Audio-Visual Speech Recognition for a Person with Severe Hearing Loss Using Deep Canonical Correlation Analysis

Yuki Takashima[1], *Tetsuya Takiguchi*[1], *Yasuo Ariki*[1], *Kiyohiro Omori*[2]

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

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

Abstract

Recently, we proposed an audio-visual speech recognition system based on a neural network 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 that of people without hearing loss, making a speaker-independent acoustic model for unimpaired persons more or less useless for recognizing it. Our proposed system has shown high performance; however, some problems remain. Although the feature extraction networks are trained using the phone labels as the target class, it is difficult to obtain the correct alignment for their speech. Also, it is necessary to consider a gap between audio and visual feature spaces to treat the different modalities. In this paper, we propose a feature extraction method using deep canonical correlation analysis to tackle these weaknesses. The effectiveness of this approach was confirmed through word-recognition experiments in noisy environments, where our feature extraction method outperformed the conventional methods.

Index Terms: Speech recognition, multimodal, deep canonical correlation analysis, 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 enable him or her to communicate with 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 of little use for recognizing such speech as described in Section 5.1. Matsumasa *et al.* [3] researched an ASR system for articulation disorders resulting from cerebral palsy, and reported the same problem. Najnin*et al.* [4] investigated the relationship between a hearing-impaired individual's speech and his hearing loss.

The performance of speech recognition systems generally degrades 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 their speaking style in a noisy environment, and it is difficult, those who are physically unimpaired, to recognize utterances using only the speech signal. In such cases, we try to read the lips of the speaker to compensate for the reduction in 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 [5, 6, 7]. In this paper, we investigate an audio-visual speech recognition approach for articulation disorders resulting from severe hearing loss.

Recently, we proposed bottleneck feature extraction [8] from audio and visual features for a hearing-impaired person using convolutive bottleneck networks (CBN), which stack multiple layers of various types (such as a convolution layer, a pooling layer, and a bottleneck layer) [9] 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 tandem 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. However, in the case of an articulation disorder, the phone label estimated by forced alignment may not be correct. An approach based on CBN [10] uses a bottleneck layer as a feature vector for a recognition system, where the number of units is extremely small compared to the adjacent layers, following the CNN layers. 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 the propagated information and extract fundamental features included in an input map. In this paper, we investigate another approach to tackle this alignment problem—unsupervised learning.

In multi-view learning, deep canonical correlation analysis (DCCA) [11], which is nonlinearly-extended canonical correlation analysis (CCA), has been proposed. DCCA has two deep neural networks and simultaneously learns nonlinear mappings (both networks) of two modalities that are maximally correlated. CCA is a statistical method for dealing with the correlation between sets of two variables, finding linear projection vectors. Unlike CCA, DCCA is a parametric method, and it can learn the complex transformations of two views. DCCA has been applied to several audio classification tasks [12, 13], and improved [14]. The DCCA objective function is optimized in an unsupervised manner over the actual data; therefore, it is not necessary to use some wrong phone labels for training networks.

In most multimodal speech recognition systems, audio and visual features are integrated by just concatenating these features. Because the audio and visual features are intrinsically different, and a gap between audio and visual feature spaces may cause undesirable effects in speech recognition. Applying

DCCA, gaps between two feature spaces are reduced, and we expect to obtain more complementary features for speech recognition. We will show in this paper that our proposed feature can achieve better recognition performance in noisy environments.

The rest of this paper is organized as follows: In Section 2, we review CCA and DCCA. In Section 3, our proposed method is explained. In Section 4, the experimental data are evaluated, and the final section is devoted to our conclusions.

2. Preliminaries

In this section, we review CCA and DCCA, where two views represent the audio and visual features.

2.1. Canonical Correlation Analysis

Let $\mathrm{X}_{audio} \in \mathbb{R}^{d_1 \times N}$, $\mathrm{X}_{visual} \in \mathbb{R}^{d_2 \times N}$ denote audio and visual features with N samples where the sample mean of these matrices is normalized to zero, and d_1 and d_2 represent the dimension of the audio and visual features, respectively. In CCA, the correlation coefficient is calculated as follows:

$$\begin{aligned} \rho(\boldsymbol{a}, \boldsymbol{b}) &= \mathrm{corr}(\boldsymbol{a}^\top \mathrm{X}_{audio}, \boldsymbol{b}^\top \mathrm{X}_{visual}) && (1) \\ &= \frac{\boldsymbol{a}^\top \Sigma_{av} \boldsymbol{b}}{\sqrt{\boldsymbol{a}^\top \Sigma_{aa} \boldsymbol{a}} \sqrt{\boldsymbol{b}^\top \Sigma_{vv} \boldsymbol{b}}}, && (2) \end{aligned}$$

where $\boldsymbol{a} \in \mathbb{R}^{d_1}$, $\boldsymbol{b} \in \mathbb{R}^{d_2}$ are the projection vectors, which are parameters of CCA, and $\Sigma_{av} \in \mathbb{R}^{d_1 \times d_2}$, $\Sigma_{aa} \in \mathbb{R}^{d_1 \times d_1}$, $\Sigma_{vv} \in \mathbb{R}^{d_2 \times d_2}$ are the cross-covariance matrices of X_{audio} and X_{visual}, the covariance matrix of X_{audio} and X_{visual}, respectively. Since $\rho(\boldsymbol{a}, \boldsymbol{b})$ is invariant to scalling of $\boldsymbol{a}$ and $\boldsymbol{b}$, we assume that each standard variance of denominator in Eq. (2) has one; that is the projections are constrained to have unit variance,

$$\max_{\boldsymbol{a}, \boldsymbol{b}} \boldsymbol{a}^\top \Sigma_{av} \boldsymbol{b} \text{ subject to } \boldsymbol{a}^\top \Sigma_{aa} \boldsymbol{a} = \boldsymbol{b}^\top \Sigma_{vv} \boldsymbol{b} = 1 \qquad (3)$$

If we use $L \leq min(d_1, d_2)$ pairs of linear projection vectors, the projection matrices for audio and visual features are formed as $\mathrm{U} \in \mathbb{R}^{d_1 \times L}$ and $\mathrm{V} \in \mathbb{R}^{d_2 \times L}$, respectively. We obtain the following formulation to identify the projection matrices A and B:

$$\begin{aligned} &\text{maximize } \mathrm{tr}(\mathrm{A}^\top \Sigma_{av} \mathrm{B}) && (4) \\ &\text{subject to } \mathrm{A}^\top \Sigma_{aa} \mathrm{A} = \mathrm{B}^\top \Sigma_{vv} \mathrm{B} = \mathrm{I}, \end{aligned}$$

where $\mathrm{tr}(\cdot)$ and I indicate the sum of the elements on the main diagonal and the unit matrix, respectively.

The optimal objective value is the sum of the top k singular values of $\mathrm{T} = \Sigma_{aa}^{-1/2} \Sigma_{av} \Sigma_{vv}^{-1/2}$. The optimal projection matrices are given by $(\mathrm{A}, \mathrm{B}) = (\Sigma_{aa}^{-1/2} \mathrm{U}_k, \Sigma_{vv}^{-1/2} \mathrm{V}_k)$, where $\mathrm{U}_k \in \mathbb{R}^{d_1 \times k}$ and $\mathrm{V}_k \in \mathbb{R}^{d_2 \times k}$ are the first k left- and right-singular vectors of T. Indeed, the covariance matrices Σ_{aa} and Σ_{vv} are estimated from data using regularization so that they are constrained to the nonsingular matrix.

2.2. Deep Canonical Correlation Analysis

DCCA computes the representations of the two views by passing them through multiple stacked layers of nonlinear transformation. Given the audio and visual features (X_{audio},X_{visual}), the outputs of the audio and visual neural networks are written as $f(\mathrm{X}_{audio}; \theta_1) \in \mathbb{R}^{o \times N}$, $f(\mathrm{X}_{visual}; \theta_2) \in \mathbb{R}^{o \times N}$, respectively. θ_1, θ_2 indicate parameters of the audio and visual networks, respectively. DCCA computes the total correlation as follows:

$$corr(\boldsymbol{a}^\top f(\mathrm{X}_{audio}; \theta_1), \boldsymbol{b}^\top f(\mathrm{X}_{visual}; \theta_2)) = \mathrm{tr}(\mathrm{T}^\top \mathrm{T})^{\frac{1}{2}}, \qquad (5)$$

where $\mathrm{T} = \hat{\Sigma}_{aa}^{-1/2} \hat{\Sigma}_{av} \hat{\Sigma}_{vv}^{-1/2}$ as reviewed in section 2.1. $\hat{\Sigma}_{av} = \frac{1}{N-1} \mathrm{X}_{audio} \mathrm{X}_{visual}^\top$ and $\hat{\Sigma}_{aa} = \frac{1}{N-1} \mathrm{X}_{audio} \mathrm{X}_{audio}^\top + r_1 \mathrm{I}$ are the covarince matrices with regularization constant $r_1 > 0$, similarly for $\hat{\Sigma}_{vv}$. The goal of DCCA is to jointly learn parameters $\{\theta_1, \theta_2, \boldsymbol{u}, \boldsymbol{v}\}$ for both views, such that the correlation is as high as possible. The parameters $\{\theta_1, \theta_2\}$ are trained using back-propagation. The gradient of Eq. 5 can be computed as follows:

$$\begin{aligned} &\frac{\partial corr(\boldsymbol{a}^\top f(\mathrm{X}_{audio}; \theta_1), \boldsymbol{b}^\top f(\mathrm{X}_{visual}; \theta_2))}{\partial f(\mathrm{X}_{audio}; \theta_1)} \\ &= \frac{1}{N-1} (2 \nabla_{aa} \mathrm{X}_{audio} + \nabla_{av} \mathrm{X}_{visual}) \qquad (6) \end{aligned}$$

where $\nabla_{ab} = \hat{\Sigma}_{aa}^{-1/2} \mathrm{U} \mathrm{V}^\top \hat{\Sigma}_{vv}^{-1/2}$ and $\nabla_{aa} = -\frac{1}{2} \hat{\Sigma}_{aa}^{-1/2} \mathrm{U} \mathrm{D} \mathrm{V}^\top \hat{\Sigma}_{aa}^{-1/2}$, and the derivative with respect to $f(\mathrm{X}_{visual}; \theta_2)$ has a symmetric expression.

General DNN objective functions are written as the expectation (or sum) of error functions (e.g., squared loss) calculated for each training sample. This property naturally suggests stochastic gradient descent (SGD) for optimization, where gradients are estimated for a few training examples (a mini-batch) and iteratively updated parameters. However, in DCCA (Eq. 5), it is necessary to estimate the covariance matrices for the training samples. Andrew et al. [11] used a full-batch algorithm (L-BFGS) for optimization. This is undesirable for applications with large training sets, as each gradient step computed on the entire training set can be very expensive in both memory and time. To mitigate this problem, Wang et al. [12] showed that it works well, even for this type of objective, if larger mini-batches are used. It is considered that a large mini-batch has enough information to estimate covariances. Hence, in this paper, we also configure a larger mini-batch size.

3. Related Works

Deep learning has had recent successes for acoustic modeling [15]. 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.* [16] proposed multimodal DNNs that learn features over audio and visual modalities. Mroueh *et al.* [17] improved this method and proposed an architecture considering the correlations between modalities. Ninomiya *et al.* [6] investigated integration of bottleneck features using multi-stream hidden Markov models (HMMs) for audio-visual speech recognition.

CNNs also have demonstrated impressive performance on several tasks, such as image analysis [18, 19, 20] and spoken language [21] and music recognition [22]. In our previous work [8], we showed that the features extracted from CNNs lead to effective results for speech recognition thanks to the properties of the local receptive field and the shift invariant. Therefore, in this paper we do not use DNNs, but CNNs, for nonlinear mappings of two modalities.

Recently, multimodal learning has been researched in relation to discovering useful information about the world. If such methodology can be used to develop an accurate system, we would be able to obtain non-verbal information that cannot, at

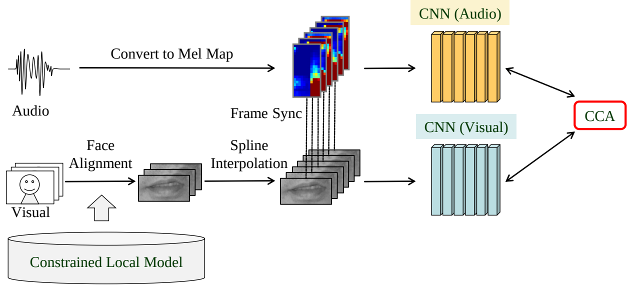

Figure 1: *Deep CCA using CNNs*

this time, be explained expressly and cannot be obtained from discriminative models. Unsupervised learning is an approach to can be used to handle this problem. For multimodal fusion tasks, several approaches have been proposed [23, 24] where modalities are modeled using a generative model that is based on an RBM. For speech recognition tasks, the reproducibility of the input data is not necessary due to the fact that a DCCA approach is concise. In this paper, we employ DCCA with CNNs as a robust feature extractor for the fluctuation of the speech uttered by a person with cerebral palsy.

4. Multimodal Feature Extraction Using DCCA

4.1. Flow of the Proposed Method

Figure 1 shows the flow of our proposed feature extraction. To employ advantages of our previous work [8], we use CNNs for the mappings of DCCA instead of DNNs. Hereafter, $f(\cdot;\theta)$ in section 2.2 indicates a CNN operation where the input is two-dimensional.

First, we prepare the input features for training a CNN 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). Then, a lip image is extracted, and 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 spline interpolation to the lip images.

The parameters of audio and visual CNNs are jointly learned by back-propagation with SGD where the gradients are calculated by DCCA objective function, starting from random values. Following the training of both CNNs, the input mel-map and lip images are transformed to the output units through each CNN, and projected linearly as follows:

$$\boldsymbol{\alpha}_t = \hat{\Sigma}_{aa}^{-1/2}\mathrm{U}_k f(\mathrm{X}_t;\theta_1) \quad (7)$$

$$\boldsymbol{\beta}_t = \hat{\Sigma}_{vv}^{-1/2}\mathrm{V}_k f(\mathrm{Y}_t;\theta_2), \quad (8)$$

where $(\mathrm{X}_t, \mathrm{Y}_t)$ are two-dimensional input feature for audio and visual at time t, and ($\boldsymbol{\alpha}_t \in \mathbb{R}^k$, $\boldsymbol{\beta}_t \in \mathbb{R}^k$) are the corresponding features, respectively. Then these features are concatenated, and $[\boldsymbol{\alpha}_t^\top\ \boldsymbol{\beta}_t^\top]^\top \in \mathbb{R}^{2k}$ is used as the feature in the training of HMMs for speech recognition.

4.2. Application to Speech Uttered by a Person with Hearing Loss

DCCA has an advantage for speech uttered by a person with hearing loss. In the case of an articulation disorder, the phone label estimated by forced alignment may not be correct. However, several approaches based on DNN use the phone label as the target class to learn parameters. The DCCA accomplishes the training procedure in an unsupervised fashion to find the maximal correlation between two sets of modalities. Therefore, the feature extracted from networks trained by DCCA is not influenced by some wrong phone labels. By using DCCA, audio and visual features are transformed through networks so that output units have a high correlation. In noisy environments, we expect that even if the audio feature is degraded, the transformed feature has adequate robustness because this feature has a high correlation to the visual feature which is noise-invariant.

5. Experiments

5.1. Recognition Results Using a Speaker-independent Acoustic Model

At the beginning, we attempted to recognize utterances using a speaker-independent acoustic model for unimpaired people (This model is included in Julius [25]). The acoustic model consists of a triphone HMM set with 25-dimensional MFCC features (12-order MFCCs, their delta and energy) and 16 mixture components for each state. Each HMM has three states and three self-loops. For a person with hearing loss, a recognition rate of only 3.24% was obtained, but for a physically-unimpaired person, a recognition rate of 88.89% was obtained for the same task. It is clear that the speaking style of a person with hearing loss differs considerably from that of a physically-unimpaired person. Therefore, it is considered that a speaker-dependent acoustic model is necessary for recognizing speech from a person with hearing loss.

5.2. Word Recognition Experiments

5.2.1. *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 [26]. 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. We compare our audio-visual feature with conventional MFCC+Δ+ΔΔ (36-dimensions) and MFCC+Δ+ΔΔ+ discrete cosine transform (DCT) (66-dimensions). Then, our proposed method and audio-visual features were evaluated in noisy environments. White noise was added to audio signals and their SNR is set to 20dB, 10dB, and 5dB. Audio CNN and HMMs are trained using the clean audio feature.

5.2.2. *Architecture of the Networks*

We construct deep networks, which consist of a convolution layer, a pooling layer, and fully-connected MLPs. For the input layer of audio CNN, we use a mel-map of subsequent 13-frames with 39-dimensional mel spectrum, and the frame shift is 5 msec. For the input layer of visual CNN, frontal face videos are recorded at 60 fps. Luminance images are extracted from

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
Audio CNN	39×13	4×2	3×3	13	108, 30, 108
Visual CNN	12×24	5×5	2×2	13	108, 30, 108

the image using CLM and resized to 12×24 pixels. Finally, the images are up-sampled by spline interpolation and input to the CNN.

Table 1 shows parameters used in experiments. We set the bottleneck layer into networks to investigate the performance of bottleneck features. In the training procedure, a learning rate and a momentum are set to be 0.0001 and 0.99, respectively.

5.2.3. Number of Mini-batch Sizes

In the preliminary experiment, we compared the effects of changing the number of mini-batches with 50 epochs in a clean environment. Table 2 shows the results when changing the number of mini-batches as 1,200, 1,500, 1,800, 2,100 and 2,400. Through the experiments, we found that the performance improves as the number of mini-batches increased. The reason for the improvement is attributed to being able to estimate the covariance matrix more accurately when using larger mini-batch sizes. In the future experiments, we will use a mini-batch size of 2,100.

Table 2: *Word recognition accuracy for each mini-batch size*

# of mini-batches	1,200	1,500	1,800	2,100	2,400
Recognition accuracy [%]	63.89	65.28	66.20	71.76	71.76

5.2.4. Results and Discussion

Figure 2 shows the word recognition accuracies in noisy environments. We compared the audio-visual feature extracted from our proposed method with two conventional features: MFCC+Δ+ΔΔ (MFCC), MFCC+Δ+ΔΔ+DCT (MFCC+DCT). In Figure 2, DCCA and DCCA bottleneck denote the features extracted from the final projection layer and the bottleneck layer (30-dimensions). Comparing DCCA bottleneck with DCCA, the former shows better accuracies. This is because the information that the audio feature has is lost when it is transformed to near the visual space. The DCCA bottleneck feature is better than MFCC+DCT in SNR of 10dB. This might be because the DCCA bottleneck feature obtained more noise-robustness compared with the conventional feature. These results show our proposed method improves performance.

Figure 3 shows the word recognition accuracies comparing our proposed method with the previous method [8]. The DCCA framework is the unsupervised learning that is applied to the actual data in order to find the maximal correlation between two sets of modalities without other information. Therefore, the extracted feature might not be able to present the phonological information. Our previous work employed supervised learning using the phone labels. In our experiments, the accuracy of the DCCA bottleneck degraded on average 14% compared to using supervised learning.

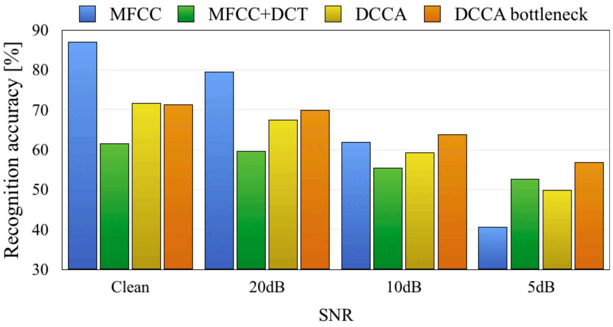

Figure 2: *Word recognition accuracy using HMMs*

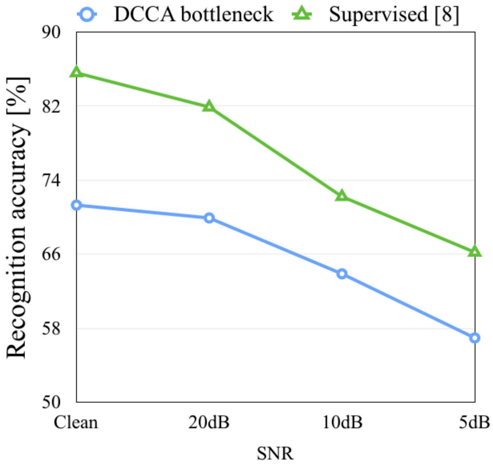

Figure 3: *Word recognition accuracy of unsupervised and supervised training procedure*

6. Conclusions

In this paper, we discussed an audio-visual speech recognition system for a person with an articulation disorder resulting from severe hearing loss based on CNNs. We proposed a feature extraction method using CNNs trained by deep CCA which is optimized in an unsupervised manner. In the DCCA procedure, audio and visual CNNs are trained maximizing the correlation between the units of each output layer. When a noisy input signal is fed to CNNs, although the audio feature is degraded, the visual feature compensates for the degraded audio feature data to increase accuracy. Then, the degradation of accuracy is restrained in high-noisy environments.

In comparison, in experiments between the proposed feature and a conventional unsupervised feature (MFFC+DCT), the proposed feature showed better performances than the conventional one. The improvement was more significant in high-noisy environments. However, the performance of the proposed method was lower than that of the supervised method. This result suggests that using DCCA, the phonological information are not necessarily extracted.

In future work, we will further investigate a better DCCA-based feature extraction which is also highly correlated to the phonological information. Person with an articulation disorder resulting from severe hearing loss need the various applications for communication, for example, voice-to-signal conversion system. Although their speech style is so different from that of people without hearing loss, they can make appropriate lip shapes. Therefore, the voice-to-signal conversion system is able to help the interaction with others. We will also research this system in future work.

7. Acknowledgements

This work was supported in part by PRESTO, JST (Grant No. JPMJPR15D2).

8. 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] S. Najnin, B. Banerjee, L. L. Mendel, M. H. Kapourchali, J. K. Dutta, S. Lee, C. Patro, and M. Pousson, “Identifying hearing loss from learned speech kernels,” in *INTERSPEECH*, 2016, pp. 243–247.

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

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

[7] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, “Lipnet: Sentence-level lipreading,” *CoRR*, vol. abs/1611.01599, 2016.

[8] Y. Takashima, R. Aihara, T. Takiguchi, Y. Ariki, N. Mitani, K. Omori, and K. Nakazono, “Audio-visual speech recognition using bimodal-trained bottleneck features for a person with severe hearing loss,” in *INTERSPEECH*, 2016, pp. 277–281.

[9] 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

[10] 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.

[11] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, “Deep canonical correlation analysis,” in *ICML*, 2013, pp. 1247–1255.

[12] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, “Unsupervised learning of acoustic features via deep canonical correlation analysis,” in *ICASSP*, 2015, pp. 4590–4594.

[13] N. E.-D. El-Madany, Y. He, and L. Guan, “Multiview learning via deep discriminative canonical correlation analysis,” in *ICASSP*, 2016, pp. 2409–2413.

[14] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, “On deep multi-view representation learning: Objectives and optimization,” *CoRR*, vol. abs/1602.01024, 2016.

[15] 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.

[16] 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.

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

[18] 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.

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

[20] 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.

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

[22] T. Nakashika, C. Garcia, and T. Takiguchi, “Local-feature-map integration using convolutional neural networks for music genre classification,” in *INTERSPEECH*, 2012.

[23] N. Srivastava and R. Salakhutdinov, “Multimodal learning with deep boltzmann machines,” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2949–2980, 2014.

[24] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in *ICML*, 2011, pp. 689–696.

[25] “Open-Source Speech Recognition Software Julius,” http://julius.sourceforge.jp/.

[26] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, “ATR Japanese speech database as a tool of speech recognition and synthesis,” *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.