INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Audio-Visual Speech Recognition Using Bimodal-Trained Bottleneck Features for a Person with Severe Hearing Loss

Yuki Takashima[1], *Ryo Aihara*[1], *Tetsuya Takiguchi*[1], *Yasuo Ariki*[1],
Nobuyuki Mitani[2], Kiyohiro Omori[2], Kaoru Nakazono[2]

[1]Graduate School of System Informatics, Kobe University, Japan
[2]Hyogo Institute of Assistive Technology, Kobe, Japan

y.takasima@me.cs.scitec.kobe-u.ac.jp, aihara@me.cs.scitec.kobe-u.ac.jp,
takigu@kobe-u.ac.jp, ariki@kobe-u.ac.jp

Abstract

In this paper, we propose an audio-visual speech recognition system for a person with an articulation disorder resulting from severe hearing loss. In the case of a person with this type of articulation disorder, the speech style is quite different from those of people without hearing loss that a speaker-independent acoustic model for unimpaired persons is hardly useful for recognizing it. The audio-visual speech recognition system we present in this paper is for a person with severe hearing loss in noisy environments. Although feature integration is an important factor in multimodal speech recognition, it is difficult to integrate efficiently because those features are different intrinsically. We propose a novel visual feature extraction approach that connects the lip image to audio features efficiently, and the use of convolutive bottleneck networks (CBNs) increases robustness with respect to speech fluctuations caused by hearing loss. The effectiveness of this approach was confirmed through word-recognition experiments in noisy environments, where the CBN-based feature extraction method outperformed the conventional methods.

Index Terms: multimodal, lip reading, deep-learning, assistive technology

1. Introduction

In recent years, a number of assistive technologies using information processing have been proposed; for example, sign language recognition using image recognition technology [1] and text reading systems from natural scene images [2]. In this study, we focused on communication assistive technology for a physically unimpaired person to communicate a person with an articulation disorder resulting from severe hearing loss.

Some people with hearing loss who have received speech training or who lost their hearing after learning to speak can communicate using spoken language. However, in the case of automatic speech recognition (ASR), their speech style is so different from that of people without hearing loss that a speaker-independent (audio-visual) ASR model for unimpaired persons is hardly useful for recognizing such speech. Matsumasa *et al.* [3] researched an ASR system for articulation disorders resulting from cerebral palsy and reported the same problem.

The performance of speech recognition is generally degraded in a noisy environment. For people with hearing loss, because they do not hear ambient sound, they cannot control the volumes of their voices and the speaking style in a noisy environment, and it is difficult to recognize utterances using only the speech signal for us. Then, we use the lip image of speaker to compensate for recognition accuracy. For people with hearing problems, lip reading is one communication skill that can help them communicate better. In the field of speech processing, audio-visual speech recognition has been studied for robust speech recognition under noisy environments [4, 5]. In this paper, we propose an audio-visual speech recognition for articulation disorders resulting from severe hearing loss.

We employ a bottleneck feature extraction method from audio-visual features using convolutive bottleneck networks (CBN), which stack multiple layers of various types (such as a convolution layer, a subsampling layer, and a bottleneck layer) [6] forming a deep network. Thanks to the convolution and pooling operations, we can train the convolutional neural network (CNN) robustly to deal with the small local fluctuations of an input feature map. In some approaches using deep learning, an output layer plays a classification role, and output units are used as a feature vector for a recognition system, where phone labels are used as a teaching signal for an output layer. On the other hands, an approach based on CBN [7] uses a bottleneck layer as a feature vector for a recognition system, where the number of units is extremely small compared with the adjacent layers, following the CNN layers. In the case of an articulation disorder, the phone label estimated by forced alignment may not be correct. Therefore, the bottleneck layer is a better feature than an output layer which is strongly influenced by some wrong phone labels because it is expected that the bottleneck layer can aggregate propagated information and extract fundamental features included in an input map.

In the most multimodal speech recognition system, audio and visual features are integrated by just concatenating these features. Because the audio and visual features are intrinsically different, a gap between audio and visual feature spaces may cause undesirable effects in speech recognition. Therefore, we propose a novel visual feature extraction method that converts the lip image into an audio feature in a covolutive network, which has an affinity with an audio feature.

Experimental results confirmed that our bottleneck features increase robustness for small local fluctuations that are caused by the utterances of those who have a hearing loss. Moreover, we confirmed that the visual feature extracted by our method compensates for the difference between audio and visual spaces in speech recognition.

http://dx.doi.org/10.21437/Interspeech.2016-721

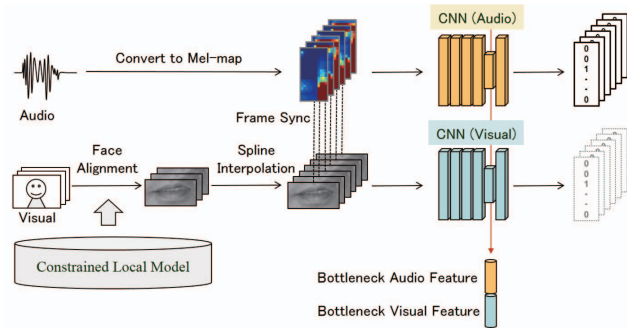

Figure 1: Flow of the feature extraction

2. RELATED WORKS

As one of the techniques used for robust speech recognition under noisy environments, audio-visual speech recognition, which uses lip dynamic visual information and audio information, has been studied. In audio-visual speech recognition, there are mainly three integration methods: early integration [8], which connects the audio feature vector with the visual feature vector; late integration [9], which weights the likelihood of the result obtained by a separate process for audio and visual signals; and synthetic integration [4], which calculates the product of output probability in each state.

In audio-visual speech recognition, detecting face parts (for example, eyes, mouth, nose, eyebrows, and outline of face) is an important task. The detection of these points is referred to as face alignment. In this paper, we employed a constrained local model (CLM) [10]. A CLM is a subject-independent model that is trained from a large number of face images.

In recent years, an ASR system has been applied as assistive technology for people with articulation disorders. During the last decades, an ASR system for a person with cerebral palsy has been researched. In [3], robust feature extraction based on principal component analysis (PCA) with more stable utterance data was proposed. In [11], multiple acoustic frames (MAF) was used as an acoustic dynamic feature to improve the recognition rate of a person with an articulation disorder, especially in speech recognition using dynamic features only.

Deep learning has had recent successes for acoustic modeling [12]. Deep neural networks (DNNs) contain many layers of nonlinear hidden units. The key idea is to use greedy layer-wise training with restricted Boltzmann machines (RBMs) followed by fine-tuning. Ngiam *et al.* [13] proposed multimodal DNNs that learn features over audio and visual modalities. Mroueh *et al.* [14] improved this method and proposed an architecture considering the correlations between modalities. Ninomiya *et al.* [5] investigated integration of bottleneck features using multi-stream hidden Markov models (HMMs) for audio-visual speech recognition.

In this paper, we employ a convolutional neural network (CNN) [6]-based approach to extract robust features from audio and visual features. The CNN is regarded as a successful tool and has been widely used in recent years for various tasks, such as image analysis [15, 16, 17] and spoken language [18], music recognition [19].

3. FEATURE EXTRACTION USING CBN

3.1. Flow of The Feature Extraction

Figure 1 shows the flow of feature extraction. First, we prepare the input features for training a CBN from lip images and speech signals uttered by a person with hearing loss. For the audio signals, after calculating short-term mel spectra from the signal, we obtain mel-maps by merging the mel spectra into a 2D feature with several frames, allowing overlaps.

The visual signals of the eyes, mouth, nose, eyebrows, and outline of the face are aligned using the point distribution model (PDM) and its model parameter is estimated by constrained local model (CLM) and a lip image is extracted. The extracted lip image is interpolated to fill the sampling rate gap between visual features with respect to audio features. In this paper, We adopted the spline interpolation to the lip images.

For the output units of the CBN, we use phoneme labels that correspond to the input mel-map and lip images. Audio and visual CBNs are separately trained, and the parameters of the CBN are trained by back-propagation with stochastic gradient descent, starting from random values. Following the training of CBNs, the input mel-map and lip images are converted to the bottleneck feature by using each CBN. Then these features are concatenated, and used in the training of HMMs for speech recognition.

In the test stage, we extract features using each CBN, which tries to produce the appropriate phoneme labels in the output layer. Again, note that we do not use the output (estimated) labels for the following procedure, but we use the BN features in the middle layer, where it is considered that information in the input data is aggregated. Finally, extracted bottleneck audio and visual features are simply concatenated and used as the input features of HMMs to audio-visual speech recognition. In our previous work [20], we evaluated the early and late integration for the similar system. Then, because the performances are equivalent between two integrations, we empoy the early integration in this paper.

3.2. Convolutive Bottleneck Network

A CBN [21] consists of an input layer, a pair of a convolution layer and a pooling layer, fully-connected Multi-Layer Perceptrons (MLPs) with a bottleneck structure, and an output layer. The MLP stacks some layers, and the number of units in a middle layer is reduced as “bottleneck features”. The number of units in each layer is discussed in the experimental section. Since the bottleneck layer has reduced the number of units for the adjacent layers, we can expect that each unit in the bottleneck layer aggregates information and behaves as a compact feature descriptor that represents an input with linear discriminant analysis (LDA) or PCA. In this paper, audio and visual features are input to each CBN and extracted bottleneck features are used for multimodal speech recognition.

4. PROPOSED VISUAL FEATURE EXTRACTION NETWORK

We propose a novel visual feature extraction method that has an affinity with the audio feature. Note that our proposed method is motivated by the difference in the feature spaces between two modalities. Figure 2 depicts the proposed visual CBN that consists of two netwroks, where C, S, and M denote the convolutional layer, subsampling layer, and MLPs, respectively.

First, we train a network from the input layter to M2 layter

in Figure 2. In this network, lip images are fed to the input layter and segment acoustic features are used as the teaching signal. This network has a roll to convert a lip image into an acoustic feature. Next, we train a network from M2 layter to the output layer in Figure 2. In this network, the mel-frequency cepstral coefficients (MFCC) segment features are fed to the input layter and phoneme labels are used as teaching signals. Finally, by coupling the above two networks, the whole network is composed, that is our proposed visual CBN. Then we fine-tune this network where lip images and phoneme labels are set for input and output layers. By transforming the lip image into the acoustic feature in the middle layer, it is expected that the gap between the visual feature and the audio feature is compensated for. In the evaluation step, we use this visual CBN to extract the visual bottleneck features at the M4 layer.

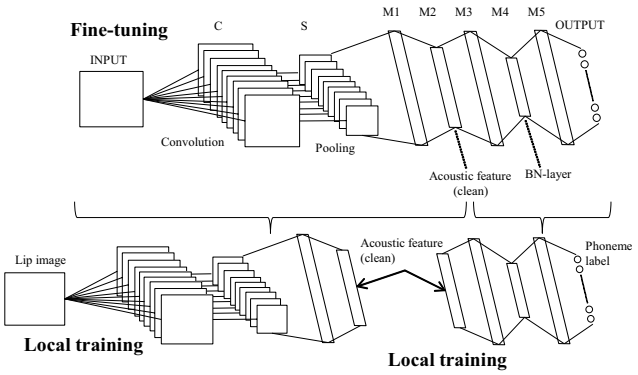

Figure 2: Proposed Visual CBN

5. Experiment

5.1. Experimental conditions

Our proposed method was evaluated on word recognition tasks. We recorded utterances of one male person with hearing loss, where the text is the same as the ATR Japanese speech database A-set. We used 2,620 words as training data, and 216 words as test data. The utterance signal was sampled at 16 kHz and windowed with a 25-msec Hamming window every 5 msec. For the acoustic-visual model, we used the monophone-HMMs (54 phonemes) with 3 states and 6 mixtures of Gaussians. The number of units of bottleneck features is 30. Therefore, input features of HMM are 30-dimensional acoustic features and 30-dimensional visual features. We compare our audio-visual feature with conventional MFCC+Δ+$\Delta\Delta$ (36-dimensions) and MFCC+Δ+$\Delta\Delta$+ discrete cosine transform (DCT) (66-dimensions). First, we compared our proposed CBN-based visual features with DNN-based visual features in lip reading. The numbers of units in each layer of the DNN are set to 100, 30, 100, 54, and the input vector is a concatenated lip image. Then, our proposed method and audio-visual features are evaluated in noisy environments. White noise is added to audio signals and its SNR is set to 20dB, 10dB, and 5dB. Audio CBN and HMMs are trained by using the clean audio feature.

5.2. Architecture of CBN

As shown in Figure 2, we use deep networks, which consist of a convolution layer, a pooling layer, and fully-connected MLPs. For the input layer of audio CBN, we use a mel-map of subsequent 13-frames with 39-dimensional-melspectrum, and the frame shift is 5 msec. For the input layer of visual CBN, frontal face videos are recorded at 60 fps. Luminance images are extracted from the image by using CLM and resized to 12×24 pixels. Finally, the images are up-sampled by spline interpolation and input to the CBN.

Table 1 shows parameters used in experiments, and Figure 3 depicts architectures of evaluated CBNs. The audio-visual bottleneck feature I (AV_BNF I) indicates that the audio and visual features are extracted by each CBN that has the same architecture (Arch 1 in Table 1). The audio-visual bottleneck feature II (AV_BNF II), which was described in section 4, indicates that the audio feature and the visual feature are extracted by Arch 1 and Arch 2 in Table 1, respectively. MFCCs of ± 2 frames (subsequent 5-frames, 12-dimentions) are used as an acoustic feature of a proposed visual CBN in Figure 2.

Table 1: Filter size, number of feature maps and number of MLPs units for each architecture. The value for C indicates the filter size of the convolution layer that has #1 maps. The convolution layer is associated with the pooling layer. The value of S means the pooling factor. The value for M indicates the number of units for each layer in the MLP part.

	Input	C	S	#1	M
Arch 1	39×13	4×2	3×3	13	108, 30, 108
Arch 2	12×24	5×5	2×2	13	108, 60, 108, 30, 108

5.3. Experimental Results

Figure 4 shows the experimental results using the lip-based feature. As shown in the figure, the feature extracted by CNN obtained a better result than the feature extracted by DNN. Our proposed visual network slightly outperformed the simple CBN structured by Arch 1 in Table 1.

We compared the audio-visual feature using our proposed visual feature with four conventional features: MFCC+Δ+$\Delta\Delta$, MFCC+Δ+$\Delta\Delta$+DCT, audio bottleneck features (BN_Audio), and AV_BNF I. In the integration step, an audio feature and a visual feature are combined into a single frame, and the combined feature is used as an input feature for the HMMs. Figure 5

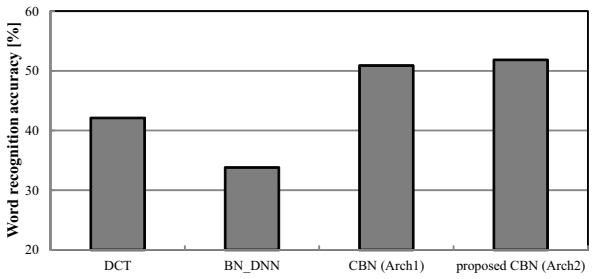

Figure 4: Word recognition accuracy using HMMs trained by the lip-based features

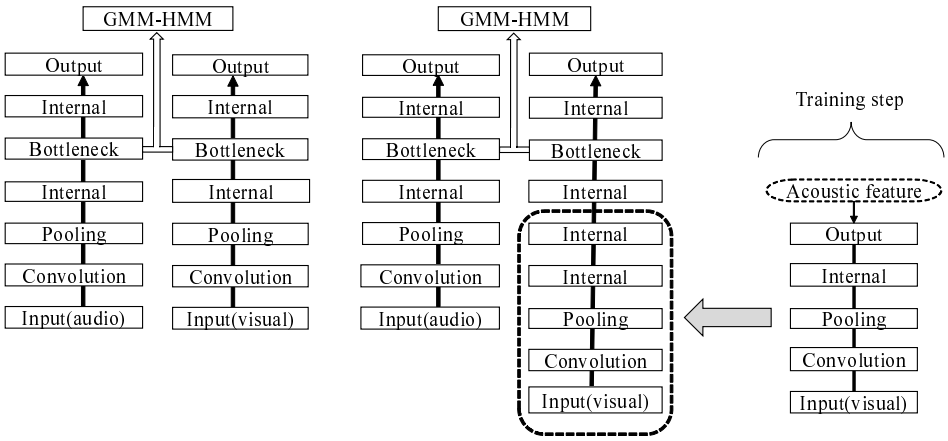

Figure 3: Evaluation network architectures: AV_BNF I (left), AV_BNF II (right)

shows the word recognition accuracies in noisy environments. The bottleneck audio feature shows the best results compared to other features at the clean environment and SNR of 20dB. This result shows that CBN features have a robustness to the small local fluctuations in a time-mel-frequency map that is caused by the articulation disordered speech. Our proposed audio-visual feature outperforms the AV_BNF I in the clean environment and SNR of 20dB, where the integrated features between the audio and the proposed visual bottleneck features improved 3.3% and 3.8% compared with the AV_BNFs I, respectively. This is because our method compensates for the difference between audio and visual spaces. However, at the SNRs of 10dB and 5dB, the integrated feature using our proposed feature could not improve the accuracy in comparison with that of the AV_BNF I. One of the reasons is that a visual feature in our proposed method is trained using clean acoustic features. Considering these results of the clean and SNR 20dB environments, if the visual feature were trained with noisy conditions, the proposed method might achieve a better result compared with the method using the same structured audio and visual CBN.

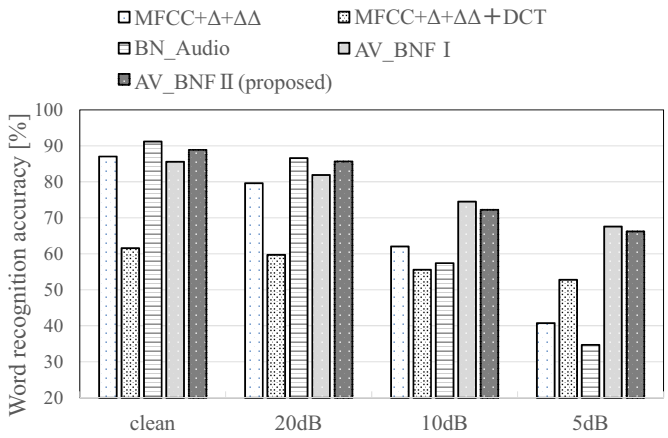

Figure 5: Word recognition accuracy using HMMs

6. Conclusions

In this paper, we proposed a visual feature, which has an affinity to the audio feature, extracted from a CBN for articulation disorders resulting from severe hearing loss. In recognition experiments, we confirmed that our proposed audio-visual feature obtained a better result than audio-visual features extracted by the same structured audio and visual CBNs in clean or high-SNR environments. However, because our proposed features are trained by clean audio features, the recognition rate using our proposed features are slightly lower than a method using the same structured audio and visual CBNs in the low SNR. In future work, we will further investigate a better audio and visual feature integration method.

7. References

[1] J. Lin, Y. Wu, and T. S. Huang, "Capturing human hand motion in image sequences," in *Workshop on Motion and Video Computing*, 2002, pp. 99–104.

[2] N. Ezaki, M. Bulacu, and L. Schomaker, "Text detection from natural scene images: Towards a system for visually impaired persons," in *ICPR*, 2004, pp. 683–686.

[3] H. Matsumasa, T. Takiguchi, Y. Ariki, I. chao Li, and T. Nakabayashi, "Integration of metamodel and acoustic model for dysarthric speech recognition," *Journal of Multimedia*, pp. 254–261, 2009.

[4] M. Tomlinson, M. Russell, and N. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," in *ICASSP*, 1996, pp. 821–824.

[5] H. Ninomiya, N. Kitaoka, S. Tamura, Y. Iribe, and K. Takeda, "Integration of deep bottleneck features for audio-visual speech recognition," in *INTERSPEECH*, 2015.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, 1998, pp. 2278–2324. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.7665

[7] T. Nakashika, T. Yoshioka, T. Takiguchi, Y. Ariki, S. Duffner, and C. Garcia, "Dysarthric speech recognition using a convolutive bottleneck network," in *ICSP*, 2014, pp. 505–509.

[8] G. Potamianos and H. P. Graf, "Discriminative training of hmm stream exponents for audio-visual speech recognition," in *ICASSP*, 1998, pp. 3733–3736.

[9] A. Verma, T. Faruquie, C. Neti, S. Basu, and A. Senior, "Late integration in audio-visual continuous speech recognition," in *ASRU*, 1999, pp. 71–74.

[10] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models." in *BMVC*, vol. 1, no. 2, 2006, p. 3.

[11] C. Miyamoto, Y. Komai, T. Takiguchi, Y. Ariki, and I. Li, "Multimodal speech recognition of a person with articulation disorders using aam and maf," in *Multimedia Signal Processing*, 2010, pp. 517–520.

[12] G. Hinton, D. Li, Y. Dong, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82 – 97, 2012.

[13] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *International Conference on Machine Learning*, 2011, pp. 689–696.

[14] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition." in *ICASSP*, 2015, pp. 2130–2134.

[15] C. Garcia and M. Delakis, "Convolutional face finder: A neural architecture for fast and robust face detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 11, pp. 1408–1423, 2004.

[16] M. Delakis and C. Garcia, "Text detection with convolutional neural networks." in *VISAPP (2)*, 2008, pp. 290–294.

[17] R. Hadsell, P. Sermanet, J. Ben, A. Erkan, M. Scoffier, K. Kavukcuoglu, U. Muller, and Y. LeCun, "Learning long-range vision for autonomous off-road driving," *Journal of Field Robotics*, vol. 26, no. 2, pp. 120–144, 2009.

[18] G. Montavon, "Deep learning for spoken language identification," in *NIPS Workshop on deep learning for speech recognition and related applications*, 2009.

[19] T. Nakashika, C. Garcia, T. Takiguchi, and I. De Lyon, "Local-feature-map integration using convolutional neural networks for music genre classification." in *INTERSPEECH*, 2012.

[20] Y. Takashima, Y. Kakihara, R. Aihara, T. Takiguchi, Y. Ariki, N. Mitani, K. Omori, and K. Nakazono, "Audio-visual speech recognition using convolutive bottleneck networks for a person with severe hearing loss," *IPSJ Transactions on Computer Vision and Applications*, vol. 7, no. 0, pp. 64–68, 2015.

[21] K. Veselỳ, M. Karafiát, and F. Grézl, "Convolutive bottleneck network features for LVCSR," in *ASRU*, 2011, pp. 42–47.