Lipreading using deep bottleneck features for optical and depth images

Satoshi Tamura[1], *Koichi Miyazaki*[1,2] *and Satoru Hayamizu*[1]

[1]Faculty of Engineering, Gifu University, Japan
[2]Graduate School of Information Science, Nagoya University, Japan
tamura@info.gifu-u.ac.jp, miyazaki@asr.info.gifu-u.ac.jp, hayamizu@gifu-u.ac.jp

Abstract

This paper investigates a lipreading scheme employing optical and depth modalities, with using deep bottleneck features. Optical and depth data are captured by Microsoft Kinect v2, followed by computing an appearance-based feature set in each modality. A basic feature set is then converted into a deep bottleneck feature using a deep neural network having a bottleneck layer. Multi-stream hidden Marcov models are used for recognition. We evaluated the method using our connected-digit corpus, comparing to our previous method. It is finally found that we could improve lipreading performance by employing deep bottleneck features.

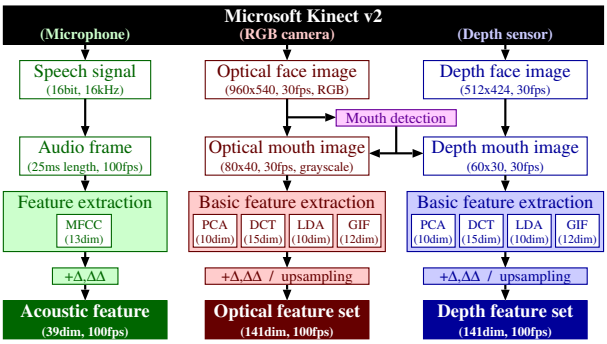
Index Terms: lipreading, deep bottleneck feature, depth information, multi-stream HMM.

1. Introduction

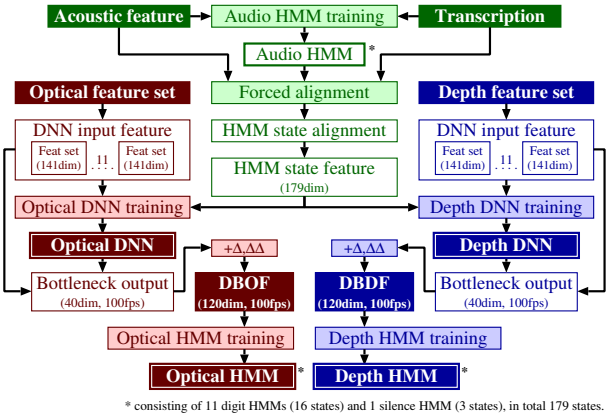
Lipreading is a technique to recognize visual-only speech activities [1, 2, 3]. In conventional methods, at first some preprocessing methods such as face detection were conducted followed by extracting visual features from an image sequence. Lipreading was then performed using some image and speech recognition technology, for instance using a Hidden Markov Model (HMM) that has been employed in many speech recognizers. Recently, lipreading approaches adopting Deep Neural Network (DNN) techniques have attracted many researchers. For example, an end-to-end lipreading method employing long short-term memory architectures was proposed, achieving high accuracy [3].

We have already investigated a lipreading method using optical and depth data [4], like [5]. In our previous work, we firstly extracted optical and depth features applying Principal Component Analysis (PCA) referring to [6]. Multi-stream HMMs were employed to balance optical and depth streams, like conventional Audio-Visual Speech Recognition (AVSR). In terms of AVSR, we have also developed a recognizer utilizing Deep Bottle-Neck Feature (DBNF) technology [7]. Mel-Frequency Cepstral Coefficient (MFCC) features were extracted from the audio modality. For each frame in the visual stream, an appearance-based feature set was obtained including PCA, Discrete Cosine Transform (DCT), Linear Discriminant Analysis (LDA), and our original feature Genetic-algorithm-based Informative Feature (GIF) [8]. One shape-based feature set was also added to the feature set. Then we applied a DNN to the audio and visual features respectively, to obtain DBNFs. Audio DBNF and Visual DBNF were subsequently concatenated, followed by performing AVSR with multi-stream HMMs.

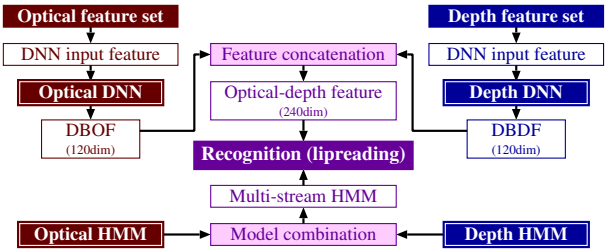
In this paper, we improve our lipreading method using optical and depth information, by choosing the DBNF technique. In each modality, DBNFs are computed as we did for our AVSR, excluding the shape information. Both DBNFs are combined frame by frame to obtain visual features. Optical and depth HMMs are also merged as multi-stream HMMs for lipreading.



(a) basic feature extraction

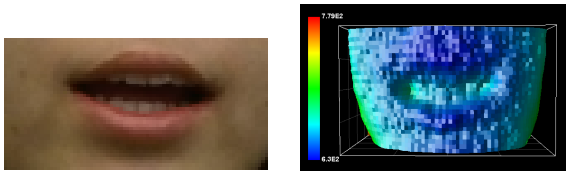


(b) HMM and DNN training



(c) recognition

Figure 1: *Our lipreading method.*



(a) optical image (b) depth image

Figure 2: *Sample images in optical and depth modalities.*

Table 1: *Experimental setup of DNN.*

	Input (1)	Hidden (2,3,4)	Bottleneck (5)	Hidden (6)	Output (7)
Layer size	1,551	2,048	40	2,048	179

	Pre-training	Fine-tuning
Epochs	10	50
Minibatch size	256	256
Learning ratio	0.00004	0.00006
Momentum	0.9	0.0

2. Lipreading

2.1. Recording and basic feature extraction

Figure 1 illustrates a flow of our lipreading scheme. We chose Microsoft Kinect v2 as a recoding device, capturing RGB and depth face data shown in Figure 2. Mouth detection is applied to optical images using Haar-like features. After cropping images, converting into monochrome images and reducing the resolution, we obtain optical and depth mouth images. Note that speech signals are also recorded simultaneously for model training. For more details about data acquisition, please refer to [4].

In the optical modality, four kinds of appearance-based features (PCA, DCT, LDA, GIF) are extracted from an image [7]. After calculating Δ and $\Delta\Delta$ coefficients as well as upsampling, a 141-dimensional optical feature set is obtained. Similarly, a depth feature set is computed from a depth image.

2.2. Training

Training DNNs and HMMs is the same as [7]. MFCCs are calculated from speech signals, subsequently audio HMMs are generated. Forced alignment is applied to get HMM state alignment, that is converted into a vector sequence for DNN training.

For the optical modality, a seven-layer full-connected DNN having a bottleneck layer is built; its input layer corresponds to the basic feature set, while the output layer corresponds to an HMM state vector. In this case, the input vector consists of previous, current and following frames. The DNN is then used for feature extraction from the basic feature set; output values of the bottleneck layer are composed into an optical feature vector, Deep Bottleneck Optical Feature (DBOF). Using DBOFs in a training set, optical HMMs are consequently trained.

For the depth modality, a sequence of Deep Bottleneck Depth Feature (DBDF) is obtained as well. Depth HMMs are finally made from DBDFs in the training data set.

2.3. Recognition

Before recognition, both optical and depth HMMs which are obtained in the above model training are combined into multi-stream HMMs. DBOF and DBDF vectors for test data are concatenated frame by frame. Lipreading is then conducted using the multi-stream HMMs and concatenated DBNFs. Similar to AVSR, we must properly set a stream weight ($0 \leq \lambda \leq 1$) for each modality, to balance contribution of optical and depth modalities. Note that we manually optimize balancing parameters in this paper, because we simply want to compare discriminative ability of conventional and proposing features.

3. Experiment

3.1. Database and experimental setup

In order to evaluate effectiveness and usefulness of our new lipreading scheme, we conducted lipreading experiments. A database used in this work was the same as [4]. We recorded speech signals and visual data including optical and depth images from 10 subjects (A–J), in acoustically and visually clean condition. Similar to an audio-visual corpus CENSREC-1-AV [9], in our database each speaker uttered 77 connected digits.

We conducted recognition experiments in a leave-one-out manner; for a test set including only one speaker's data, we used optical and depth data from the other nine speakers when training models. The DNN setup are shown in Table 1, and the other setup should be referred to [4, 7].

Table 2: *Lipreading accuracy [%] using conventional and proposed methods.*

Spkr ID	Conv. (PCA)	Prop. (DBNF)	Spkr ID	Conv. (PCA)	Prop. (DBNF)
A	23.32	35.57	F	20.95	40.71
B	18.58	35.97	G	23.23	38.98
C	18.65	36.90	H	35.57	60.08
D	33.20	42.29	I	25.69	44.66
E	22.92	26.88	J	22.05	45.28
			Ave.	24.42	40.73

3.2. Result and discussion

Table 2 shows recognition accuracy, considering deletion, substitution, and insertion errors. Stream weight factors were manually set for each speaker, so that the highest accuracy could be obtained. From Table 2, it is obviously found that the proposed scheme utilizing DBNFs achieved better performance for all the subjects compared to the conventional one adopting PCA only.

The difference between the previous and proposed methods is adding several basic features (DCT, LDA and GIF), and applying the DBNF technology. We also did an additional experiment using PCA, DCT, LDA and GIF, but not applying DBNF. It is then found that there is not enough difference compared to the method only choosing PCA. That means the improvement in this work mainly comes from the DBNF architecture.

Since the optimal stream weight for each speaker is quite different, automatic stream weight optimization is expected. We will also try to collect much more data in near future.

4. Acknowledgment

A part of this work was supported by JSPS KAKENHI Grant No. 16H03211.

5. References

[1] J. S. Chung et al., "Out of time: automated lip sync in the wild," Proc. ACCV2016 Workshop W9 (2016).

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

[3] Y. M. Assael et al., "LipNet: end-to-end sentence-level lipreading," arXiv:1611.01599v2 (2016).

[4] S. Tamura et al., "Visual speech recognition using optical and depth image features," Proc. FCV2016, pp.17-21 (2016).

[5] A. Rekik et al., "A new visual speech recognition approach for RGB-D cameras," Proc. ICIAR2014, pp.21-28 (2014).

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

[7] S. Tamura et al., "Audio-visual speech recognition using deep bottleneck features and high-performance lipreading," Proc. APSIPA ASC 2015 (2015).

[8] N. Ukai et al., "GIF-LR: GA-based informative feature for lipreading," Proc. APSIPA ASC 2012, PS.3-IVM.7.5 (2012).

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