Lipreading using Convolutional Neural Network

Kuniaki Noda[1], *Yuki Yamaguchi*[2], *Kazuhiro Nakadai*[3], *Hiroshi G. Okuno*[2], *Tetsuya Ogata*[1]

[1]Graduate School of Fundamental Science and Engineering, Waseda University, Tokyo, Japan
[2]Graduate School of Informatics, Kyoto University, Kyoto, Japan
[3]Honda Research Institute Japan Co., Ltd., Saitama, Japan

kuniaki.noda@akane.waseda.jp, yamaguch@kuis.kyoto-u.ac.jp,
nakadai@jp.honda-ri.com, okuno@kuis.kyoto-u.ac.jp, ogata@waseda.jp

Abstract

In recent automatic speech recognition studies, deep learning architecture applications for acoustic modeling have eclipsed conventional sound features such as Mel-frequency cepstral coefficients. However, for visual speech recognition (VSR) studies, handcrafted visual feature extraction mechanisms are still widely utilized. In this paper, we propose to apply a convolutional neural network (CNN) as a visual feature extraction mechanism for VSR. By training a CNN with images of a speaker's mouth area in combination with phoneme labels, the CNN acquires multiple convolutional filters, used to extract visual features essential for recognizing phonemes. Further, by modeling the temporal dependencies of the generated phoneme label sequences, a hidden Markov model in our proposed system recognizes multiple isolated words. Our proposed system is evaluated on an audio-visual speech dataset comprising 300 Japanese words with six different speakers. The evaluation results of our isolated word recognition experiment demonstrate that the visual features acquired by the CNN significantly outperform those acquired by conventional dimensionality compression approaches, including principal component analysis.

Index Terms: Lipreading, Visual Feature Extraction, Convolutional Neural Network

1. Introduction

Incorporation of speakers' lip movements as visual information for automatic speech recognition (ASR) systems is known to contribute to robustness and accuracy, especially in environments where audio information is corrupted by noise, provided use of suitable visual features. In previous studies, several different approaches have been proposed for extracting visual features from the region-of-interest [1, 2], which broadly fall into two representative categories. The first approach is a top-down approach, where an a priori lip shape representation framework is embedded in a model; for e.g., active shape models (ASMs) [3] and active appearance models (AAMs) [4]. ASMs and AAMs extract higher-level, model-based features derived from the shape and appearance of the mouth area images. The second approach is a bottom-up approach, where visual features are directly estimated from the image; for e.g., several dimensionality compression algorithms such as discrete cosine transform [5, 6], principal component analysis (PCA) [5, 7], and discrete wavelet transform [5]. They are commonly utilized for extracting lower-level, image-based features. The former, model-based features are suitable for explicitly analyzing internal representations; however, some elaboration of lip shape models and precise hand-labeled training data is required for constructing a statistical model that represents valid lip shapes. While the latter, image-based features are advantageous because they do not require dedicated lip shape models or hand-labeled data for training, however, they are vulnerable to changes in lighting conditions, translations, or rotations of input images. The main objective of this study is to adopt the bottom-up approach for a lipreading system that overcomes the weaknesses of image-based feature extraction mechanisms. This is achieved by introducing a deep learning approach.

In the machine learning community [8], deep learning approaches have been successfully applied to feature learning for various modalities, such as images [9] and audio [10]. Achievements in deep learning technologies have facilitated in making advanced applications available to the public. For example, competition results from ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [11] have led to significant improvements in web image search engines [12]. The following three factors have contributed to recent progress in the application of neural networks to the problems of image classification and speech recognition systems. First, the popularization of low-cost, high performance computational environments: high-end consumer personal computers equipped with general-purpose graphics processing units (GPGPUs) have allowed a wider range of users to conduct brute-force numerical computation with large datasets. Second, the public accessibility improvement of large databases: these have enabled unsupervised learning mechanisms to self-organize highly generalized features that can outperform conventional handcrafted features. Third, the invention of powerful machine learning techniques: recent improvements in optimization algorithms have enabled large-scale neural-network models to be efficiently trained with large datasets, which has made possible deep neural networks generating highly generalized features.

In this paper, we propose a novel visual speech recognition (VSR) system based on a deep learning approach. Specifically, we propose to apply a convolutional neural network (CNN), one of the most successfully utilized neural-network architectures for image clustering problems, as the visual feature extraction mechanism for a lipreading system. This is achieved by training the CNN with over a hundred thousand mouth area image frames in combination with corresponding phoneme labels as the target. The CNN parameters are learned in order to maximize the average across training cases of log-probability of the correct label under the prediction distribution. Through supervised training, multiple layers of convolutional filters, required for extracting primitive visual features, and predicting phonemes from the raw image inputs are acquired. The two main advantages of our approach are as follows: (1) our proposed model is easy to implement because dedicated lip shape models or hand-labeled data are not required. (2) CNN guarantees shift- and rotation- resistant image recognition. Our proposed mechanism is tested on 40 kinds of phoneme recognition evaluation experiments and attains a 58% recognition rate. Furthermore, the outputs of the CNN are regarded as visual feature sequences, and a hidden Markov model with Gaussian mixture observation model (GMM-HMM) is applied for an isolated word recognition task. By comparing its word recognition performance with the results utilizing other image-based visual feature extraction mechanisms, the advantages of CNN are demonstrated.

2. Convolutional neural network

The CNN is a variant of an artificial neural network commonly utilized for image classification problems [13, 14, 15]. CNNs integrate three architectural ideas to ensure spatial invariance: local receptive fields, shared weights, and spatial subsamplings. Accordingly, CNNs are advantageous compared to ordinary fully connected feed-forward networks in the following three ways.

First, the local receptive fields in the convolutional layers extract local image features by connecting each unit to just small local regions of an input image. Local receptive fields can extract elementary visual features, such as oriented-edges, end-points, and corners. By nature, nearby pixels are highly correlated and faraway pixels are weakly correlated. Thus, the stack of convolutional layers is structurally advantageous for recognizing images by effectively extracting and combining the acquired features. Second, CNNs can guarantee some degree of spatial invariance with respect to shift, scale, or local distortion of inputs by forcing the sharing of same weight configurations across the input space. Units in a plane are forced to perform the same operation on different parts of the image. As CNNs are equipped with several local receptive fields, multiple features are extracted at each location. In principle, fully-connected networks are also able to perform similar invariances; however, learning such weight configurations requires a very large number of training datasets to cover all possible variations. Third, sub-sampling layers, which perform local averaging and sub-sampling, are utilized to reduce the resolution of the feature map and the sensitivity of the output to input shifts and distortions (for more details on concrete implementation, see [13]).

In terms of computational scalability, shared weights allow CNNs to possess fewer connections and parameters, compared to standard feed-forward neural networks with similar-sized layers. Moreover, current improvements in computational resources availability, especially with highly-optimized implementations of 2D convolution algorithms processed with GPGPUs, has facilitated the efficient training of remarkably large CNNs with millions of image datasets [9, 11].

Regarding application of CNNs for speech recognition studies, several approaches have been proposed. Abdel-Hamid et al. [16] applied their original functionally-extended CNNs for sound spectrogram inputs, and showed that their CNN architectures outperform earlier basic forms of fully-connected, deep neural networks on phone recognition and large vocabulary speech recognition tasks. Palaz et al. [17] applied a CNN for phoneme sequence recognition by estimating phoneme class conditional probabilities from raw speech signal inputs, yielding comparable or better phoneme recognition performances, compared to conventional approaches. Thus, CNNs are attracting increasing attention in speech recognition studies; however, their applications are limited to audio signal processing, and their applications towards lipreading remain unaddressed.

3. VSR system overview

Figure 1 shows the schematic diagram of our VSR system. Our proposed system comprises a CNN and a GMM-HMM for visual feature extraction and isolated word recognition, respectively. For visual feature extraction, a seven-layered CNN is utilized to recognize phonemes from the mouth area image sequences. The CNN models nonlinear mappings from the raw grayscale image inputs to the corresponding posterior probability distribution of the phoneme labels. The time-series acquired from these outputs are regarded as the visual features for lipreading. Then, by processing the acquired visual feature sequences, left-to-right GMM-HMMs are utilized for isolated word recognition.

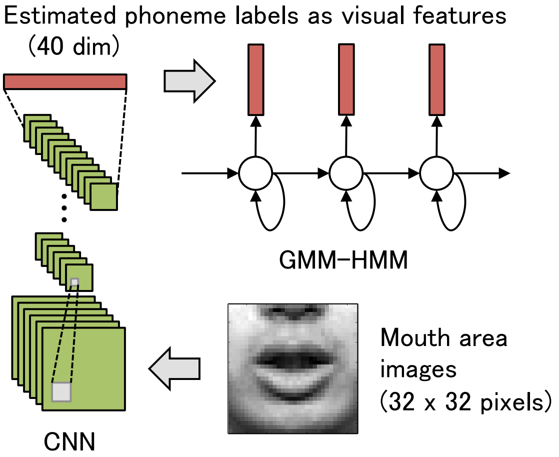

Figure 1: Architecture of our VSR system

4. The dataset

A Japanese audiovisual dataset [18] was used for the evaluation of our proposed model. In the dataset, six male's speech data with 300 words (216 phonetically-balanced words, and 84 important words, from the ATR speech database [19]) were used. Audio data was recorded with a 16 kHz sampling rate, 16-bit depth, and a single channel. For training the acoustic model utilized for assigning phoneme labels to the image sequences, we extracted 39 dimensions of audio features, composed of 13 Mel-frequency cepstral coefficients (MFCCs) and their first and second temporal derivatives. To synchronize the acquired features between audio and video, the MFCCs were sampled at 100 Hz. Additionally, to exactly match the numbers of audio and video frames, video frames were resampled with reference to the number of MFCC frames. Visual data was a full-frontal 640×480 pixels facial view with 8 bit monochrome, recorded at 100 Hz. For visual model training and evaluation, we prepared a trimmed dataset by manually cropping 128×128 pixels of mouth area from the original visual data and resizing them to 32×32 pixels. In the first (visual feature extraction) experiment, image sequences from all 300 words were used. In the second (isolated word recognition) evaluation experiment, 216 phonetically-balanced words and the residual 84 words were used as the training data and the test data, respectively.

Table 1: Construction of convolutional neural network

INPUT DIM*	OUTPUT DIM*	LAYERS*
1024	40	C1-P2-C3-P4-C5-P6-F7**

* INPUT DIM, OUTPUT DIM, and LAYERS give the input dimensions, the output dimensions, and the layer-wise construction of the network, respectively.

** C, P, and F stand for the layer types corresponding to the convolutional layer, the local-pooling layer, and the fully-connected layer, respectively. The numbers after the layer types represent layer indices.

5. Visual feature extraction by CNN

To assign phoneme labels to each frame of mouth area image sequences, we trained a set of monophone HMMs, one for each phoneme, with the MFCCs utilizing the Hidden Markov Model Toolkit (HTK) [20] and assigned 40 phoneme labels including short pause by conducting a forced alignment function of the HVite command of the HTK. For the network training and evaluation, the mouth area images of six speakers' 300 words (approximately 1.23×10^5 samples) were used. The image data were shuffled and 5/6 of the data were used for training, while the rest was used for evaluation. From our preliminary experiment, we confirmed that phoneme recognition precision degrades if images from all six speakers are modeled with a single CNN.[1] Therefore, we prepared an independent CNN for each speaker. The higher-level visual features (phoneme label posterior probabilities) for the further isolated word recognition experiment were generated by recording the neuronal outputs from the last layer of the CNN when the mouth area image sequences corresponding to the 216 training words were provided as inputs to the CNN.

A seven-layered CNN is used in reference to the work by Krizhevsky et al. [11]. Table 1 summarizes the construction of the network containing four weighted layers (approximately 1.19×10^5 parameters): three convolutional (C1, C3, and C5) and one fully-connected (F7). The first convolutional layer (C1) filters the 32×32 pixels input image with 32 kernels of 5×5 pixels with a stride of one pixel. The second and the third convolutional layers (C3 and C5) take the response-normalized and pooled output of the previous convolutional layers (P2 and P4) as inputs, and filter them with 32 and 64 filters of 5×5 pixels, respectively. The fully-connected layer (F7) takes the pooled output of the previous convolutional layer (P6) as input and outputs a 40-way soft-max, regarded as a posterior probability distribution over the 40 classes of phoneme label. Response-normalization layers follow the first and second convolutional layers. A max-pooling layer follows the first response-normalization layer. Average-pooling layers follow the second response-normalization layer, as well as the third convolutional layer. The Rectified linear unit non-linearity is applied to the output of every convolutional layer.

Our network maximizes the multinominl logistic regression objective, which is equivalent to maximizing the average across training cases of log-probability of the correct label under the prediction distribution. The parameters for the network structures are empirically determined in reference to previous studies [11, 13]. We utilized an open-source software "cuda-convnet" [11] for our visual feature extraction experiment on a consumer-class personal computer with an Intel Core i7-3930K processor (3.2 GHz, 6 cores), 32 GB RAM, and a single Nvidia GeForce GTX Titan graphic processing unit with 6 GB on-board graphics memory.

6. Results

6.1. Phoneme recognition

After training the CNN, the phoneme recognition performance is evaluated by recording the neuronal outputs from the last layer of the CNN when the mouth area image sequences corresponding to the test image data are provided as the inputs to the CNN. The average phoneme recognition performance for the 40 phonemes over six speakers is 58%. Figure 2 shows the mean and the standard deviation of phoneme-wise recognition rate from six different speakers. This result demonstrates that the visual phoneme recognition works better for recognizing vowels than consonants. The result derives from the fact that, the mean recognition rate for all vowels is 60-100%, whereas the mean recognition rate for all other phonemes is 20-80%. This result may be attributed to the fact that the generation of vowels is strongly correlated with visual cues represented by the movements of lips or jaw [21, 22].

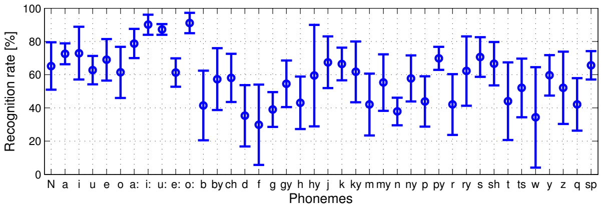

Figure 2: Phoneme recognition rate

Figure 3 shows the confusion matrix of the phoneme recognition evaluation result. It should be noted that the wrongly recognized consonants are classified as vowels in most cases. This indicates that the articulation of consonants is mostly attributed not only to the motion of the lips but also to the dynamic interaction of interior oral structures, such as tongue, teeth, oral cavity, etc., invisible from frontal facial images.

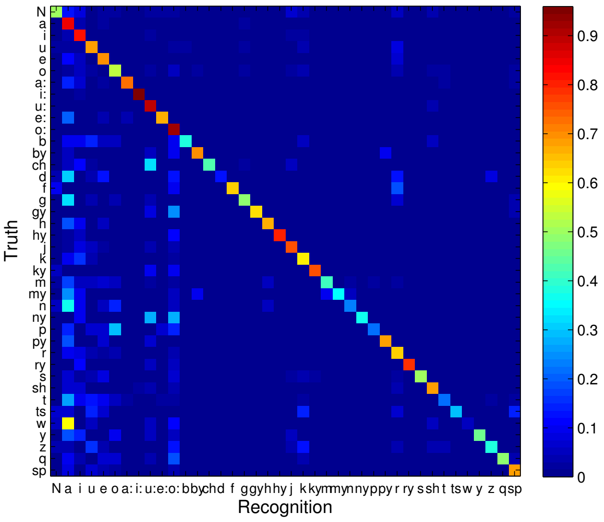

Figure 3: Phoneme recognition confusion matrix

[1] We think this degradation is mainly due to the limited variations of lip region images we prepared for the training of CNN. To generalize the higher-level visual features that enable a CNN to attain speaker invariant phoneme recognition, we consider more image samples from different speakers are needed.

Visually explicit phonemes, such as bilabial consonants (e.g., /m/, /p/, or /b/), are expected to be relatively well discriminated by a VSR system. However, recognition performances is not as high as it was expected to be in our results. To improve the recognition rate, the procedure to obtain phoneme target labels for CNN training should be improved. In general pronunciation, consonant sounds are shorter than vowel sounds, therefore, the labeling for consonants is more time critical than vowels. In addition, consonant label accuracy directly affects recognition performance because the number of training samples for consonants is much smaller than for vowels.

6.2. Visual feature space analysis

To analyze how the acquired visual feature space is structured, the trained CNN is used to generate phoneme posterior probability sequences from test image sequences. The acquired sequences are processed by PCA, and the first three principal components are extracted to visualize the feature space. Figure 4 shows the visual feature space corresponding to the five representative Japanese vowel phonemes, /a/, /i/, /u/, /e/, and /o/. As demonstrated in the graph, raw mouth area images corresponding to the five vowel phonemes are clearly discriminated by the CNN, and the clusters corresponding to the phonemes are structured in the visual feature space. This result indicates that the acquired phoneme posterior probability sequences can be effectively utilized as visual feature sequences for the further isolated word recognition task.

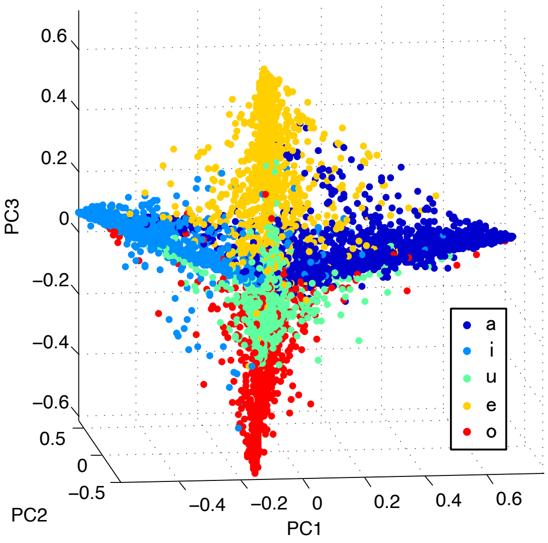

Figure 4: Visual feature space for the five representative Japanese vowel phonemes

6.3. Isolated word recognition

The acquired visual features are evaluated by conducting an isolated word recognition experiment. To recognize words from the phoneme label sequences generated by the CNN, monophone HMMs with 8, 16, and 32 GMM components are utilized. Evaluation is conducted with the 84 test words from the same speaker, yielding a closed-speaker and open-vocabulary evaluation. To compare with the baseline performance, two other visual features with similar dimensionalities are prepared. One feature has 36 dimensions, generated by simply rescaling the images to 6×6 pixels, and the other feature has 40 dimensions, generated by compressing the raw images by PCA.

Figure 5 shows the word recognition rates acquired from nine different models with a combination of three kinds of visual feature and three kinds of component for the GMMs. These results demonstrate that the visual features acquired by the CNN attain higher word recognition capabilities than the other two. The highest word recognition rate was about 37%, and this result shows that visual inputs can be a reliable information source for a speech recognition task when a visual feature extraction mechanism is properly utilized.

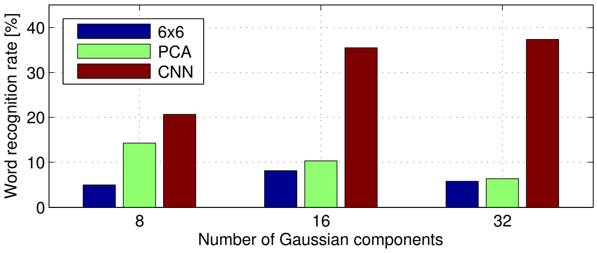

Figure 5: Word recognition rate

7. Conclusions

In this work, we proposed a novel visual feature extraction approach for a VSR system utilizing a CNN. Our experimental results demonstrate that a supervised learning approach to recognize phonemes from raw mouth area image sequences could discriminate 40 phonemes by six speakers at 58% recognition accuracy. Moreover, the acquired phoneme sequences are utilized as a visual feature for an isolated word recognition task, and this significantly outperforms features acquired by other dimensionality compression mechanisms, such as simple image rescaling and PCA. In the current approach, we apply speaker dependent models for phoneme recognition, and a common model for isolated word recognition. The main reason we prepare speaker dependent models is due to significant variations in mouth area appearance, depending on the speaker, and the prepared training dataset is insufficient to acquire a speaker-independent model by covering all possible appearance variations. Considering the generalization ability of a CNN to be successfully utilized for the ILSVRC contest, it has the potential to acquire a speaker independent model for the VSR task. The next step for our future work is to investigate the possibility of building a speaker-independent phoneme recognition model by preparing a larger dataset, increasing the number of speakers, and applying artificial deformation for the image dataset. This research objective can also lead to the fundamental understanding of existing viseme models from a computer science study approach.

8. Acknowledgements

This work has been supported by JST PRESTO “Information Environment and Humans” and MEXT Grant-in-Aid for Scientific Research on Innovative Areas “Constructive Developmental Science” (24119003) and (S) (24220006).

9. References

[1] I. Matthews, T. Cootes, J. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198–213, 2002.

[2] Y. Lan, B.-j. Theobald, R. Harvey, E.-j. Ong, and R. Bowden, "Improving Visual Features for Lip-reading," in *Proceedings of the International Conference on Auditory-Visual Speech Processing*, Hakone, Japan, Oct. 2010.

[3] J. Luettin, N. Thacker, and S. Beet, "Visual speech recognition using active shape models and hidden Markov models," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, no. 95, Atlanta, GA, USA, May 1996, pp. 817–820.

[4] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[5] I. Matthews, C. N. Gerasimos Potamianos, and J. Luettin, "A Comparison of Model and Transform-based Visual Feature for Audio-visual LVCSR," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, Tokyo, Japan, Aug. 2001.

[6] P. Scanlon and R. Reilly, "Feature analysis for automatic speechreading," in *Proceedings of the IEEE 4th Workshop on Multimedia Signal Processing*, Cannes, France, Oct. 2001, pp. 625–630.

[7] P. S. Aleksic and A. K. Katsaggelos, "Comparison of Low- and High-level Visual Features for Audio-visual Continuous Automatic Speech Recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, Montreal, Canada, May 2004, pp. 917–920.

[8] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, Jan. 2009.

[9] Q. V. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Dean, and A. Y. Ng, "Building high-level features using large scale unsupervised learning," in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, Jul. 2012, pp. 81–88.

[10] G. Hinton, L. Deng, D. Yu, G. Dahl, A.Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, pp. 82–97, Nov. 2012.

[11] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, 2012.

[12] Rosenberg, Chuck, "Improving Photo Search: A Step Across the Semantic Gap," http://googleresearch.blogspot.jp/2013/06/improving-photo-search-step-across.html, Jun. 2013.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[14] Y. LeCun and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, Washington, D.C., USA, Jun. 2004, pp. 97–104.

[15] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, Jun. 2009, pp. 609–616.

[16] O. Abdel-Hamid and H. Jiang, "Rapid and Effective Speaker Adaptation of Convolutional Neural Network Based Models for Speech Recognition," in *14th Annual Conference of the International Speech Communication Association*, Lyon, France, Aug. 2013.

[17] D. Palaz, R. Collobert, and M. Magimai.-Doss, "Estimating Phoneme Class Conditional Probabilities from Raw Speech Signal using Convolutional Neural Networks," in *14th Annual Conference of the International Speech Communication Association*, Lyon, France, Aug. 2013.

[18] T. Yoshida, K. Nakadai, and H. G. Okuno, "Automatic speech recognition improved by two-layered audio-visual integration for robot audition," in *Proceedings of the 9th IEEE-RAS International Conference on Humanoid Robots*, Paris, France, Dec. 2009, pp. 604–609.

[19] H. Kuwabara, K. Takeda, Y. Sagisaka, S. Katagiri, S. Morikawa, and T. Watanabe, "Construction of a Large-scale Japanese Speech Database and its Management System," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, Scotland, UK, May 1989, pp. 560–563.

[20] S. Young, G. Evermann, M. Gales, T. Hain, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2009.

[21] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, pp. 23–43, 1998.

[22] J. Barker and F. Berthommier, "Evidence of correlation between acoustic and visual features of speech," in *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, CA, USA, Aug. 1999, pp. 5–9.