Exemplar-based Lip-to-Speech Synthesis Using Convolutional Neural Networks

Yuki Takashima, Tetsuya Takiguchi, Yasuo Ariki
Graduate School of System Informatics, Kobe University, Kobe, Japan
takashima@stu.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

***Abstract*—We propose in this paper a neural network-based lip-to-speech synthesis approach that converts "unvoiced" lip movements to "voiced" utterances. In our previous work, a lip-to-speech conversion method based on exemplar-based non-negative matrix factorization (NMF) was proposed. However, this approach has several problems. First, the unnatural pre-processing of visual features is required to satisfy the non-negativity constraint of NMF. Next, there is a possibility that an activity matrix cannot be shared between the visual and the audio feature in an NMF-based approach. To tackle these problems, in this paper, we use convolutional neural networks to convert visual features into audio features. Furthermore, we integrate an exemplar-based approach into the neural networks in order to adopt an advantage associated with our previous work. Experimental results showed that our proposed method produced more natural speech than conventional methods.**

***Index Terms*—lip reading, speech synthesis, multimodal, assistive technology, neural networks**

I. Introduction

An assistive technology is a system or a product that is used to improve the functional capabilities of individuals with disabilities. Over the past few decades, some speech processing techniques have been adopted in assistive technology. As a result of recent advances in statistical text-to-speech synthesis (TTS), hidden Markov model (HMM)-based TTS is used for reconstructing the voice of individuals with degenerative speech disorders [1]. Voice conversion (VC) is also applied to assistive technology. A Gaussian mixture model (GMM)-based VC method has been applied to reconstruct a speaker's individuality in electrolaryngeal speech [2] and speech recorded by non-audible murmur (NAM) microphones [3].

In this paper, we propose lip-to-speech synthesis using convolutional neural networks (CNN [4]). Lip images without a voice recording are converted to a voice utterance. We assume our proposed method will be an assistive technology for those who have a speech impediment. Moreover, our approach can be applied to voice reconstruction of videos that lack sound tracks or communication tools in noisy environments.

Lip reading is a technique of understanding speech by visually interpreting the movements of the lips, face, and tongue when the spoken sounds cannot be heard. For example, for people with hearing problems, lip reading is one communication skill that can help them communicate better. McGurk *et al.* [5] reported that we human beings perceive a phoneme not only from the auditory information of the voice but also from visual information associated with the movement of the lips and face. Moreover, it is reported that we try to catch the movement of lips in a noisy environment and we misunderstand the utterance when the movements of the lips and the voice are not synchronized. In the field of speech processing, audio-visual speech recognition (lip reading) has been researched [6]–[9]. Lip reading has the goal of classifying words or short phonemes from the lip movements.

Recently, some techniques have been introduced for lip-to-speech synthesis that generate speech from the lip movements. Aihara *et al.* [10] proposed an NMF-based lip-to-speech synthesis approach, using a high-speed camera, where the text was numbers. A high-speed camera is able to deal with fine-grained imaging [11], but the effectiveness of the time resolution has not been verified. In this work, we investigate the relationship between the performance and the number of input video frames. Akbari *et al.* [12] proposed a neural network-based method using the GRID audio-visual corpus [13]. A deep autoencoder was used for coding speech, and a deep lip-reading network extracted the encoded features from the face.

Previously, we adopted an exemplar-based NMF for lip-to-speech synthesis [10]. NMF [14] is a well-known approach that utilizes sparse representations that decompose the input into a linear combination of a small number of bases. An exemplar-based NMF decomposes the input observation into a parallel exemplars "dictionary" and the weight matrix "activity". However, there are several problems with this approach. First, the unnatural pre-processing of visual features is required to satisfy the non-negativity constraint of NMF because the discrete cosine transform (DCT) feature was used as the visual feature. This approach assumes that an activity is shared between an audio feature frame and the corresponding visual feature frame. This assumption is wrong, as shown in Section II-B. Therefore, in this paper, we investigate a novel lip-to-speech synthesis method that converts the lip image into the speech spectrum directly. This neural network-based lip-to-speech synthesis has been proposed and its effectiveness [12] has been suggested. Moreover, we integrate an exemplar-based approach into the neural networks. An exemplar-based NMF VC approach can convert speech with high naturalness [15]. In our model, the weight of the last layer is assigned to the audio dictionary and the parameters of the previous layers only are trained from the training data. In NMF VC, the speaker identity is controlled by the dictionary. Therefore, it is considered that our model is able to generate various speakers' voices by changing the dictionary of the last layer (into a

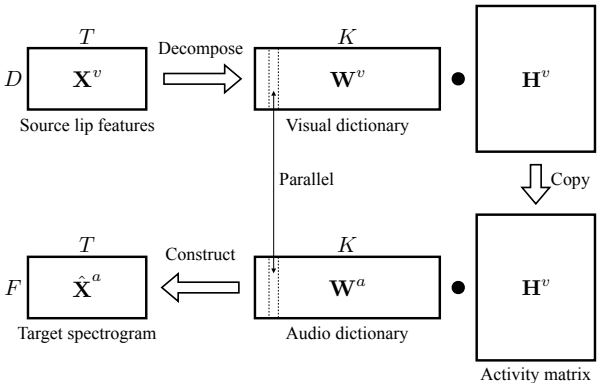

Fig. 1. Basic approach of NMF-based lip-to-speech conversion

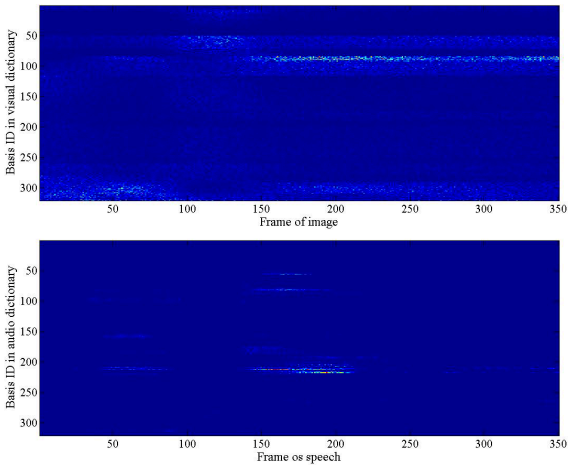

Fig. 2. Activity matrices for lip images (top) and spectrogram (bottom). Red and blue indicate large and small values of activity, respectively.

dictionary for another speaker).

II. Preliminary

A. NMF-based conversion

Fig. 1 shows the basic approach of lip-to-speech synthesis using an exemplar-based NMF [10], where D, F, T, and K represent the numbers of visual feature dimensions, audio feature dimensions, frames, and bases, respectively. This method needs two dictionaries that are phonemically parallel. $\mathbf{W}^v$ represents a visual dictionary, and $\mathbf{W}^a$ represents an audio dictionary. In an exemplar-based approach, these two dictionaries consist of the same words and are aligned with dynamic time warping (DTW), just as conventional NMF-based VC is. In this work, because we use a high-speed camera, we can obtain lip images and audio features with the same sampling rate without DTW. These dictionaries have the same number of bases.

At first, for the source visual features $\mathbf{X}^v$, the visual activity $\mathbf{H}^v$ is estimated using NMF while fixing a visual dictionary $\mathbf{W}^v$. The cost function of NMF is defined as follows:

$$d_{KL}(\mathbf{X}^v, \mathbf{W}^v\mathbf{H}^v) + \lambda||\mathbf{H}^v||_1 \ s.t.\ \mathbf{H}^v \geq 0 \qquad (1)$$

where the first term is the Kullback-Leibler (KL) divergence between $\mathbf{X}^v$ and $\mathbf{W}^v\mathbf{H}^v$ and the second term is the sparsity constraint with the L1-norm regularization term that causes the activity matrix to be sparse. λ represents the weight of the sparsity constraint. This function is minimized by iteratively updating.

This method assumes that when the source and target observations (which are the same words, with one being lip images and the other being speech) are expressed with sparse representations of the source dictionary and the target dictionary, respectively, the obtained activity matrices are approximately equivalent. The estimated visual activity $\mathbf{H}^v$ is multiplied to the audio dictionary $\mathbf{W}^a$, and the target spectrogram $\hat{\mathbf{X}}^a$ is constructed.

$$\hat{\mathbf{X}}^a = \mathbf{W}^a\mathbf{H}^v \qquad (2)$$

B. Problems

This conventional NMF-based lip-to-speech synthesis approach has several problems. First, in [10], DCT features are calculated from lip images, and used as the visual feature. To use DCT features as visual features, the unnatural preprocessing of visual features is required to satisfy the non-negativity constraint of NMF. Next, this approach assumes that activity matrices estimated from the audio feature and the visual feature are equivalent to each other, and can be shared for a voice and corresponding lip movements. However, there is a possibility that this assumption is wrong. Fig. 2 shows an example of the activity matrices estimated from the word /nana/ (/seven/ in English). These activity matrices are not similar to each other at all. For this reason, the distribution of the visual feature is different from the distribution of the audio feature. For example, an interval from 210 to 350 frames is a short pause. In this interval, the spectrograms often have very small values. For this reason, the estimated activity matrices have values close to zero for all bases. On the other hand, DCT features calculated from lip images in a corresponding period represent some characteristics for the lip shape, so the estimated activity matrices have some values. Therefore, it seems that this assumption that activity matrices estimated from the audio feature and the visual feature are equivalent to each other is not established.

III. Proposed method

A. Flow of the proposed method

Fig. 3 shows the flow of our proposed method. First, we construct an audio dictionary. We employ WORLD [16] for feature extraction and speech synthesis, and use the WORLD spectrum for audio features. The WORLD spectra calculated from the speech data are merged, and we obtain an audio dictionary. Next, we calculate a high level representation from lip images using CNNs. Unlike a previous work [10], we use the lip image as an input feature directly. Our model has

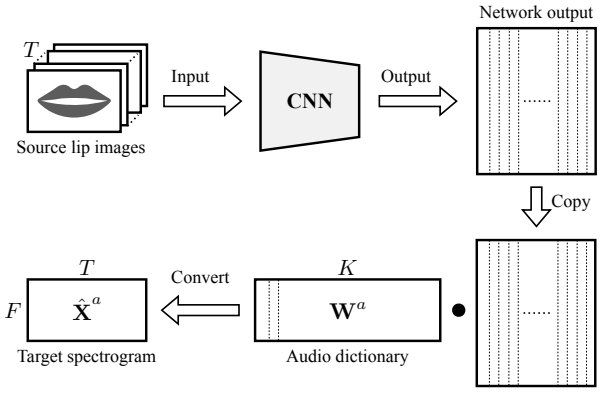

Fig. 3. Flow of the proposed method.

been arranged in M successive frames as input. Finally, the estimated high-level representation is multiplied to the audio dictionary $\mathbf{W}^a$ and the target spectra $\hat{\mathbf{X}}^a$ are constructed. In our proposed method, because the converted spectra are calculated using an audio dictionary, it seems that the converted speech has high naturalness.

B. Architecture

Table I shows the architecture of CNNs where conv_MT and lReLU indicate the convolution with multiple towers [8] and leaky ReLU (lReLU) parameter $a = 0.2$, respectively. All convolution layers are pre-activation batch-normalized. For the first to fourth layers, we set convolution layers with multiple towers. There is no time-domain connectivity between frames. Each convolution layer is associated with shared weights between frames and takes an input frame. In the fifth layer, there are shared weights within input channels (time-domain), and the time information is merged. In the last layer, we obtain K-dimensional representation where K is the number of bases in an audio dictionary.

C. Training

Given the input feature $\boldsymbol{x}_t^v = \{x_{t-M/2}^v, \cdots, x_t^v, \cdots, x_{t+M/2}^v\}$ where x_t^v is an input image at frame t, the output $\boldsymbol{y}_t \in \mathbb{R}^K$ at frame t is defined as follows:

$$\boldsymbol{y}_t = f(\boldsymbol{x}_t^v; \theta), \tag{3}$$

where $f(\cdot;\theta)$ indicates CNNs that have parameters θ. We calculate outputs along time from the input image sequence, and then the converted spectrogram is written as follows:

$$\hat{\mathbf{X}}^a = \mathbf{W}^a \mathbf{Y}, \tag{4}$$

where $\mathbf{Y} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_t, \cdots, \boldsymbol{y}_T\}$. The objective function used to train CNNs is as follows:

$$\min_{\theta} \mathcal{L}(\mathbf{X}^a, \hat{\mathbf{X}}^a) + \lambda ||\mathbf{Y}||_1, \tag{5}$$

where $\mathcal{L}(\mathbf{X}, \mathbf{Y})$ indicates the mean square error between $\mathbf{X}$ and $\mathbf{Y}$. $\mathbf{X}^a$ is training speech data that corresponds to input lip images. The training speech $\mathbf{X}^a$ and the dictionary speech $\mathbf{W}^a$ are not duplicated. The second term in Eq. (5) is the sparse constraint with an L1-norm regularization term that causes $\mathbf{Y}$ to be sparse.

TABLE I
NETWORK ARCHITECTURE.

Layer index	Operation
0	30 × 45 × M Image
1	3×3 conv_MT, 64, lReLU
2	3×3 conv_MT, 64, lReLU, 2×2 max-pool
3	3×3 conv_MT, 128, lReLU
4	3×3 conv_MT, 128, lReLU 2×2 max-pool
5	3×3 conv w/ shared weights, 256, lReLU
6	3×3 conv, 256, lReLU
7	K dense, ReLU

IV. EXPERIMENTS

A. Conditions

We recorded 158 utterances of clean continuous speech consisting of Japanese numbers and ○ (/maru/) of one Japanese male by using a high-speed camera. A high-speed camera is MEMRECAM GX-1. We used 10 utterances as test data. To construct an audio dictionary, we used ten isolated numbers and ○. To train CNNs, we used the remaining utterances. The number of frames in the audio dictionary was 2,469. Audio and visual data were recorded at the same time in a quiet room.

The frame rate of the visual data was 500 fps and the image size was 640×480, which was converted to grayscale, and the 30×45 mouth area was extracted. Fig. 4 shows examples of lip images. Sampling frequency of speech was 12kHz. The audio spectrum was extracted by WORLD from the speech data with a 2ms frame shift. The number of dimensions of the audio spectrum was 257. In this work, we focus on spectrum conversion, so the other information, such as aperiodic components, is synthesized using that of target speech.

The proposed method was evaluated by comparing it with a conventional NMF-based method [10] ("Conv") and a CNN-based method without using an audio dictionary ("CNN"). For the conventional system, we used 200-dimensional DCT coefficients of lip motion images of the source speaker ' s utterances as input features. We introduced the segment features for the DCT coefficient, which consist of its consecutive frames (the 2 frames coming before and the 2 frames coming after). Therefore, the total dimension of visual feature is 1,000. For the CNN-based system, we estimated the spectrum without using an audio dictionary. In the last layer (layer index 7 in Table I), we set a dense layer of 257 units. The batch size was set to 128, and we used SGD with a learning rate of 1.0×10^{-5}, a momentum of 0.9, and a weight decay of 2.0×10^{-4} to optimize the CNNs with 30 epochs. A sparse coefficient was set to 5.0×10^{-3}. We set the length of input frames M to 5.

Fig. 4. Sample of successive frames (an interval of 22 ms). The mouth motion is shown left to right.

In order to evaluate our proposed method, we conducted an objective evaluation. We used mel-cepstrum distortion (MCD) [dB] as a measure of the objective evaluations, defined as follows:

$$MCD = (10/\ln 10)\sqrt{2\sum_{d=1}^{24}(mc_d^{conv} - mc_d^{tar})^2} \quad (6)$$

where mc_d^{conv} and mc_d^{tar} denote the d-th dimension of the converted and target mel-cepstral coefficients, respectively.

B. Results and discussion

Table II shows the average MCD values for each method with 95% confidence intervals where "eCNN" and "eCNN w/ sparse" indicate our proposed exemplar-based CNN with or without the sparse constraint. Here, a lower value indicates a better result. Neural network-based methods outperformed the conventional NMF-based method using DCT features as input features. Moreover, we also confirmed that our proposed exemplar-based approach has comparable performance to the method that does not use a dictionary.

TABLE II
AVERAGE MCD OF EACH METHOD.

Method	Conv	CNN	eCNN	eCNN w/ sparse
MCD [dB]	8.79±2.22	8.17±0.90	8.27±0.50	8.14±0.45

Fig. 5 shows examples of target spectrograms and converted spectrograms. As shown in the middle panel, the spectrogram converted using the conventional method was blurring in the low-frequency portion. For this reason, it seems that the activity cannot be shared between the audio and the visual features. Our proposed method generated the proper spectrum in the low-frequency portion. The high-frequency portion in the spectrum was not generated sufficiently, and this is one of our next challenges.

We evaluated the performance of our proposed method using different numbers of input frames. The results are shown in Fig. 6. We changed the number of input frames as 5, 9, 13, and 17. As shown in this figure, the best performance was achieved with 5 frames (an interval of 5ms). Therefore, there are significant movements contained in the interval of 5ms that cannot be captured by normal cameras.

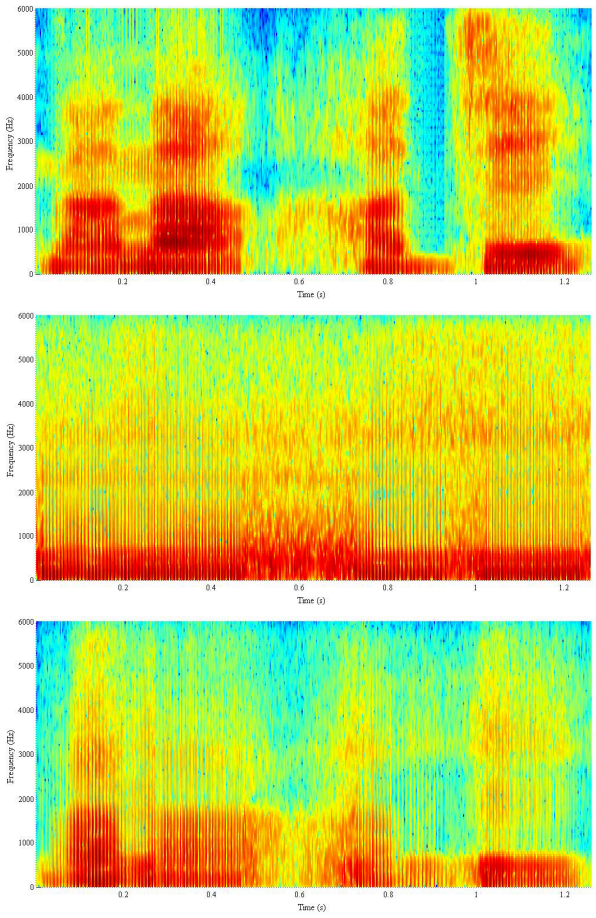

Fig. 5. Example of spectrograms for /nana hachi/ (/seven eight/ in English) uttered by an evaluation speaker (top), converted using "Conv" (middle), and converted using "eCNN w/ sparse" (bottom). The red and the blue indicate the high and the low amplitude, respectively.

V. CONCLUSION

In this paper, we proposed a lip-to-speech synthesis method that uses exemplar-based neural networks. In past works, exemplar-based NMF-VC has shown that converted speech has high-naturalness. Our previous work utilized this advantage in lip-to-speech synthesis; however, there are several problems with this approach. First, visual features are processed using unnatural pre-processing to satisfy the non-negativity constraint of NMF. Next, there is a possibility that activity matrices cannot be shared between the visual and the audio domains in an NMF-based approach. To tackle these problems, we introduced an exemplar-based approach into neural networks that converts the lip images into a spectrum while fixing weights in the last layer as exemplars. During training, the model learns the relationship between lip movements and sounds. In our experiments, we confirmed that our proposed method outperforms the conventional NMF-based method. Moreover, our proposed method achieved a

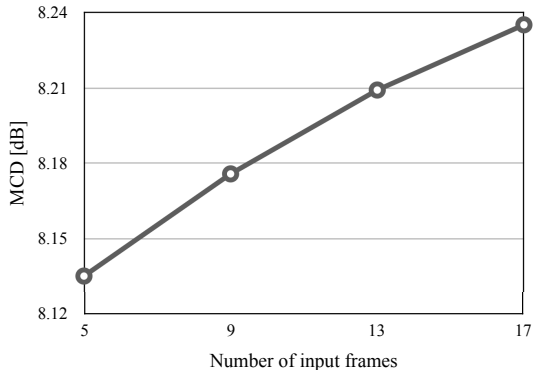

Fig. 6. MCD of our proposed method with sparse constraint with varying the number of input frames.

performance that is comparable to the CNN-based method that does not use a dictionary. Although the method without using the dictionary generates only the training speaker's voice, the proposed method is able to generate the various voices of the speaker by changing the dictionary. In future experiments, we will evaluate the proposed method on a speaker conversion task. By using a high-speed camera, we found that important information is included in a very short time interval to generate a voice that cannot be captured by ordinary video cameras. For our future work, we will investigate some constraints to generating more natural voices.

ACKNOWLEDGMENT

This work was supported in part by PRESTO, JST (Grant No. JPMJPR15D2) and JSPS KAKENHI (Grant No. 17H01995 and No. JP17J04380).

REFERENCES

[1] C. Veaux, J. Yamagishi, and S. King, "Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders," in *INTERSPEECH*, pp. 967–970, 2012.

[2] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[3] ——, "Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech." in *INTERSPEECH*, 2006.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[5] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 12 1976.

[6] A. Verma, T. Faruquie, C. Neti, S. Basu, and A. Senior, "Late integration in audio-visual continuous speech recognition," in *Automatic Speech Recognition and Understanding*, vol. 1, pp. 71–74, 1999.

[7] K. Palecek and J. Chaloupka, "Audio-visual speech recognition in noisy audio environments," in *International Conference on Telecommunications and Signal Processing*, pp. 484–487, 2013.

[8] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*, 2016.

[9] M. Wand and J. Schmidhuber, "Improving speaker-independent lipreading with domain-adversarial training," in *INTERSPEECH*, pp. 3662–3666, 2017.

[10] R. Aihara, K. Masaka, T. Takiguchi, and Y. Ariki, "Lip-to-speech synthesis using locality-constraint non-negative matrix factorization," in *International Workshop on Machine Learning in Spoken Language Processing*, 2015.

[11] A. Davis, M. Rubinstein, N. Wadhwa, G. Mysore, F. Durand, and W. T. Freeman, "The visual microphone: Passive recovery of sound from video," *ACM Transactions on Graphics (Proc. SIGGRAPH)*, vol. 33, no. 4, pp. 79:1–79:10, 2014.

[12] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, "Lip2audspec: Speech reconstruction from silent lip movements video," in *ICASSP*, pp. 2516–2520, 2018.

[13] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, November 2006.

[14] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, pp. 556–562, 2000.

[15] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *SLT*, pp. 313–317, 2012.

[16] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions*, vol. 99-D, no. 7, pp. 1877–1884, 2016.