T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

Title	Multimodal Speech Recognition Using Mouth Images from Depth Camera
Authors	Yuki Yasui, Nakamasa Inoue, Koji Iwano, Koichi Shinoda
Citation	Proc. APSIPA, , , pp. 1233-1236
Pub. date	2017, 12
DOI	https://doi.org/10.1109/APSIPA.2017.8282227
URL	http://www.ieee.org/index.html
Copyright	(c)2017 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.
Note	このファイルは著者（最終）版です。 This file is author (final) version.

Powered by T2R2 (Science Tokyo Research Repository)

Multimodal Speech Recognition Using Mouth Images from Depth Camera

Yuki Yasui*, Nakamasa Inoue*, Koji Iwano†, Koichi Shinoda*,
* Tokyo Institute of Technology, Japan
E-mail: {yyasui,inoue}@ks.c.titech.ac.jp, shinoda@c.titech.ac.jp
† Tokyo City University, Japan
E-mail: iwano@tcu.ac.jp

***Abstract*—Deep learning has been proved to be effective in multimodal speech recognition using facial frontal images. In this paper, we propose a new deep learning method, a trimodal deep autoencoder, which uses not only audio signals and face images, but also depth images of faces, as the inputs. We collected continuous speech data from 20 speakers with Kinect 2.0 and used them for our evaluation. The experimental results with 10dB SNR showed that our method reduced errors by 30%, from 34.6% to 24.2% from audio-only speech recognition when SNR was 10dB. In particular, it is effective for recognizing some consonants including /k/, /t/.**

I. Introduction

Nowadays automatic speech recognition has been used in various environments, but its accuracy decreases under noisy environment. Many speech enhancement methods robust against ambient noises have been proposed. Their example includes spectral subtraction, VTS, beamforming using multi-microphones. In this paper, we focus on multimodal speech recognition where we use not only audio signals (speech) but also the images of mouths. It is expected to be effective even when the signal-to-noise ratio (SNR) is very low, since images are not affected by acoustic noise [1]. Its main application is car navigation. There have been many studies on multimodal speech recognition. There are two major problems to solve; one is how to extract features and the other is how to fuse the two modes [2], [3], [4].

Most studies have used only frontal face images in the image mode. From the frontal view, the shape of mouths for some phones are quite similar, and thus, it is difficult to discriminate them [5]. Some methods proposed the use of depth images of mouths. Nakamura et, al. [6] reported that it significantly improved the accuracy of consonant recognition. But it was not feasible in real applications since it needed special sensors attached to users' faces. Kumar et, al. [7] utilized images of faces in profile, but their method needed to shoot a face at the same time with multiple cameras, which is difficult in real use. Recently, depth cameras such as Microsoft Kinect has become available with a small cost. They were also used for multimodal speech recognition [8], [9], [10]. In this study, we aim to improve the performance of multimodal speech recognition using depth cameras.

In these few years, deep learning has become popular and significantly improved performance of the conventional methods in various tasks such as image recognition and speech recognition It has also applied to multimodal speech recognition [11], [12], [13]. One interesting example is Deep Autoencoder (DAE). It is a neural network trained to reconstruct inputs in the output layer, and its hidden layer outputs are extracted as a bottle-neck feature. By using both audio signals and images as inputs, it can not only extract features from them but also fuse them in the same network. However, there have been no deep learning researches using depth images until now, to the best of our knowledge. The conventional multimodal speech recognition research with Kinect used multi-stream HMMs for multimodal fusion, and did not employ deep leaning for feature extraction.

In this paper, we propose a deep learning method, Trimodal Deep AutoEncoder (TDAE), for multimodal speech recognition using a depth camera. It compresses audio features, color features, and depth features non-linearly to extract bottle-neck features, which effectively represent comprehensive features of all the modes. We collected continuous speech data from 20 speakers with Kinect 2.0 and used them for our evaluation. The experimental results show that our method outperforms the conventional bimodal approaches in speech recognition accuracy.

This paper is organized as follows. Section 2 introduces some related studies. Section 3 explains the proposed TDAE. Section 4 shows the experimental results. Section 5 concludes this paper.

II. Related Works

Neti et al. [1] proved that multimodal approaches are effective in noisy conditions. Neffian et al. [2] proposed a coupled HMM, a multimodal fusion method to replace popular multi-stream HMMs. Kolossa et al. [3] dealt with a problem that even visual data are often unreliable by computing uncertainties of visual features. Borde et al. [4] proposed a visual feature extraction method based on Zernike moments with principal component analysis (PCA). These methods using 2D face images are effective especially in small vocabulary tasks, but it is difficult to discriminate the mouth shapes for some phones, which are quite similar from the frontal view.

Nakamura et, al. [6] reported that 3D coordinates information significantly improved the accuracy of consonant recognition. But it was not feasible in real applications since it needed special sensors attached to users' faces. Galatas

et al. [8] used Kinect 1.0, a depth camera, for audio-visual speech recognition. They presented a depth feature extraction method based on discrete cosine transform (DCT) of the region of interest (ROI). In their method, audio and visual features are fused by multi-stream HMMs. They improved word error rate in digit speech recognition tasks under noisy condition. Palecek [9] proposed depth-based active appearance model (AAM) features and improved the accuracy over DCT. Wang et al. [10] used the features based on 3D lip points obtained from Kinect. These methods is more suitable for real applications.

Deep learning has recently improved the performance of multimodal speech recognition. It has the advantage of optimizing not only classification, but also feature extraction. There are two methods to apply deep learning for multimodal speech recognition; one uses one deep learning model for feature extraction and feature fusion [11], [12], and the other performs feature extraction independently on each mode [13]. We assume that the first one is better because it is not necessary to manually optimize the stream weights which multi-stream HMMs require.

Ngiam et al. [11] applied a deep learning method to multimodal feature learning. Their proposed model is Bimodal Deep Autoencoder. Autoencoder is a kind of neural networks, which reproduce inputs at the output layer. Autoencoder has one hidden layer, but deep autoencoder has more hidden layers. Since deep autoencoder is trained by unsupervised manner, unlabelled data can be used in training, unlike supervised learning models such as Deep Neural Networks (DNN), Convolutional Neural Networks (CNN). They achieved significant improvements in AVLetters dataset, alphabet speech recognition. Srivastava et al. [12] used another model, deep boltzmann machines (DBM) for multimodal learning. DBM is an unsupervised model also, but it is an undirected graphical one. Tamura et al. [13] trained independently each modal DNN, and then extracted bottleneck features from DNN. These methods showed that deep learning methods significantly improved the conventional performances on multimodal speech recognition, but they have the same problem as the methods without deep learning for 2D images. In addition, deep learning models have so many parameters which we have to tune.

III. Trimodal Deep Autoencoder

To improve the performance of multimodal speech recognition with depth information, we propose a new feature learning method, Trimodal Deep Autoenocoder, which combines audio features, color features and depth features to extract new comprehensive features.

Figure 1 shows our proposed network, in the case that the number of layers is five. Features of each mode are used for inputs, and then they are reconstructed at the output layer. The shared hidden layer represents comprehensive features. The network has a weight parameter $\mathbf{W}_i^{(l)}$ at each connection for $i = 1, 2, 3$ and $l = 1, 2, 3, 4$. Here, we assume that a bias term is included in it. The following presents the learning process of network parameters.

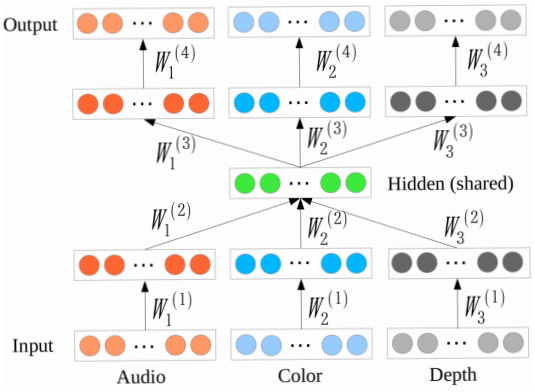


Fig. 1. Trimodal Deep Autoencoder.

Let $\mathbf{x}_i$ be input features for audio ($i = 1$), color ($i = 2$) and depth ($i = 3$). The proposed network is trained so that it reproduces $\mathbf{x}_i$ at the output layer from a corrupted input feature $\tilde{\mathbf{x}}_i = q(\mathbf{x}_i)$, where q is a function for corruption. We employ a function for corruption which randomly sets each element of $\mathbf{x}_i$ to zero with a probability p, in both pre-training and fine-tuning.

In pre-training, since our proposed model is vertically symmetric, parameters are estimated from the outside to the inside, i.e., from the top layer to the middle hidden layer and from the bottom layer to the middle layer. This can be viewed as an extension of greedy layer-wise training [14], in which parameters are estimated from the bottom to the top.

In the first step, parameters $\mathbf{W}_i^{(1)}$ and $\mathbf{W}_i^{(4)}$ are pre-trained for each type of features. By keeping the connections corresponding to these parameters and by omitting the others on the proposed network, we obtain three denoising autoencoders (denoising AEs) as shown in Figure 2 (a). These three denoising AEs are trained by audio, color, and depth features, independently and respectively, with input features $\tilde{\mathbf{x}}_i = q(\mathbf{x}_i)$.

In the second step, parameters $\mathbf{W}_i^{(2)}$ and $\mathbf{W}_i^{(3)}$ are pre-trained by using a denoising AE, which has three inputs, three outputs, and one shared hidden layer corresponding to these parameters, as shown in Figure 2 (d). The input layer accepts $\tilde{\mathbf{h}}_i = q(\mathbf{h}_i)$ as input, where $\mathbf{h}_i$ is an activation vector of the i-th denoising AE trained at the first step:

$$\mathbf{h}_i = \sigma(\mathbf{W}_i^{(1)}\mathbf{x}_i), \tag{1}$$

where σ is an activation function. The next shared hidden layer computes

$$\mathbf{h}' = \sigma(\mathbf{W}_1^{(2)}\mathbf{h}_1 + \mathbf{W}_2^{(2)}\mathbf{h}_2 + \mathbf{W}_3^{(2)}\mathbf{h}_3), \tag{2}$$

and the output layer computes

$$\mathbf{y}_i = \sigma(\mathbf{W}_i^{(3)}\mathbf{h}'), \tag{3}$$

for $i = 1, 2, 3$.

In fine-tuning, all parameters are updated from the pre-trained parameters with a loss function defined by the average

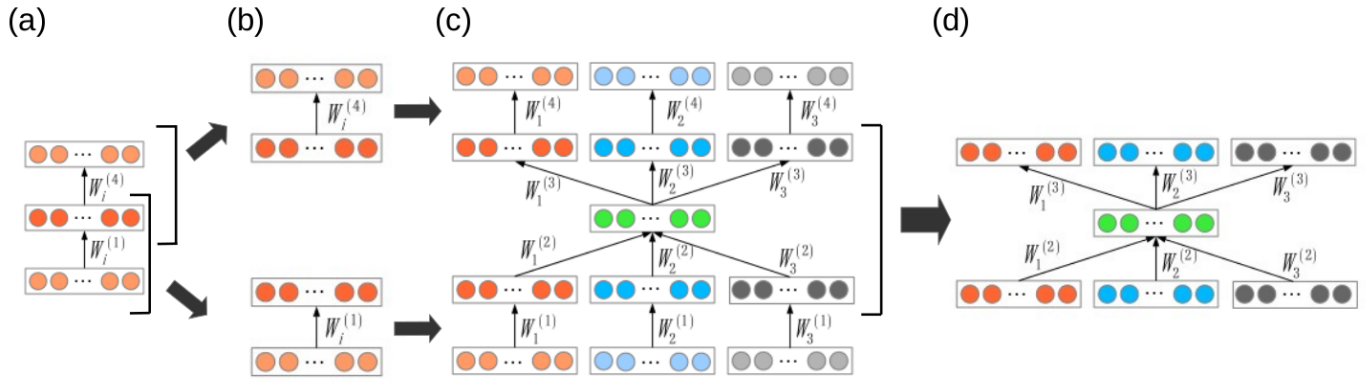


Fig. 2. Pre-training process of Trimodal Deep Autoencoder with five layers. (a) is a denoising autoencoder for the first step. After (a) is trained, it is split into two parts, a decoder part and a encoder part in (b). Next, we insert a shared hidden layer to (b), in (c). We extract middle three layers in (d), and we train this part as a denoising autoencoder including three input layers and three output layers.

of Mean Squared Errors (MSEs) over three types of features. MSE for the i-th feature type is calculated as

$$\mathbf{MSE_i} = \frac{1}{n}\sum_{j=1}^{n}(y_{i,j} - x_{i,j})^2 \quad (4)$$

where n is the dimension of features, $y_{i,j}$ is the j-th element of $\mathbf{y}_i$, and $x_{i,j}$ is the j-th element of $\mathbf{x}_i$.

IV. Experiments

A. Dataset

We collected video data for evaluation of our method. It consists of 15 hours of video with 20 speakers. The recorded sentences are ATR 503 [15], which is often used for Japanese speech recognition research. It has 503 sentences extracted from newspapers, journals, and novels. They are phonetically balanced. The audio data is recorded at 44.1 kHz sampling rate, and downsampled to 16 kHz before feature extraction. The recording place is a student room in a laboratory. Thus, recorded audio data include life noise such as keyboard typing, and coughing.

We use Microsoft Kinect 2.0 for recording video. The frame rate is 30 Hz. The distance between a camera and a speaker is not fixed, but it is more than 500 mm, the minimum distance that Kinect 2.0 can detect. To extract a mouth Region of Interest (ROI), we used *Kinect for windows SDK 2.0*, which is a free software development tool provided by Microsoft. It can track a lot of facial landmarks, and we use the points corresponding to the left corner of the mouth, the right corner of the mouth, the top of the upper lip and the bottom of the lower lip. After extracting mouth regions, the size of mouth images is normalized to 96 × 48.

B. Experimental conditions

The audio features are extracted from 40-dimensional log mel-filter banks with 10 msec frame shift. The color features are 32-dimensional PCA scores of RGB mouth images. The depth features are 32-dimensional PCA scores of depth mouth images. We adjusted the video frame rate to the audio frame rate by three-dimensional spline interpolation. The input/output of Trimodal Deep Autoencoder (TDAE) is composed from the continuous 11 frames features.

TABLE I
The numbers of units in Deep Autoencoder.

DAE	Mode	The number of units of each layer
DAE_audio	Audio	440-200-120-80-120-200-440
DAE_audio_color	Audio	440-200-120-80-120-200-440
	Color	352-200-120-80-120-200-352
DAE_audio_depth	Audio	440-200-120-80-120-200-440
	Depth	352-200-120-80-120-200-352
TDAE	Audio	440-200-120-120-120-200-440
	Color	352-200-120-120-120-200-352
	Depth	352-200-120-120-120-200-352

Table I shows the number of units of the layers corresponding to each mode. The center hidden layer (fourth layer) is the shared layer of each mode, and its outputs form new multimodal features. We use the DNN-HMM based speech recognition frameworks [16]. The unit of the DNN-HMM is triphone. The number of hidden units is 1024, and the number of hidden layers is three. A language model we used is forward 3-gram from the Mainichi newspapers data. The number of vocabulary is 63,465. We evaluate the recognition performance by Word Error Rate (WER) in 4-fold cross validation. The training step takes 12 hours per one fold.

The optimization method of Autoencoder is Adam [17]. The noise addition ratio on Denoising Autoencoder (p) is 0.2, and batch size on mini-batch training is 100. The activation function is sigmoid function. The computational time to train Trimodal Deep Autoencoder is 48 hours by 1 CPU:Intel Xeon X5670 2.93GHx and 1 GPU:NVIDIA Tesla K20X 1.31 Tflops.

For comparison, we evaluate the speech recognition performance when using audio features, consisting of 12-dimensional mel-frequency cepstral coefficients (MFCC) and log-energy, without using Deep Autoencoder. In addition, we conduct experiments using the features obtained from audio only Deep Autoencoder (DAE_audio) and two bimodal Deep Autoencoders integrating audio and color (DAE_audio_color) and audio and depth (DAE_audio_depth). Table I also shows

TABLE II
WORD ERROR RATES (%) ON CONTINUOUS SPEECH RECOGNITION WITH WHITE NOISE.

Feature	SNR		
	clean	20dB	10dB
MFCC	20.3	23.7	34.6
DAE_audio	21.8	24.6	35.0
DAE_audio_color	21.9	23.4	25.2
DAE_audio_depth	21.1	22.3	23.9
TDAE	20.9	22.4	24.2

the number of units in these Deep Autoencoders.

C. Results

Table II shows the results of continuous word recognition. Our proposed method is TDAE. To evaluate robustness of our method against acoustic noise, we conduct the experiments when the raw audio data are contaminated with white noise at SNRs of 10 and 20dB. The SNR condition of the training data for Deep Autoencoder is same with that of the evaluation data. Our method improves the accuracy by 2.2% from audio only deep autoencoder when SNR is 20dB, and by 10.8% when SNR is 10dB. In the noisy conditions, our method TDAE outperforms the baseline method using the standard audio feature MFCC. However, DAE_audio_depth yields the better performance than TDAE.

D. Analysis

Table II shows that audio only deep autoencoder performs worse than MFCC. We employed a random corruption function in training, which may be largely different from the noise used in the test set.

When SNR is low, the color and depth information are especially useful. In our experiments, depth features are more effective than color features. We assume that depth information is robust to differences of speakers. However, since the PCA-based color features used in our method are primitive, the performance can be improved by using other color features such as DCT or HOG.

We analyzed individual recognition results on 10dB SNR. Compared with DAE_audio results, we found that DAE_audio_color improves recognition accuracy especially for vowels and some consonants, such as /m/, which are pronounced with speakers' lips closed. In comparison of DAE_audio_color and TDAE, we found that depth information is effective to recognize some consonants, such as /k/ and /t/, which are pronounced with their mouth open.

V. CONCLUSIONS

We proposed a deep learning method, Trimodal Deep Autoencoder, for multimodal speech recognition using three-dimensional images. We collected original audio-visual data by depth camera (Kinect 2.0). The experimental results with 10dB SNR showed that our method reduced errors by 30%, from 34.6% to 24.2% from audio-only speech recognition when SNR was 10 dB.

At this stage we have only 15 hours of data for training, which may be the main reason that our method is not effective in some cases. In future, we need to collect more data to verify the effectiveness of our method over the 2D autoencoders. While our method can be used for any recognition schemes, we also plan to compare our method with the end-to-end framework where both feature extraction and speech recognition are simultaneously optimized. We are also interested in exploring the case where not only audio data, but also image/depth data are contaminated by noise.

REFERENCES

[1] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari and J. Zhou, "Audio visual speech recognition," Workshop 2000 Final Report, Technical Report 764, October 2000.

[2] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao and K. Murphy, "A coupled HMM for audio-visual speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002.

[3] D. Kolossa, S. Zeiler, A. Vorwerk and R. Orglmeister, "Audiovisual speech recognition with missing or unreliable data," *International Conference on Auditory-Visual Speech Processing (AVSP)*, 2009.

[4] P. Borde, A. Varpe, R. Manza and P. Yannawar, "Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition," *International Journal of Speech Technology*, vol. 18, no. 2, pp. 167–175, 2015.

[5] Y. Fukuda and S. Hiki, "Characteristics of the mouth shape in the production of Japanese," *Journal of the Acoustical Society of Japan*, vol. 3, no. 2 pp. 75–91, 1982.

[6] S. Nakamura and E. Yamamoto, "Speech-to-lip movement synthesis by maximizing audio-visual joint probability based on the EM algorithm," *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 27, no. 1-2, pp. 119–126, 2001.

[7] K. Kumar, T. Chen and R. M. Stern, "Profile view lip reading," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[8] G. Galatas, G. Potamianos and F. Makedon, "Audio-visual speech recognition incorporating facial depth information captured by the Kinect," *IEEE European Signal Processing Conference (EUSIPCO)*, 2012.

[9] K. Palecek, "Comparison of Depth-Based Features for Lipreading," *IEEE International Conference on Telecommunications and Signal Processing (TSP)*, 2015.

[10] J. Wang, J. Zhang, K. Honda, J. Wei and J. Dang, "Audio-Visual speech recognition integrating 3D lip information obtained from the Kinect," *Multimedia Systems*, pp. 1–9., 2016.

[11] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee and AY. Ng, "Multimodal Deep Learning," *International Conference on Machine Learning (ICML)*, 2011.

[12] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," *Journal of Machine Learning Research*, vol. 15, pp. 2949–2980, 2014.

[13] S. Tamura, H. Ninomiya, N. Kitaoka, S. Osuga, Y. Iribe, K. Takeda and S. Hayamizu, "Audio-visual speech recognition using deep bottleneck features and high-performance lipreading," *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2015.

[14] Y. Bengio, P. Lamblin, D. Popovici and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems (NIPS)*, 2007.

[15] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, K. Shikano, "ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis," *Speech Communication* vol. 9, pp. 357–363, 1990.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel and J. Silovsky (2011). "The Kaldi speech recognition toolkit," *In IEEE workshop on automatic speech recognition and understanding*, 2011.

[17] DP. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," *International Conference for Learning Representations (ICLR)*, 2015.