IPSJ Transactions on Computer Vision and Applications Vol.7 64–68 (July 2015)

[DOI: 10.2197/ipsjtcva.7.64]

Express Paper

Audio-Visual Speech Recognition Using Convolutive Bottleneck Networks for a Person with Severe Hearing Loss

Yuki Takashima[1,a)] Yasuhiro Kakihara[1,b)] Ryo Aihara[1,c)]
Tetsuya Takiguchi[1,d)] Yasuo Ariki[1,e)] Nobuyuki Mitani[2]
Kiyohiro Omori[2] Kaoru Nakazono[2]

Received: March 13, 2015, Accepted: April 20, 2015, Released: July 27, 2015

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 with the result that of people without hearing loss that a speaker-independent model for unimpaired persons is hardly useful for recognizing it. We investigate in this paper an audio-visual speech recognition system for a person with severe hearing loss in noisy environments, where a robust feature extraction method using a convolutive bottleneck network (CBN) is applied to audio-visual data. We confirmed the effectiveness of this approach through word-recognition experiments in noisy environments, where the CBN-based feature extraction method outperformed the conventional methods.

Keywords: 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 [8], [16] and text reading systems from natural scene images [4]. In this study, we focused on a person with an articulation disorder resulting from severe hearing loss.

In Japan alone, there are 360,000 people suffering from 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 in recognizing such speech. Matsumasa et al. [9] researched an ASR system for articulation disorders resulting from cerebral palsy and revealed the same problem.

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 [14], [18], [19]. In this paper, we propose an audio-visual speech recognition for articulation disorders resulting from severe hearing loss.

The main contribution of this paper is that we propose a bottleneck feature extracted from audio-visual features. Convolutive Bottleneck Network (CBN) [20], which stacks multiple layers of various types (such as a convolution layer, a subsampling layer, and a bottleneck layer) [6], [7] forming a deep network, is applied to audio-visual data. The bottleneck layer reduces the number of units for the adjacent layers, and, consequently, we can expect that each unit in the bottleneck layer aggregates information and behaves as a compact feature descriptor that represents an input with a small number of bases.

Our experimental results confirmed that our bottleneck features have robustness for small local fluctuations that are caused by the utterances of those who have hearing loss. Moreover, our integration of audio and visual features acquired robustness in noisy environments.

The rest of this paper is organized as follows: In Section 2, related works are introduced. In Section 3, the flow of our proposed method is explained. In Section 4, the face alignment problem in lip reading is described. In Section 5, our proposed bottleneck feature is explained. In Section 6, the experimental data are evaluated, and the final section is devoted to our conclusions.

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,

[1] Graduate School of System Informatics, Kobe University, Kobe, Hyogo 657–8501, Japan
[2] Hyogo Institute of Assistive Technology, Kobe, Hyogo 651–2134, Japan
[a)] y.takasima@me.cs.scitec.kobe-u.ac.jp
[b)] kakyhara@me.cs.scitec.kobe-u.ac.jp
[c)] aihara@me.cs.scitec.kobe-u.ac.jp
[d)] takigu@kobe-u.ac.jp
[e)] ariki@kobe-u.ac.jp

© 2015 Information Processing Society of Japan

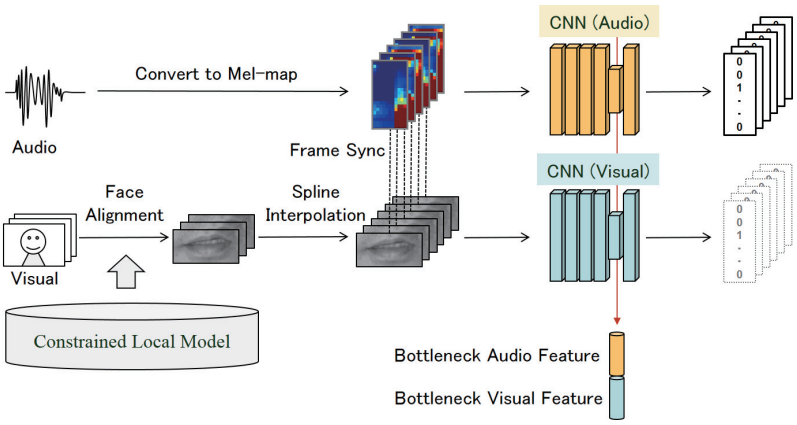

Fig. 1 Flow of the feature extraction.

has been studied. In audio-visual speech recognition, there are mainly three integration methods: early integration [14], which connects the audio feature vector with the visual feature vector; late integration [19], which weights the likelihood of the result obtained by a separate process for audio and visual signals; and synthetic integration [18], 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. The Active Appearance Model (AAM) [1] and Active Shape Model (ASM) [17] are well-known face alignment models. In this paper, we employed a Constrained Local Model (CLM) [2], [15]. 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, we have researched an ASR system for a person with cerebral palsy. In Ref. [9], we proposed robust feature extraction based on principal component analysis (PCA) with more stable utterance data instead of discrete cosine transform (DCT). In Ref. [10], we used multiple acoustic frames (MAF) 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 [5]. 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.

In this paper, we employ a Convolutional Neural Network (CNN) [6], [7]-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 [3] and spoken language [11]. In Ref. [12], CNN is employed as robust feature extraction for the fluctuation of the speech uttered by a person with cerebral palsy. Experimental results in Ref. [12] revealed that the convolution and pooling operations in CNN have a robustness to the small local fluctuation which is caused by motor paralysis resulting from athetoid cerebral palsy.

3. Flow of the Proposed Method

Figure 1 shows the flow of our proposed feature extraction. First, we prepare the input features for training a CBN from audio and visual signals. For the audio signals, after calculating short-term mel spectra from the signal, we obtain mel-maps by dividing the mel spectra into segments with several frames, allowing overlaps.

The visual signals of the eyes, mouth, nose, eyebrows, and outline of the face are aligned using a Constrained Local Model (CLM) and a lip image is extracted. The details of lip image extraction are explained in the following section. The extracted lip image is interpolated to fill the sampling rate gap between audio features.

For the output units of the CBN, we use phoneme labels that correspond to the input mel-map and lip images. Audio and visual CBN are separately trained. The input mel-map and lip images are converted to the bottleneck feature by using each CBN. Extracted features are used as the input feature of Hidden Markov Models (HMM).

4. Lip Image Extraction Using CLM

Face alignment of this paper is conducted by using the Point Distribution Model (PDM) and its model parameter is estimated by CLM. CLM consists of two steps. The first step is the face point detection and the second step is parameter estimation.

4.1 PDM

We model a facial image of a large number of people by using the PDM which models a facial image by 2-dimensional shape vectors. The position vector which corresponds to the point of the PDM is defined as follows:

$$\mathbf{X} = (\mathbf{X}_1^{\mathrm{T}}, \ldots, \mathbf{X}_M^{\mathrm{T}})^{\mathrm{T}} \tag{1}$$

where $\mathbf{X}_i = (x_i, y_i)^{\mathrm{T}}$ and M denote the i-th point of PDM and the number of points of PDM, respectively. The position vector is represented as follows:

$$\mathbf{X} = \overline{\mathbf{X}} + \mathbf{\Phi q} \tag{2}$$

where $\mathbf{\Phi}$, $\mathbf{q}$ and $\overline{\mathbf{X}}$ denote the principal vectors extracted by Prin-

© 2015 Information Processing Society of Japan

cipal Component Analysis (PCA), the parameter vector and the mean vector of the shape vector, respectively. By using PDM, the i-th point on the image, $\mathbf{X}_i(\mathbf{p})$, is represented as follows:

$$\mathbf{X}_i(\mathbf{p}) = s\mathbf{R}[\overline{\mathbf{X}}_i + \mathbf{\Phi_i q}] + \mathbf{t} \quad (3)$$

where $\mathbf{p} = \{s, \mathbf{R}, \mathbf{t}, \mathbf{q}\}$ denotes the parameter set. s denotes a scale and $\mathbf{R}$ denotes a rotation which consists of pitch α, yaw β, roll γ. $\mathbf{t}$, $\mathbf{q}$ and $\mathbf{\Phi_i}$ denote the shift vector, the parameter vector and the i-th principal vector, respectively.

4.2 CLM

The parameter of PDM is estimated by using CLM. First, feature points are detected by Support Vector Machine (SVM) which is trained by a large number of facial images.

Then, the model parameter $\mathbf{p}$ is estimated from the i-th detected feature point $\hat{\mathbf{X}}_i$ by minimizing the following equation:

$$Q(\mathbf{p}) = \sum_{i=1}^{M} \|\hat{\mathbf{X}}_i - \mathbf{X}_i(\mathbf{p})\|^2 + R(\mathbf{p}) \quad (4)$$

where $R(\mathbf{p})$ is a regularization term to avoid over fitting. In this paper, we defined $R(\mathbf{p})$ as normal distribution of $N(0, \mathbf{\Lambda})$.

5. Feature Extraction Using CBN

5.1 Convolutive Bottleneck Network

A CBN 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 as shown in **Fig. 2**. C, S, and M denote convolutional layer, subsumpling layer, and MLPs, respectively. The MLP shown in Fig. 2 stacks three layers (M1, M2, M3), and the number of units in the middle layer (M2) 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 a small number of bases, similar to other feature descriptors, such as MFCC, 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.

5.2 Bottleneck Feature Extraction

First, we train audio and visual CBN. We prepare the input features for training a CBN from an image and speech signal uttered by person with hearing loss. For the audio feature, we obtain mel-maps by dividing the mel spectra into segments with several frames, allowing overlaps. For the output units of the CBN, we use phoneme labels that correspond to the input mel-map. For example, when we have a mel-map with the label /i/, only the unit corresponding to the label /i/ is set to 1, and the others are set to 0 in the output layer. The label data is obtained by forced alignment using HMMs from the speech data.

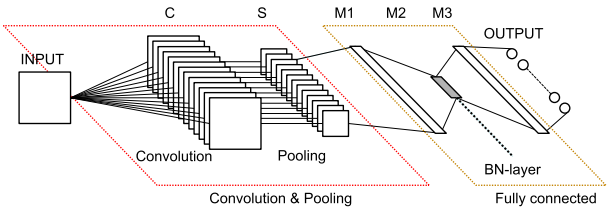

Fig. 2 Convolutional bottleneck network.

For the visual features, because its sampling rate is smaller than the audio signal, spline interpolation is adopted to the images in order to fill the sampling rate gap. The output units of the CBN are the same as that of the audio features.

The parameters of the CBN are trained by back-propagation with stochastic gradient descent, starting from random values. The bottleneck (BN) features in the trained CBN are then used in the training of an HMM for speech recognition. In the test stage, we extract features using the 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 used as the input features of audio or visual HMMs and the recognition results are integrated. Details about this integration are discussed in Section 6.3.

6. Experiment

6.1 Experimental Conditions

Our proposed method was evaluated on word recognition tasks for one male person with hearing loss. We recorded 216 words included in the ATR Japanese speech database B-set which are used as test data and 2,620 words included in the ATR Japanese speech database A-set which are used as training data. The utterance signal was sampled at 16 kHz and windowed with a 25-msec Hamming window every 10 msec. For the acoustic model, we used the monophone-HMMs (54 phonemes) with 5 states and 6 mixtures of Gaussians. For the visual model, we used the monophone-HMMs (54 phonemes) with the same states and mixtures of Gaussians to the acoustic model. 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 bottleneck feature with conventional MFCC+ΔMFCC (30-dimensions). Furthermore, we evaluated our method in noisy environments. We added white noise to audio signals and its SNR is set to 20 dB, 10 dB, and 5 dB. Audio CBN and HMMs are trained by using the clean audio feature.

6.2 Architecture of CBN

As shown in Fig. 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 39-dimensional-melspectrum × 13, and the frame shift is 1. 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 the size of each feature map. The numbers of units in each layer of MLPs are set to 108, 30, 54. Those numbers are the same to audio CBN and visual CBN.

© 2015 Information Processing Society of Japan

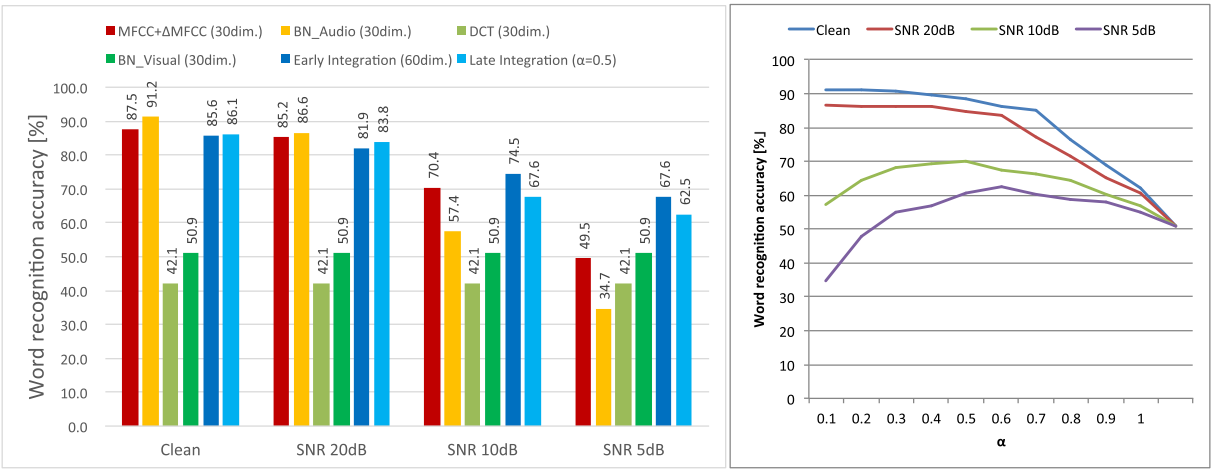

Fig. 3 Word recognition accuracy using HMMs.

Table 1 Size of each feature map. $(k, i \times j)$ indicates that the layer has k maps of size $i \times j$.

	Input	C1	S1
Audio CBN	$1, 39 \times 13$	$13, 36 \times 12$	$13, 12 \times 4$
Visual CBN	$1, 12 \times 24$	$13, 8 \times 20$	$13, 4 \times 10$

6.3 Experimental Results

Compared input features for the HMMs are listed as follows:

- MFCC+ΔMFCC
- Audio Bottleneck features (BN Audio)
- Discrete Cosine Transformation (DCT)
- Visual Bottleneck features (BN Visual)
- Early integration of BN Audio and BN Visual
- Late integration of BN Audio and BN Visual

In early integration, an audio feature and a visual feature are combined into a single frame and this frame used as an input feature for the HMMs. In late integration, an audio feature and a visual feature are input to each audio and visual HMM, and the output likelihood is integrated as follows:

$$L_{A+V} = \alpha L_V + (1-\alpha) L_A , \quad 0 \le \alpha \le 1 \tag{5}$$

where L_{A+V}, L_A, L_V and α denote integrated likelihood, likelihood of an audio feature, likelihood of a visual feature, and weights of likelihood, respectively.

The left side of **Fig. 3** shows the word recognition accuracies in noisy environments. The bottleneck audio feature shows the best results compared to conventional MFCC at the clean environment and SNR of 20 dB. This is due to the robustness of the CBN features to small local fluctuations in a time-mel-frequency map, caused by the articulation disordered speech.

The word recognition rate of lip reading using the bottleneck visual feature is 50.9%. At the SNR of 10 dB, the early integration between audio and visual bottleneck features improved 4.1% from our baseline. Moreover at the SNR of 5 dB, the early integration between audio and visual bottleneck features improved 18.1% from our baseline. t can be seen from these results that multimodal features are shown to be effective in noisy environments.

The right side of Fig. 3 shows the word recognition accuracies in the evaluation set as a function of the weight of the likelihood (α in Eq. (5)). $\alpha = 0.0$ in Fig. 3 shows the result of ASR using audio features only and $\alpha = 1.0$ in Fig. 3 shows the result of lip reading. This figure shows the best value for α under each condition. At the SNR of 10 dB and SNR 5 dB, the graph is convex, and these results show the effectiveness of multimodal features in noisy environments.

7. Conclusions

We proposed multimodal bottleneck features using CBN for articulation disorders resulting from severe hearing loss. Compared with conventional MFCC, our proposed audio bottleneck feature shows the better results. We assume that is because our bottleneck features are robust to small local fluctuations, which are caused by hearing loss. In noisy environments, our proposed method using multimodal bottleneck features shows its effectiveness in comparison to the other methods. Since the tendency of the fluctuations in articulation disordered speech depend on the speaker, we would like to apply and investigate our method to a variety of speakers with speech disorders in the future.

References

[1] Cootes, T.F.: Active Appearance Models, *Proc. European Conf. Computer Vision*, Vol.2, pp.484–498 (1998).
[2] Cristinacce, D. and Cootes, T.F.: Feature Detection and Tracking with Constrained Local Models, *Proc. British Machine Vision Conf.*, Vol.2, No.5, pp.929–938 (2006).
[3] Delakis, M. and Garcia, C.: Text detection with Convolutional Neural Networks, *Proc. Int. Conf. Computer Vision Theory and Applications*, pp.290–294 (2008).
[4] Ezaki, N., Bulacu, M. and Schomaker, L.: Text Detection from Natural Scene Images: Towards a System for Visually Impaired Persons, *Proc. Int. Conf. Pattern Recognition*, pp.683–686 (2004).
[5] Hinton, G., Li, D., Dong, Y., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N. and Kingsbury, B.: 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).
[6] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P.: Gradient-Based Learning Applied to Document Recognition, *Proc. IEEE*, Vol.86, No.11, pp.2278–2324 (1998).
[7] Lee, H., Largman, Y., Pham, P. and Ng, A.Y.: Unsupervised Feature Learning for Audio Classification using Convolutional Deep Belief Networks, *Proc. Neural Information Processing Systems*, Vol.22, pp.1096–1104 (2009).
[8] Lin, J., Ying, W. and Huang, T.S.: Capturing human hand motion in image sequences, *Proc. IEEE Motion and Video Computing Workshop*, pp.99–104 (2002).
[9] Matsumasa, H., Takiguchi, T., Ariki, Y., Li, I. and Nakabayashi, T.: Integration of Metamodel and Acoustic Model for Dysarthric Speech

© 2015 Information Processing Society of Japan

Recognition., *Journal of Multimedia*, Vol.4, No.4, pp.254–261 (2009).
[10] Miyamoto, C., Komai, Y., Takiguchi, T., Ariki, Y. and Li., I.: Multimodal Speech Recognition of a Person with Articulation Disorders Using AAM and MAF., *Proc. IEEE Int. Workshop on Multimedia Signal Processing*, pp.517–520 (2010).
[11] Montavon, G.: Deep learning for spoken language identification, *Proc. Workshop on Deep Learning for NIPS* (2009).
[12] Nakashika, T., Yoshioka, T., Takiguchi, T., Ariki, Y., Duffner, S. and Garcia, C.: Convolutive Bottleneck Network with Dropout for Dysarthric Speech Recognition, *Trans. Machine Learning and Artificial Intelligence*, Vol.2, No.2, pp.46–60 (2014).
[13] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. and Ng., A.: Multimodal deep learning, *Proc. International Conference on Machine Learning* (2011).
[14] Potamianos, G. and Graf, H.P.: Discriminative Training of HMM Stream Exponents for Audio-Visual Speech Recognition, *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp.3733–3736 (1998).
[15] Saragih, J.M., Lucey, S. and Cohn, J.F.: Deformable model fitting by regularized landmark mean-shift, *Int. Journal of Computer Vision*, Vol.91, No.2, pp.200–215 (2011).
[16] Starner, T., Weaver, J. and Pentland, A.: Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.20, No.12, pp.1371–1375 (1998).
[17] Sum, K., Lau, W., Leung, S., Liew, A.W.C. and Tse, K.W.: A new optimization procedure for extracting the point-based lip contour using active shape model, *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp.1485–1488 (2001).
[18] Tomlinson, M.J., Russell, M.J. and Brooke, N.M.: Integrating audio and visual information to provide highly robust speech recognition, *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp.821–824 (1996).
[19] Verma, A., Faruquie, T., Neti, C., Basu, S. and Senior, A.: Late Integration In Audio-Visual Continuous Speech Recognition, *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding* (1999).
[20] Vesely, K., Karafiat, M. and Grezl, F.: Convolutive Bottleneck Network features for LVCSR, *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pp.42–47 (2011).

(Communicated by *Atsushi Nakazawa*)

© 2015 Information Processing Society of Japan