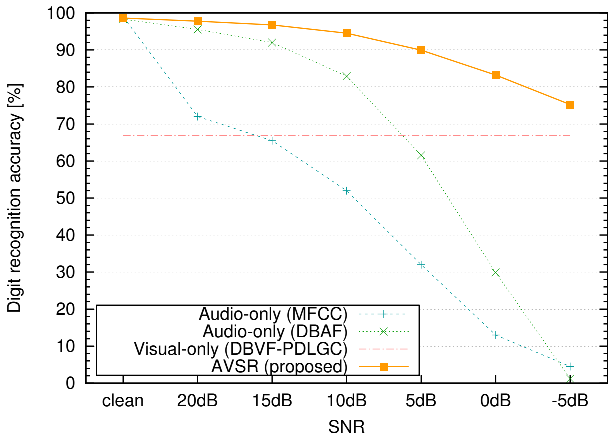


Fig. 6. Average digit recognition accuracy at each SNR level for audio-only, visual-only and audio-visual ASR methods.

The proposed AVSR system using DBAF and DBVF-PDLGC outperformed not only the audio-only baseline but

TABLE VI
DIGIT RECOGNITION ACCURACY USING VISUAL FEATURES ONLY HAVING SPEECH SEGMENTS.

Feature	Insertion penalty	
	w/o	w/
PCA	39.99	53.32
DCT	33.56	45.44
LDA	45.80	50.98
GIF	41.92	52.47
COORD	45.33	52.19
PDLG	55.63	60.91
PDLGC	62.89	64.10
DBVF-PDLGC	76.42	77.80

also the AVSR method using DBAF and DBVF-PCA, achieving 83.2% and 46.4% relative error reduction, respectively. In particular, our proposed method could improve recognition performance in heavily noisy environments keeping the advantage in noiseless conditions, due to increase of visual recognition ability by DBVF-PDLGC. In conclusion, our new AVSR approach significantly improves recognition performance in noisy conditions by employing new visual DNN-based features.

F. Experiment (3) - Importance of VAD in lipreading

As shown in Section IV-D, visual features have been drastically improved by incorporating several kinds of basic features and applying a DNN-based tandem approach. Meanwhile, in conventional ASR, it is much effective to detect and extract speech segments, i.e. to perform VAD, for improving recognition performance and reducing noise influence. On the other hand, it is unclear that VAD is useful for visual speech recognition when using DBVFs. Therefore, we conducted additional experiments for lipreading excluding non-speech segments. A image sequence in CENSREC-1-AV is designed to include 800msec silence periods before and after an utterance. We removed these silence periods from visual features.

Table VI shows experimental results using visual features that only contain speech segments. It is obvious that lipreading performance was drastically improved; 15.7-22.6% relative error reduction was observed compared to the results in Table IV, and the best accuracy was 77.80% when using DBVF-PDLGC. Such the improvement comes mainly from reducing recognition errors within or near silence periods.

We further investigated how removing silence periods affects the performance. Since our recognizer could accept not only digits but beginning and ending silence parts, ideally we can find approximately 800msec beginning and 800msec ending silence segments in recognition results. In other words, if any recognition errors related with silence periods occurred, time duration of the silence periods should vary shorter or longer. Consequently, statistically checking beginning and ending silence duration enables us to find the importance of visual VAD. Figure 7 illustrates histograms of silence duration in recognition results (corresponding to Table IV) for clean audio and some visual features. When using the audio feature, most silence periods had 700-800msec duration properly. On the other hand, when using the visual features without applying DNN, there were a lot of detection failures; many silence periods had shorter duration making insertion errors, and several periods had longer duration causing deletion errors. Compared to these visual features, DBVF-PDLGC had less errors. From these results, if we apply VAD and could correctly detect speech periods, recognition errors in lipreading must incredibly decrease. To conclude, it is important to detect visual speech activities to avoid recognition errors in silence periods, which improves lipreading and AVSR performance.

V. CONCLUSION

This paper proposes two techniques to improve audio-visual speech recognition: (1) enhanced visual features **DBVF-PDLGC** using DNN architectures, and (2) high-performance AVSR using new visual features **DBVF-PDLGC** and **DBAF**. For visual feature extraction, four appearance-based and one shape-based features are extracted from an image. After incorporating them, a tandem approach using DNN is applied to obtain our visual features. For a digit recognition task, experimental results show our visual speech recognition method could achieve 73.66% recognition accuracy in the speaker-independent open condition. Furthermore, in AVSR experiments, we obtained 89.87% average recognition accuracy over clean and noisy conditions. In both cases, we can achieve significant improvement. In addition, we also investigate (3) effectiveness of VAD in the visual modality. Through recognition experiments excluding silence periods from visual features, we finally obtained 77.80% lipreading accuracy. This means VAD is essential not only for audio but also visual modalities. In conclusion, we could obtain better AVSR performance thanks to robust visual features, deep learning techniques, and visual VAD.

With respect to our future works, we would like to further investigate visual features, in particular shape-based ones, to build a better recognition scheme. Although our new DBVF has successfully achieved, there are some speakers whose performance was quite low (roughly 30-40%). To overcome this issue, we have a plan to introduce model adaptation to lipreading [29]. Incorporating our AVSR scheme with audio-visual VAD [31], [32] is also included in our future works.

VI. ACKNOWLEDGMENTS

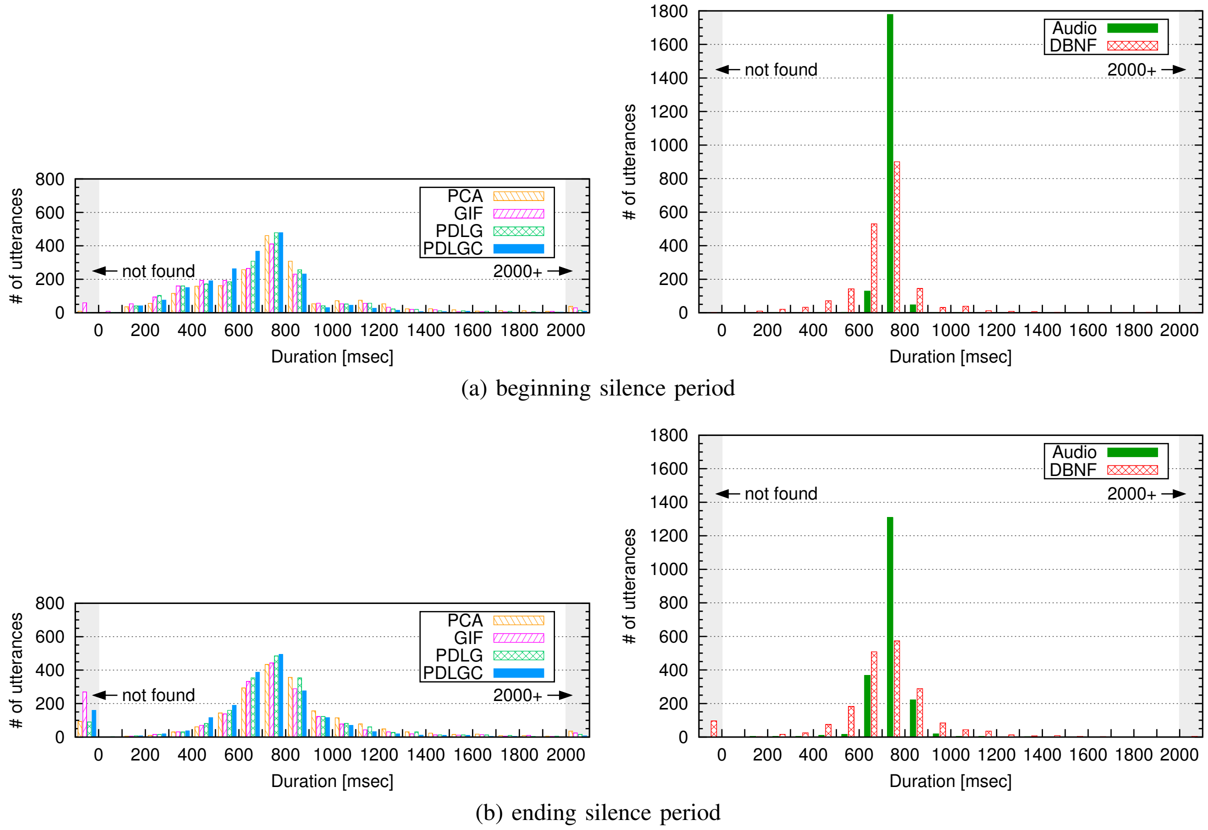
We greatly appreciate Mr. Shinichi KOJIMA (Toyota Central R&D Labs., Inc.), for his support in mouth detection and lip feature point estimation. A part of this work was supported by JSPS KAKENHI Grant Number 25730109.

REFERENCES

[1] A. Mohamed et al., “Acoustic modeling using deep belief networks,” IEEE trans. on Audio, Speech, and Language Processing, vol.20, no.1, pp.23-29 (2012).

[2] D. Yu et al., “Improved bottleneck features using pretrained deep neural networks,” Proc. INTERSPEECH2011, pp.237-240 (2011).

[3] F. Seide et al., “Conversational speech transcription using context-dependent deep neural networks,” Proc. INTERSPEECH2011, pp.437-440 (2011).



(a) beginning silence period

(b) ending silence period

Fig. 7. Histograms of beginning and ending silence duration detected using every features.

[4] J. Gehring et al., "Extracting deep bottleneck features using stacked auto-encoders," Proc. ICASSP2013, pp. 3377-3381 (2013).

[5] T. Hayashi et al., "Investigation of robustness of deep bottleneck features for speakers of a variety of ages in speech recognition," Proc. Forum Acusricum 2014 (2014).

[6] J. Ngiam et al., "Multimodal deep learning," Proc. ICML2011 (2011).

[7] J. Huang et al., "Audio-visual deep learning for noise robust speech recognition," Proc. ICASSP2013, pp.7596-7599 (2013).

[8] K. Noda et al., "Audio-visual speech recognition using deep learning," Applied Intelligence, Springer, vol.42, no.4, pp.722-737 (2015).

[9] H. Ninomiya et al., "Integration of deep bottleneck features for audio-visual speech recognition," Proc. INTERSPEECH2015 (2015, accepted).

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

[11] S. Nakamura et al., "Data collection and evaluation of AURORA-2 Japanese corpus," Proc. ASRU2003, pp.619-623 (2003).

[12] C. Bregler et al., ""Eigenlips" for robust speech recognition," Proc. ICASSP'94, pp.669-672 (1994).

[13] G. Potamianos et al., "Stream confidence estimation for audio-visual speech recognition," Proc. ICSLP2000, vol.3, pp.746-749 (2000).

[14] K. Mase et al., "Automatic lipreading by optical-flow analysis," Systems and Computers in Japan, vol.22, no.6, pp.67-75 (1991).

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

[16] C. Miyamoto et al., "Multimodal speech recognition of a person with articulation disorders using AAM and MAF," Proc. MMSP2010, pp.517-520 (2010).

[17] T. Saitoh, "Efficient face model for lip reading," Proc. AVSP2013, pp.227-232 (2013).

[18] A.H. Abdelaziz, et al., "A new EM estimationof dynamic stream weights for coupled-HMM-based audio-visual ASR," Proc. ICASSP2014, pp.1527-1531 (2014).

[19] V. Estellers et al., "On dynamic stream weighting for audio-visual speech recognition," IEEE Transaction on Audio, Speech, and Language Processing, vol.20, no.4, pp.1145-1157 (2011).

[20] Y. Bengio et al., "Greedy layer-wise training of deep networks," Proc. NIPS'06, pp.153-160 (2007)

[21] G. Hinton et al., "A fast learning algorithm for deep belief nets," Neural Computation, vol.18, no.7, pp.1527-1554 (2006).

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

[23] N. Ukai et al., "GIF-LR: GA-based informative feature for lipreading," Proc. APSIPA ASC 2012 (2012).

[24] S. Kojima, "Statistical face shape model separating inter-individual variation from intra-individual variation," IEICE technical report (IBISML), vol.113, no.197, pp.13-18 (2013, in Japanese).

[25] D. Cristinacce et al., "Feature detection and tracking with constrained local models," Proc. British Machine Vision Conference, vol.3, pp.929-938 (2006).

[26] Y. Wang et al., "Enforcing convexity for improved alignment with constrained local models," Proc. CVPR2008 (2008).

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

[28] 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).

[29] T. Seko et al., "Improvement of lipreading performance using discriminative feature and speaker adaptation," Proc. AVSP2013, pp.221-226 (2013).

[30] `http://htk.eng.cam.ac.uk/`

[31] S. Takeuchi et al., "Voice activity detection based on fusion of audio and visual information," Proc. AVSP2009, pp.151-154 (2009).

[32] C. Ishi et al., "Real-time audio-visual voice activity detection for speech recognition in noisy environments," Proc. AVSP2010, pp.81-84 (2010).