IEICE TRANS. INF. & SYST., VOL.E99–D, NO.10 OCTOBER 2016

PAPER Special Section on Recent Advances in Machine Learning for Spoken Language Processing

Investigation of DNN-Based Audio-Visual Speech Recognition

Satoshi TAMURA[†a)], ***Member***, **Hiroshi NINOMIYA**[††], ***Student Member***, **Norihide KITAOKA**[†††], ***Member***, **Shin OSUGA**[††††], ***Nonmember***, **Yurie IRIBE**[†††††], **Kazuya TAKEDA**[††], ***and*** **Satoru HAYAMIZU**[†], ***Members***

SUMMARY Audio-Visual Speech Recognition (AVSR) is one of techniques to enhance robustness of speech recognizer in noisy or real environments. On the other hand, Deep Neural Networks (DNNs) have recently attracted a lot of attentions of researchers in the speech recognition field, because we can drastically improve recognition performance by using DNNs. There are two ways to employ DNN techniques for speech recognition: a hybrid approach and a tandem approach; in the hybrid approach an emission probability on each Hidden Markov Model (HMM) state is computed using a DNN, while in the tandem approach a DNN is composed into a feature extraction scheme. In this paper, we investigate and compare several DNN-based AVSR methods to mainly clarify how we should incorporate audio and visual modalities using DNNs. We carried out recognition experiments using a corpus CENSREC-1-AV, and we discuss the results to find out the best DNN-based AVSR modeling. Then it turns out that a tandem-based method using audio Deep Bottle-Neck Features (DBNFs) and visual ones with multi-stream HMMs is the most suitable, followed by a hybrid approach and another tandem scheme using audio-visual DBNFs.

key words: *audio-visual speech recognition, deep neural network, Deep Bottleneck Feature, multi-stream HMM*

1. Introduction

Automatic Speech Recognition (ASR) has been developed for many years, and nowadays ASR is widely used on many devices such as cell phones and car navigation systems. However, ASR has been still suffering from degradation of recognition performance in noisy or real environments. To overcome this issue, many techniques have been proposed and used, achieving successful improvement; beam forming is a signal processing technique to extract a target signal using microphone arrays, which enables us to reduce background noises and obtain emphasized speech signals [1]; in acoustic feature extraction, Spectral Subtraction (SS) [2] and Cepstral Mean Normalization (CMN) [3] are often employed to remove noise influence and channel distortion; adjusting recognition models according to test data is also effective in ASR: e.g. Maximum A Posteriori (MAP) [4] and Maximum Likelihood Linear Regression (MLLR) [5].

In addition to these techniques above, Audio-Visual Speech Recognition (AVSR) also known as bimodal speech recognition or multi-modal speech recognition, has been investigated for this couple of decades [6]–[8]. Since lip movement is not basically affected by acoustic noise, visual information can play a great role in the condition where ASR performance severely decreases. There are several research topics in terms of AVSR; making audio-visual databases, e.g. [9]–[11], is essential as a first step of AVSR researches; how to extract effective audio and visual features is also important [12]–[16]; several works were devoted to investigate audio-visual integration, efficient recognition modeling and adaptation, for example [17]; many ASR systems use Voice Activity Detection (VAD) which extracts speech turns from acoustic signals, therefore, audio-visual VAD has also been explored as well [18], [19]; in addition, we should develop real-time AVSR systems and applications on mobile devices like [20].

Recently, deep learning has attracted a lot of attentions in signal processing and pattern recognition domains, including computer vision and speech recognition fields. In particular, a Deep Neural Network (DNN) is often employed in these pattern recognition fields. There are two DNN-based strategies for ASR: a hybrid approach [21] and a tandem approach [22]. In the hybrid approach, an emission probability on each Hidden Markov Model (HMM) state is computed using a DNN. In the tandem approach, a DNN is composed into a feature extraction scheme. Many works related to DNNs have been dedicated for audio-only ASR, showing both hybrid and tandem approaches are effective to improve ASR accuracy. As one of the tandem approaches, some of authors also have investigated a feature extraction method using a DNN having a bottleneck layer [23]. In this paper, we call such the feature vector Deep BottleNeck Feature (DBNF) [23].

Several studies using DNNs in AVSR have been already done as introduced in the following section. However, there are few researches investigating DBNFs to improve recognition performance of AVSR not only in clean but also noisy environments. Furthermore, we should compare an AVSR method using DBNFs with an AVSR scheme employing the hybrid strategy. Therefore, we mainly focus on how to incorporate audio and visual modalities in AVSR that uses DNNs. Particularly, in this work we investigate which

Manuscript received February 5, 2016.
Manuscript revised May 25, 2016.
Manuscript publicized July 19, 2016.

[†]The authors are with Gifu University, Gifu-shi, 501–1193 Japan.
[††]The authors are with Nagoya University, Nagoya-shi, 464–8603 Japan.
[†††]The author is with Tokushima University, Tokushima-shi, 770–8506 Japan.
[††††]The author is with Aisin Seiki Co., Ltd., Kariya-shi, 448–8650 Japan.
[†††††]The author is with Aichi Prefectural University, Nagakute-shi, 480–1198 Japan.
a) E-mail: tamura@info.gifu-u.ac.jp
DOI: 10.1587/transinf.2016SLP0019

method is suitable for DBNFs, and when integrating audio and visual modalities either before or after applying DNNs. Moreover, we also clarify which scheme is better for AVSR, a hybrid approach or a tandem approach with multi-stream HMMs. We built several AVSR methods, some of which were based on the tandem approach and the other was based on the hybrid approach. We then tested these methods using an audio-visual corpus CENSREC-1-AV [9], to find out which method is suitable for AVSR.

The rest of this paper is organized as follows. Section 2 briefly describes related works. DNN-based modeling and feature extraction investigated in this paper are introduced in Sect. 3. Section 4 shows experimental setup, result and discussion. Finally Sect. 5 concludes this paper.

2. Related Works

First, related works using DNNs in ASR are briefly introduced. As mentioned, in general there are two approaches: a hybrid method [21] and a tandem strategy [22]. A conventional HMM employs a Gaussian Mixture Model (GMM) on each state, to calculate an emission probability for a given feature vector. This kind of HMM is nowadays called GMM-HMM. In the hybrid approach, emission probabilities are obtained as posteriori probabilities instead, which are based on output scores of DNN. This approach is often called DNN-HMM. On the other hand, in the tandem approach, a DNN is exploited to extract feature vectors. Using this strategy, we can build an extracted feature set according to training data. There are two ways to obtain feature vectors in the tandem approach; output values of DNN are straightforwardly chosen; alternatively, a DNN is designed to include a certain hidden layer having relatively few perceptrons, often called a bottleneck layer, and output values of the layer are composed as a feature vector. This paper focuses on the second scheme, e.g. [23].

There are many researches related to AVSR. As mentioned in Sect. 1, several researchers have tried to develop DNN-aided AVSR schemes; a bimodal deep audoencoder was proposed to obtain multi-modal feature vectors [24]; a deep belief network was also utilized performing middle-level feature combination [25]; another research used a DNN to check audio-visual synchrony when combining audio and visual features [26]; in terms of recognition models, multi-stream HMMs, that are often employed in AVSR, were built using features obtained by deep denoising autoencoder [27].

We aim at investigating DNN-based AVSR mainly from the viewpoint of audio and visual integration. This paper firstly focuses on audio-visual feature extraction related to the previous works [24]–[26]. We have proposed an AVSR scheme to use DBNFs, achieving significant improvement [28], [29]. In general, there are two architectures to extract audio-visual features; audio features are computed using DNNs from basic audio ones, simultaneously visual features are obtained as well, before both features are combined; alternatively, basic audio and visual features are concatenated followed by applying DNNs to calculate audio-visual features. However, according to our knowledge, there is no work to compare these approaches. Secondly, we compare a hybrid-based scheme and tandem-based AVSR methods that particularly using multi-stream HMMs. Both methods respectively have advantages and disadvantages for ASR; the hybrid approach can usually achieve higher performance, however, it is difficult to apply conventional model adaptation that is much effective in real environments; in the tandem approach, we can easily employ conventional techniques including multi-stream HMMs for AVSR. Because there is no prior knowledge about these approaches for AVSR, this paper also tries to find out which strategy is better.

3. DNN-Based AVSR

We investigate several DNN-based modeling and feature extraction schemes for AVSR. In this section, these methods proposed in this paper are respectively introduced in addition to baseline systems without DNNs. Figure 1 summarizes all the methods.

3.1 Baseline(l): GMM-HMM Using MFCC and PCA

A baseline system having conventional modeling and feature extraction schemes is prepared. A conventional 39-dimensional Mel-Frequency Cepstral Coefficient (MFCC) vector is employed as an audio feature $\boldsymbol{f}_{a,t}$ where t is a frame index. A visual feature vector consists of 10-dimensional eigenlip parameters [12] as well as their Δ and $\Delta\Delta$ coefficients. Eigenlip feature extraction is based on Principal Component Analysis (PCA). Let us denote a raster-scan vector from a lip image by $\boldsymbol{r}_t = (v_{x,y,t})$ having intensity values $v_{x,y,t}$ of every pixels (x, y) in a t-th image. A covariance matrix of training feature vectors is decomposed to orthogonal vectors (eigenvectors) with corresponding variances (eigenvalues). A transformation matrix A is then obtained by choosing eigenvectors that have larger eigenvalues over a certain threshold. Now we can compute a static visual feature vector $\boldsymbol{s}_t$ from an input feature in Eq. (1):

$$\boldsymbol{s}_t = A \cdot \boldsymbol{r}_t \tag{1}$$

followed by computing time derivatives of $\boldsymbol{s}_t$ to generate a visual feature vector $\boldsymbol{f}_{v,t}$. An audio-visual feature vector $\boldsymbol{f}_{av,t}$ is finally generated by concatenating audio and visual feature vectors as Eq. (2):

$$\boldsymbol{f}_{av,t} = \left(\boldsymbol{f}_{a,t}{}^{\top}, \boldsymbol{f}_{v,t}{}^{\top}\right)^{\top} \tag{2}$$

where $\top$ indicates transpose. Note that before the above process, audio and visual frame rates must be consistent. In most cases, a visual frame rate is lower then an audio one. Thus in our work, visual feature vectors are interpolated using a spline function so that both feature vectors could be synchronized. This synchronization is based on the original baseline scheme in CENSREC-1-AV [9].

		Baseline(l)	Baseline(m)	Hybrid	Tandem(i40)	Tandem(i80)	Tandem(l)	Tandem(m)
Feature	Audio	MFCC(39)	MFCC(39)	MFCC(39)	DBAVF(40)	DBAVF(80)	DBAF(40)	DBAF(40)
	Visual	PCA(30)	PCA(30)	PCA(30)			DBVF(40)	DBVF(40)
Model		GMM-HMM	GMM-HMM (multi-stream)	DNN-HMM	GMM-HMM	GMM-HMM	GMM-HMM	GMM-HMM (multi-stream)

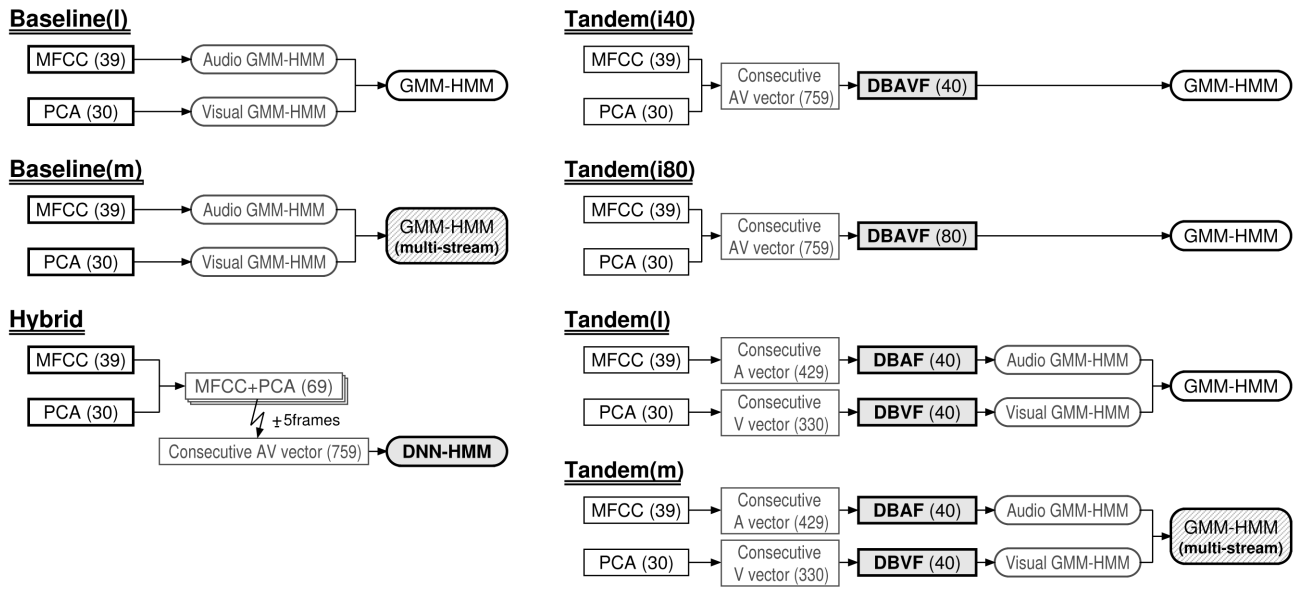


Fig. 1 AVSR methods using DNNs (numbers in brackets indicate feature dimensions).

Model training is simply carried out as follows. At first, audio GMM-HMMs are built using embedded training with the Maximum Likelihood (ML) criterion. Next, time-aligned labels are obtained using the audio HMMs and clean audio training data. Visual GMM-HMMs are subsequently estimated using ML-based bootstrap training with the time-aligned labels. Finally, both HMMs are simply combined into audio-visual HMMs so that audio and visual modalities have equal influence. This training method is also based on CENSREC-1-AV. When recognizing test data, the Viterbi algorithm is applied for audio-visual features.

3.2 Baseline(m): Multi-Stream GMM-HMM Using MFCC and PCA

Another baseline system is also considered in this work. The same feature extraction method is employed as the last baseline scheme, while a recognition model is modified. In AVSR, a multi-stream HMM is often employed which can adjust contributions of audio and visual modalities. Let us denote log likelihoods obtained from audio and visual models by $b_a(\boldsymbol{f}_{a,t})$ and $b_v(\boldsymbol{f}_{v,t})$, respectively. A log likelihood $b_{av}(\boldsymbol{f}_{av,t})$ for an audio-visual feature vector $\boldsymbol{f}_{av,t}$ in a multi-stream HMM is then formulated as Eq. (3):

$$b_{av}\left(\boldsymbol{f}_{av,t}\right) = \lambda_a b_a(\boldsymbol{f}_{a,t}) + \lambda_v b_v(\boldsymbol{f}_{v,t}) \tag{3}$$

where λ_a and λ_v are stream weight factors for audio and visual streams, respectively. By controlling these factors properly, we can improve recognition performance of AVSR.

Note that model training is almost the same as Baseline(l), except introducing multi-stream HMMs in the last step. Therefore, stream weights do not affect model training, and are only in effect when recognizing test data.

3.3 Hybrid: DNN-HMM Using Conventional Features

Instead of GMM-HMM, this hybrid method employs a DNN-HMM. At a t-th frame, a concatenated vector $\boldsymbol{c}_{av,t}$ is prepared as Eq. (4), by combining consecutive audio-visual features:

$$\boldsymbol{c}_{av,t} = \left(\boldsymbol{f}_{av,t-T}{}^{\top}, \boldsymbol{f}_{av,t-T+1}{}^{\top}, \cdots, \boldsymbol{f}_{av,t+T}{}^{\top}\right)^{\top} \tag{4}$$

The concatenated vector corresponds to an input layer of DNN. Each perceptron on an output layer generates an emission probability for an HMM state. As mentioned later, this paper employs 11 HMMs each having 16 states and one HMM including three states. Thus the number of states is 179 in total, meaning the output layer has 179 units. Figure 2 depicts a DNN structure for the hybrid approach. DNN training consists of two stages: pre-training and fine-tuning; unsupervised pre-training is conducted in a layer-wise manner [30], before all parameters are fine-tuned [31].

3.4 Tandem(i): GMM-HMM Using DBAVF

In this method, a conventional GMM-HMM is adopted while feature extraction is based on deep learning. From an audio-visual feature vector $\boldsymbol{c}_{av,t}$ consisting of MFCCs and

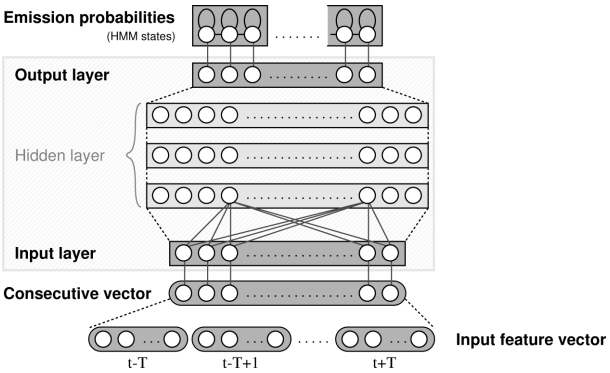


Fig. 2 A DNN for a hybrid approach.

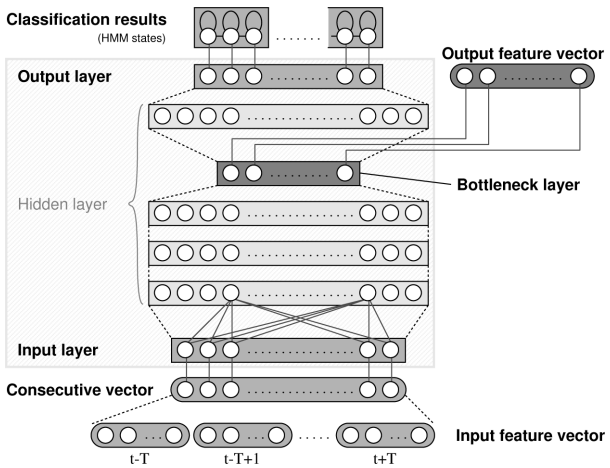


Fig. 3 A DNN for a tandem approach.

eigenlip parameters, a DBNF vector $\boldsymbol{d}_{av(i),t}$ is obtained using a DNN. We call this DBNF vector derived from an *initially* combined feature Deep Bottleneck Audio-Visual Feature (**DBAVF**). Similar to the hybrid approach, a DNN is built where an input layer corresponds to $\boldsymbol{c}_{av,t}$ and an output layer corresponds to HMM states. The difference from the hybrid approach is that, the DNN has a bottleneck layer. Using the bottleneck layer, we can obtain DBAVF. Figure 3 illustrates a structure of DNN. Once DBAVFs are obtained, model training and recognition can be conducted in the same manner as the baseline schemes; audio-visual HMMs are built applying ML-based embedded training using DBAVFs in a training data set. In this work, two methods (i40) and (i80) are prepared to investigate influence of dimensionality.

3.5 Tandem(l): GMM-HMM Using DBAF and DBVF

This scheme also chooses GMM-HMMs and DBNFs. In contrast to the last method, a DBNF vector is computed in each modality before both DBNFs are fused in the *late* stage. To obtain an audio DBNF $\boldsymbol{d}_{a,t}$, consecutive MFCC vectors are incorporated into one vector $\boldsymbol{c}_{a,t}$ as:

$$\boldsymbol{c}_{a,t} = \left(\boldsymbol{f}_{a,t-T}{}^{\top}, \boldsymbol{f}_{a,t-T+1}{}^{\top}, \cdots, \boldsymbol{f}_{a,t+T}{}^{\top}\right)^{\top} \quad (5)$$

Fig. 4 An example of lip image.

before the vector is converted to a DBNF using an audio DNN. A visual DBNF $\boldsymbol{d}_{v,t}$ is computed from a concatenated visual vector $\boldsymbol{c}_{v,t}$ as well. We call the audio feature Deep Bottleneck Audio Feature (**DBAF**), and the visual feature Deep Bottleneck Visual Feature (**DBVF**). An audio-visual feature vector $\boldsymbol{d}_{av(l),t}$ is then composed as:

$$\boldsymbol{d}_{av(l),t} = \left(\boldsymbol{d}_{a,t}{}^{\top}, \boldsymbol{d}_{v,t}{}^{\top}\right)^{\top} \quad (6)$$

Note that DNN training, model training and recognition are the same as the previous methods.

3.6 Tandem(m): Multi-Stream GMM-HMM Using DBAF and DBVF

We also prepare a tandem-based approach employing multi-stream HMMs. Using log likelihoods in audio and visual modalities $b_a(\boldsymbol{d}_{a,t})$ and $b_v(\boldsymbol{d}_{v,t})$, an audio-visual log likelihood $b_{av}(\boldsymbol{d}_{av(l),t})$ for an audio-visual feature vector $\boldsymbol{d}_{av(l),t}$ can be obtained as Eq. (7):

$$b_{av}\left(\boldsymbol{d}_{av(l),t}\right) = \lambda_a b_a(\boldsymbol{d}_{a,t}) + \lambda_v b_v(\boldsymbol{d}_{v,t}) \quad (7)$$

4. Experiment

In order to compare AVSR methods in Sect. 3, we conducted recognition experiments using an audio-visual corpus. In this section, database and experimental setup are introduced, followed by experimental result and discussion.

4.1 Database

In this paper, a Japanese audio-visual corpus CENSREC-1-AV was used [9]. CENSREC-1-AV is designed to evaluate AVSR, providing training and test data. Each utterance in this database consists of 1-7 connected digit(s). The training data set includes 3,234 utterances spoken by 20 female and 22 male subjects. There are 1,963 utterances in the test set, made by 26 female and 25 male speakers. All the speech data were recorded at the sampling rate of 16kHz. Assuming that we would apply AVSR to in-car environments where illumination condition is drastically changed, infrared gray-scale mouth images were chosen, of which size is 81×55. A sample image is shown in Fig. 4. In order to obtain eigenlip features, all mouth images were resized into 40×27 gray-scale ones. Note that the cumulative contribution ratio for 10-dimensional eigenlip features was 76%.

Table 1 Specification of training and test data sets.

(a) Summary

	Training set	Test set
# of speakers	42	51
# of utterances	3,234	1,963
Audio	clean speech, 15 noisy speeches[(b)]	clean speech, 30 noisy speeches[(c)]
Visual	infrared	infrared

(b) Noise condition for training data set

SNR	20dB	15dB	10dB	5dB	0dB	−5dB
Cityroad	x	x	x	x	x	
Expressway	x	x	x	x	x	
Music	x	x	x	x	x	
Music+Cityroad						
Music+Expressway						

(c) Noise condition for test data set

SNR	20dB	15dB	10dB	5dB	0dB	−5dB
Cityroad	x	x	x	x	x	x
Expressway	x	x	x	x	x	x
Music	x	x	x	x	x	x
Music+Cityroad	x	x	x	x	x	x
Music+Expressway	x	x	x	x	x	x

To train DNNs and recognition models, and to evaluate AVSR in different noise conditions, not only clean data but also noisy data were prepared. Interior car noises recorded on cityroads and expressways provided in CENSREC-1-AV (**Cityroad** and **Expressway**), as well as musical waveforms (**Music**) were respectively added to clean speech data at several SNR levels (20dB, 15dB, 10dB, 5dB, 0dB and −5dB). In addition, two kinds of noisy speech data were also prepared; cityroad noise and musical sound were simultaneously added to speech data (**Music+Cityroad**), and similarly, expressway noise and musical sound were overlapped to speech waveforms (**Music+Expressway**). As a result, clean speech data and 30 kinds of noisy speech data were prepared. All the kinds of speech data were included in the test set, while the training data set consisted of clean speech data as well as cityroad-, expressway-, and music-overlapped (**Cityroad**, **Expressway**, and **Music**) noisy speeches, excluding −5dB data. Both data sets are summarized in Table 1.

4.2 Experimental Setup

Feature extraction setup for MFCC and eigenlip was the same as those used in CENSREC-1-AV; in this work, we did not apply noise reduction techniques such as SS and CMN. A current feature vector in addition to previous five and incoming five feature vectors were concatenated ($T = 5$) to make an input feature vector of DNN, when computing DNN-HMM, DBAVF, DBAF and DBVF, respectively. We have investigated some DBAFs, and found 40-dimensional one can achieve good performance. In order to easily discuss and compare DBNFs, we chose the same dimension for DBVFs and DBAVFs, except 80-dimensional ones.

Experimental setup of DNN is shown in Table 2. A sigmoid function was basically chosen as an activation function. A DNN used in the hybrid method had three hidden layers, while DNNs used for the tandem methods had five hidden layers when training. When computing DBNFs, a bottleneck layer is located as a fourth hidden layer, just as Fig. 3. Therefore, the number of layers from the input layer to the output (hybrid) or bottleneck (tandem) layer was the same. We simply adopted the number of hidden layers based on the conventional setting that has been used in most speech recognition. For an audio DNN, we have tested a couple of conditions in our past works and found this condition is the best. Regarding DBAVFs and a hybrid approach, we assumed that the same architecture would be suitable because the number of training data samples was the same. It is actually true that input feature dimensions were different, however, only hidden layers next to input layers was changed and might be strongly affected. Finally, in the visual modality, the amount of training data for a visual DNN was much less which might be insufficient to investigate the optimal condition. Thus in this work, we chose the same unified condition. We consequently believe influence of employing the same architecture among these DNNs is limited in this work, even though we should explore the best DNN setting for each method.

Table 2 Experimental setup for DNNs.

(a) The number of units on each layer.

	Input	Hidden	Bottleneck	Output
DNN-HMM	759	2,048	-	179
DBAVF(80)	759	2,048	80	179
DBAVF(40)	759	2,048	40	179
DBAF(40)	429	2,048	40	179
DBVF(40)	330	2,048	40	179

(b) Pre-training and fine-tuning settings.

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

In terms of modeling, a left-to-right HMM was prepared for each word (digit) and silence. A digit HMM consisted of 16 states, while a silence HMM had 3 states. For GMM-HMMs, each state in the digit HMM contained 20 Gaussian components, while there were 36 components on each state in the silence HMM. Note that there were 11 digit HMMs (one, two, $\cdots$, nine, zero and oh) in the following experiments. Model adaptation such as MAP and MLLR was not applied. In this paper, we set stream weight factors empirically; we tested 11 pairs $(\lambda_a, \lambda_v) = (1.0, 0.0), (0.9, 0.1), \cdots, (0.0, 1.0)$ and chose the weights that achieved the best recognition performance for all test data. Optimizing stream weights according to test condition is of course important, however, is also one of crucial challenges in AVSR. Because we would like to separate the issues, in this paper we kept stream weights for all test conditions.

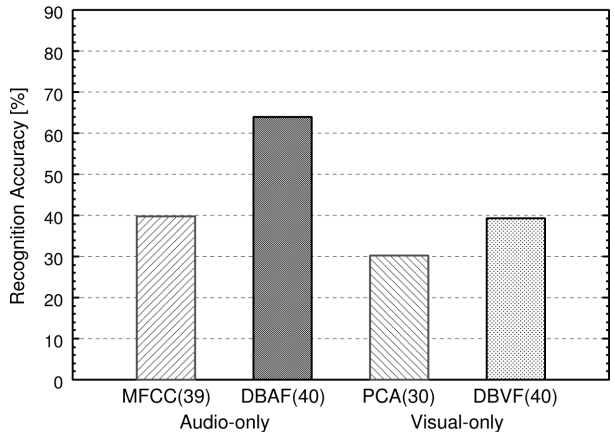


Fig. 5 Digit recognition accuracy for unimodal methods.

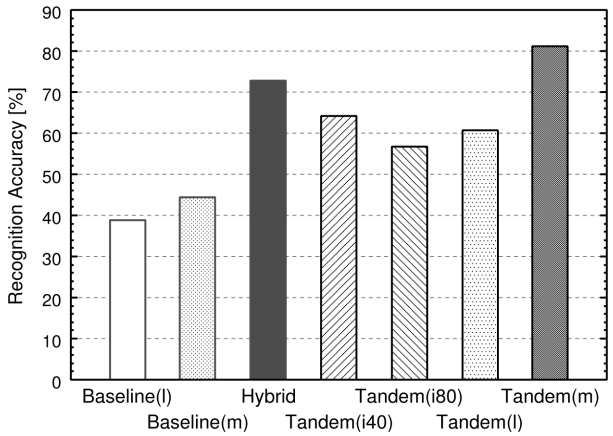


Fig. 6 Digit recognition accuracy for AVSR methods.

4.3 Experimental Result and Discussion

We evaluated each recognition method by digit recognition accuracy for all the test data. Figure 5 shows recognition results of audio-only and visual-only recognition schemes. From Fig. 5, it is obvious that using DBNFs (DBAF and DBVF) can improve recognition accuracy. Particularly in the audio modality, approximately 40% relative error reduction was observed, showing effectiveness of DBNF in speech recognition.

Next, we compared audio-only, visual-only, and AVSR methods. Figure 6 indicates recognition results of AVSR approaches. It is interestingly found that some bimodal results were slightly degraded from the unimodal results, in particular from the audio ones: for example, MFCC (39.7%) vs Baseline(l) (38.8%), and, DBAF (63.9%) vs Tandem(l) (60.7%). Since the visual performance was lower than the audio one, the visual modality could not contribute to performance improvement as long as simply concatenating both features and using conventional GMM-HMMs. In other words, in order to improve recognition performance by adopting AVSR, a certain framework to effectively incorporate audio and visual modalities, or to balance audio and visual contributions in a recognition model is essential.

Here, we investigated audio-visual integration frameworks. Firstly, two DBAVFs (i40) and (i80) were compared. It is observed that the former approach (i40) is superior to the latter one (i80). This means that we can efficiently compact audio-visual information by using the feature (i40). Secondly, we discussed tandem(i40) and tandem(l) schemes. Compared with the tandem(i40) method, the tandem(l) approach was a little worse. When training a feature extraction scheme for the tandem(i40) method, audio and visual information were simultaneously and complementarily used. This means audio information affected the visual modality and vice versa. When building a GMM-HMM in the tandem(l) method, however, audio information was only used for audio model parameters, and visual data also affected visual model parameters only. This might cause such the difference. Thirdly, we compared the tandem(i40) method that was the best method so far, with the hybrid scheme. As a result, recognition performance of the hybrid method was higher. The difference between both methods is how to use DNN outputs; in the hybrid method DNN outputs were used as emission probabilities directly, whereas in the tandem(i40) approach outputs on a bottleneck layer were composed as a feature vector for GMM-HMMs. This indicates using DNN-HMMs can strongly enhance the performance of AVSR.

It is finally observed that the tandem(m) method can significantly improve recognition performance compared to the other AVSR approaches including the hybrid approach. It is difficult to employ audio-visual balancing frameworks like multi-stream HMMs in the hybrid and tandem(i) schemes. This means that it is the best for AVSR to use DNNs for feature extraction and to employ a multi-stream framework. In terms of stream weights, the best recognition performance in the tandem(m) approach was obtained when using $\lambda_a = 0.6 \sim 0.7$ and $\lambda_v = 0.3 \sim 0.4$. We also found visual information can contribute even in the clean condition. These facts indicate that the visual stream always improves the performance. In terms of noises, we found that the best stream weight pair depends more strongly on SNRs than noise type.

To briefly conclude, incorporating audio and visual modalities after applying DNNs and adopting multi-stream HMMs to balance both contributions are quite effective in AVSR. However, combing both modalities before applying DNNs is also useful to some extent. If we have numerous training data, such the balancing architecture can be also included in DNNs, and thus there is still a potential to employ a hybrid method in AVSR.

5. Conclusion

This paper investigated several kinds of AVSR methods using DNNs, mainly focusing on how to incorporate audio and visual modalities in AVSR that uses DNNs. We prepared two baseline methods using conventional features.

A hybrid-based AVSR scheme as well as four tandem-based AVSR approaches exploiting DBNFs were also built. Evaluation experiments were conducted using CENSREC-1-AV. Recognition results tell us integrating audio and visual modalities after obtaining DBNFs from DNNs with a balancing framework i.e. multi-stream HMMs is the most suitable. A hybrid approach and a tandem method using DBAVFs may be also useful if we can get more audio-visual data.

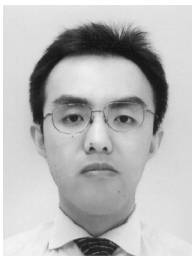
Finally as our future work, further investigating DNN-based AVSR frameworks using a large-scale audio-visual database will be held to verify the potential of the hybrid and DBAVF methods and to determine the optimal condition of DNNs. We will also apply our scheme to the other tasks such as Large Vocabulary Continuous Speech Recognition (LVCSR). In addition, involving better audio and visual features for DNNs like [29] is also expected.

Acknowledgments

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

References

[1] B.D. Van Veen and K.M. Buckley, "Beamforming: A versatile approach to spatial filtering," IEEE Acoust. Speech Signal Process. Mag., vol.5, no.2, pp.4–24, 1988.
[2] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Signal Process., vol.27, no.2, pp.113–120, 1979.
[3] B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Am., vol.55, no.6, pp.1304–1312, 1974.
[4] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Trans. Speech Audio Process. vol.2, no.2, pp.291–298, 1994.
[5] C.J. Leggette and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, vol.9, no.2, pp.171–185, 1995.
[6] G. Potamianos and C. Neti, "Stream confidence estimation for audio-visual speech recognition," Proc. ICSLP2000, vol.3, pp.746–749, 2000.
[7] C. Miyajima, K. Tokuda, and T. Kitamura, "Audio-visual speech recognition using MCE-based HMMs and model-dependent stream weights," Proc. ICSLP2000, vol.2, pp.1023–1026, 2000.
[8] K. Iwano, S. Tamura, and S. Furui, "Bimodal speech recognition using lip movement measured by optical-flow analysis," Proc. HSC2001, pp.187–190, 2001.
[9] S. Tamura, C. Miyajima, N. Kitaoka, T. Yamada, S. Tsuge, T. Takiguchi, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, S. Matsuda, T. Ogawa, S. Kuroiwa, K. Takeda, and S. Nakamura, "CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition," Proc. AVSP2010, pp.85–88, 2010.
[10] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," IEEE Trans. Multimed., vol.17, no.5, pp.603–615, 2015.
[11] D. Burnham, "Big data and resource sharing: A speech corpus and a virtual laboratory for facilitating human communication science research," Proc. Oriental COCOSDA 2014 (Keynote talk), p.10, 2014.
[12] C. Bregler and Y. Konig, ""Eigenlips" for robust speech recognition," Proc. ICASSP '94, vol.2, pp.669–672, 1994.
[13] C. Miyamoto Y. Komai, T. Takiguchi, Y. Ariki, and I. Li, "Multimodal speech recognition of a person with articulation disorders using AAM and MAF," Proc. MMSP2010, pp.517–520, 2010.
[14] Y. Lan, R. Harvey, B.-J. Theobald, E.-J. Ong, and R. Bowden, "Comparing visual features for lipreading," Proc. AVSP2009, pp.102–106, 2009.
[15] S. Tamura, Y. Tagami, and S. Hayamizu., "GIF-SP: GA-based informative feature for noisy speech recognition," Proc. APSIPA ASC 2012, PS.5-SLA.18.10, pp.1–4, 2012.
[16] N. Ukai, T. Seko, S. Tamura, and S. Hayamizu, "GIF-LR: GA-based informative feature for lipreading," Proc. APSIPA ASC 2012, PS.3-IVM.7.5, pp.1–4, 2012.
[17] S. Tamura, M. Oonishi, and S. Hayamizu, "Audio-visual interaction in model adaptation for multi-modal speech recognition," Proc. APSIPA ASC 2011, Thu-PM.PS2.7, pp.1–4, 2011.
[18] S. Takeuchi, T. Hashiba, S. Tamura, and S. Hayamizu, "Voice activity detection based on fusion of audio and visual information," Proc. AVSP2009, pp.151–154, 2009.
[19] C.T. Ishi, M. Sato, N. Hagita, and S. Lao, "Real-time audio-visual voice activity detection for speech recognition in noisy environments," Proc. AVSP2010, pp.81–84, 2010.
[20] S. Tamura, T. Seko, and S. Hayamizu, "Data collection for mobile audio-visual speech recognition in various environments," Proc. Oriental COCOSDA 2014, pp.134–139, 2014.
[21] A.-R. Mohamed, G.E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," IEEE Trans. Audio Speech Language Process., vol.20, no.1, pp.14–21, 2012.
[22] D. Yu and M.L. Selzer, "Improved bottleneck features using pretrained deep neural networks," Proc. INTERSPEECH2011, pp.237–240, 2011.
[23] T. Hayashi, N. Kitaoka, and K. Takeda, "Investigating the robustness of deep bottleneck features for recognizing speech of speakers of various ages," Proc. Forum Acusticum 2014, pp.1–6, 2014.
[24] J. Ngiam, A. Khosla, M. Kim, and A.Y. Ng, "Multimodal deep learning," Proc. ICML2011, 2011.
[25] J. Huang and B. Kingsbury, "Audio-visual deep learning for noise robust speech recognition," Proc. ICASSP2013, pp.7596–7599, 2013.
[26] E. Marcheret, G. Potamianos, J. Vopicka, and V. Goel, "Detecting audio-visual synchrony using deep neural networks," Proc. INTERSPEECH2015, pp.548–552, 2015.
[27] K. Noda, Y. Yamaguchi, K. Nakadai, H.G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," Applied Intelligence, vol.42, no.4, pp.722–737, Springer, 2015.
[28] H. Ninomiya, N. Kitaoka, S. Tamura, Y. Iribe, and K. Takeda, "Integration of deep bottleneck features for audio-visual speech recognition," Proc. INTERSPEECH2015, pp.563–566, 2015.
[29] S. Tamura, H. Ninomiya, N. Kitaoka, S. Osuga, Y. Iribe, K. Takeda, and S. Hayamizu, "Audio-visual speech recognition using deep bottleneck features and high-performance lipreading," Proc. APSIPA ASC 2015, pp.575–582, 2015.
[30] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," Proc. NIPS '06, pp.153–160, 2007
[31] G.E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," Neural Comput., vol.18, no.7, pp.1527–1554, 2006.



Satoshi Tamura received his M.S. and Ph.D. degrees in information science and engineering from Tokyo Institute of Technology, in 2002 and 2005 respectively. He became a research associate at Department of Computer Science, Gifu University, in 2005. He has been an assistant professor in Gifu University since 2007. His research interests are speech information processing, such as multimodal (audio-visual) speech recognition, robust speech recognition and application of speech recognition to real system. He is also interested in music information processing, natural language processing, computer vision, medical information processing, and service engineering.

Hiroshi Ninomiya received his B.S. degree in Electrical, Electronic Engineering and Information Engineering from Nagoya University in 2015. He is currently pursuing his M.S. degree in Graduate School of Information Science at Nagoya University. His research interests include speech recognition and image recognition.

Norihide Kitaoka received his B.S. and M.S. degrees from Kyoto University. In 1994, he joined Denso Corporation. In 2000, he received his Ph.D. degree from Toyohashi University of Technology (TUT). He joined TUT as a research associate in 2001 and was a lecturer from 2003 to 2006. He became an associate professor in Nagoya University in 2006. Since 2015 he has been a professor in Tokushima University.

Shin Osuga became an engineer in Aisin Seiki Co., Ltd., Japan in 2003. He received his M.S. degree in Graduate School of Arts and Sciences from the University of Tokyo in 2009. He has developed a camera based in-cabin monitoring system. His interest is human machine interface.

Yurie Iribe received her B.E. degree in Systems Engineering from Nagoya Institute of Technology and M.S. degree in Human Informatics from Nagoya University in 1999 and 2001, respectively. She became a research associate in the Information and Media Center at Toyohashi University of Technology in 2004. She received her Ph.D. degree from Nagoya University in 2007. She is currently an assistant professor in Aichi Prefectural University from 2013. Her research interests include education support and human interface.

Kazuya Takeda received his B.E., M.E., and Ph.D. degrees from Nagoya University, in 1983, 1985, and 1994, respectively. From 1986 to 1989 he was a researcher at Advanced Telecommunication Research laboratories (ATR). From 1989 to 1995, he was a researcher and research supervisor at KDD Research and Development Laboratories. From 1995 to 2003, he was an associate professor at Faculty of Engineering, and since 2003 he has been a professor at Graduate School of Information Science at Nagoya University. His research interests are media signal processing and its applications include; spatial audio, robust speech recognition, behavior modeling and interfaces.

Satoru Hayamizu received his B.E., M.E., and Ph.D. degrees from the University of Tokyo in 1978, 1981, and 1993, respectively. From 1981 to 2001, he worked at Electrotechnical Laboratory, AIST, Ministry of International Trade and Industry. From 2001 to 2002, he worked at National Institute of Advanced Industrial Science and Technology. Since 2002, he has been a professor at Faculty of Engineering, Gifu University. His research interests are machine learning and media informatics.