AUDIOVISUAL SPEAKER CONVERSION: JOINTLY AND SIMULTANEOUSLY TRANSFORMING FACIAL EXPRESSION AND ACOUSTIC CHARACTERISTICS

Fuming Fang[1], *Xin Wang*[1], *Junichi Yamagishi*[1,2], *Isao Echizen*[1]

[1]National Institute of Informatics, Tokyo, Japan
[2]The University of Edinburgh, Edinburgh, UK

ABSTRACT

An audiovisual speaker conversion method is presented for simultaneously transforming the facial expressions and voice of a source speaker into those of a target speaker. Transforming the facial and acoustic features together makes it possible for the converted voice and facial expressions to be highly correlated and for the generated target speaker to appear and sound natural. It uses three neural networks: a conversion network that fuses and transforms the facial and acoustic features, a waveform generation network that produces the waveform from both the converted facial and acoustic features, and an image reconstruction network that outputs an RGB facial image also based on both the converted features. The results of experiments using an emotional audiovisual database showed that the proposed method achieved significantly higher naturalness compared with one that separately transformed acoustic and facial features.

Index Terms— Audiovisual speaker conversion, multi-modality transformation, machine learning

1. INTRODUCTION

With the development of information processing technology and the spread of Internet, voice and facial expression-based methods are being used more and more in our everyday lives. Two prominent methods are voice conversion [1] and face transformation [2], which change a person's voice or facial expressions into that or those of another person. These methods can be used for privacy protection, film/animation production, games, and other voice/facial signal-based transformations.

Changing both voice and facial expressions is important for certain applications, such as video games. An obvious way to achieve this is to separately transform the voice and facial expressions using two separate methods [3, 4]. This approach can result, however, in loss of naturalness due to asynchronous voice and facial movements due to transformation errors, delays (when considering context information), and other factors. Naturalness can be improved by using a synchronization method.

Another way to improve naturalness is to utilize the correlation between speech and facial movements [5] and jointly transform them so that they are always associated together. We have developed such a method, an audiovisual speaker conversion (AVSC) method that simultaneously transforms acoustic and facial characteristics. It uses three neural networks: a conversion network that fuses and transforms the acoustic and facial features of a source speaker into those of a target speaker, a waveform generation network that produces the waveform given both the converted acoustic and facial features, and an image generation network that outputs the RGB facial image also based on both the converted features. With the proposed method, we observed higher naturalness and quality than when the acoustic and facial features were separately transformed. This appears to be the first ever research on integrating voice conversion and face transformation in one system[1].

The rest of this paper is organized as follows. Section 2 discusses the differences between the proposed method and related ones. Section 3 describes the proposed method. Section 4 describes the experimental conditions, and Section 5 presents and discusses the results. Finally, Section 6 summarizes the key points and mentions future work.

2. RELATED WORK

The proposed method is related to work in four areas: audiovisual voice conversion, audiovisual speech enhancement, lip movement-to-speech synthesis, and speech-to-lip movement synthesis.

Tamura et al. [6] proposed an audiovisual voice conversion method that learns highly correlated acoustic and lip movement features by using deep canonical correlation analysis [7] and then ties them together as a new feature. They reported significant improvement in terms of speech quality under noisy conditions compared with using acoustic feature only. Gabbay et al. [8] proposed an audiovisual speech enhancement method that fuses lip feature and a noisy spectrum using a neural network and directly predicts a clean spectrum. Gogate et al. [9] and Afouras et al. [10] developed similar speech enhancement methods that predict a mask from lip images and noisy acoustic features using a neural network to filter out noise.

A lip movement-to-speech synthesis system was developed by Kumar et al. [11] that pairs mouth image sequence obtained using multi-view cameras with the corresponding audio to learn a mapping function. Kumar et al. [12] and Suwajanakorn et al. [13] developed similar speech-to-lip movement synthesis methods that generate mouth keypoints from audio information and then render RGB images of the mouth from the mouth shape (represented by the keypoints). Taylor et al. [14] used speech to control an avatar.

The proposed method differs from these methods in that it uses not only lip movements but also facial expressions and movements. Furthermore, it transforms both audio and facial expressions.

3. PROPOSED METHOD

The main idea of the proposed AVSC method is to correlate acoustic and facial characteristics so that they can compensate for each other during transformation and achieve high naturalness. Figure 1 illustrates an implementation of the proposed method, which uses three networks: an audiovisual transformation network, a WaveNet [15], and an image reconstruction network. The audiovisual transformation network is used to convert acoustic and facial features from a

[1]A demonstration is available at `https://nii-yamagishilab.github.io/avsc/index.html`

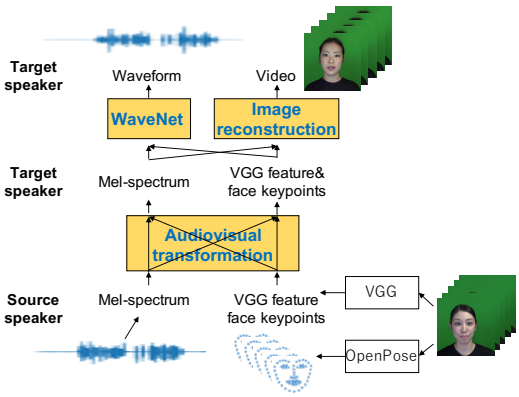

Fig. 1. An implementation of proposed audiovisual speaker conversion method.

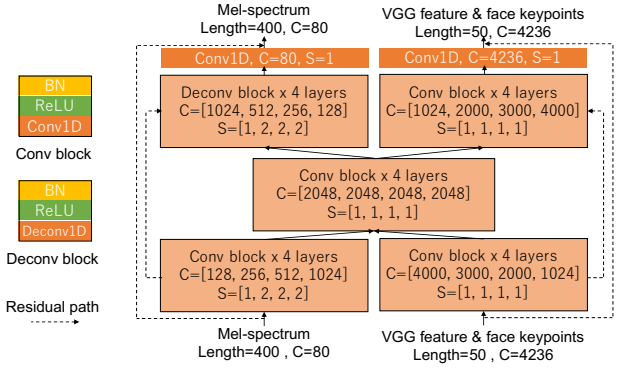

Fig. 2. Audiovisual transformation network. "BN" denotes batch normalization, "Conv1D" denotes 1-D CNN layer, "Deconv1D" means transposed 1-D CNN layer, "C" means number of channels for each layer, and "S" means stride for each layer. Kernel size for all convolution layers is five. Activation function of hidden layers is ReLU and that of output layers is linear function. Batch normalization is performed in all hidden layers and not in output layer. Residual path connects feature maps having same shape.

source speaker to a target speaker. The WaveNet and the image reconstruction network synthesize speech and RGB images from the transformed acoustic and facial features, respectively. Finally, a video of the target speaker is created using the synthesized speech and image sequence.

3.1. Audiovisual transformation network

The acoustic feature is the mel-spectrum with 80 dimensions. It is extracted from the waveform by setting the window size to 25 ms and the hop length to 5 ms. The facial feature is extracted using a pre-trained VGG-19 [16] network. The 17th hidden layer's output is used (denoted as "VGG feature"). Because the VGG-19 is stacked by convolutional neural network (CNN) [17] and max-pooling layers followed by fully connected layers, most of the original geometric information is lost, and only high-level features are preserved. To enable facial geometric information to be used, facial keypoints are extracted from each video frame by using OpenPose [18]. These keypoints describe the location and shape of a face. For example, they mark the contours of the jaw, mouth, nose, eyes, and eyebrows. The VGG feature (4096 dimensions) and face keypoints ($70 \times 2 = 140$ dimensions) are concatenated to create a new facial feature (4236 dimensions). In addition, since video data has a frame rate of 25 fps (40 ms per frame), a facial feature corresponds to eight mel-spectrum frames.

Figure 2 shows the architecture of the audiovisual transformation network, which contains a stack of 1-D convolutional layers. Its design was inspired by the work of Afouras et al. [10]. The convolutions are performed along the temporal dimension, and the feature dimension is treated as channel. This makes it possible to match the sampling rates for the acoustic and facial features by adjusting the stride. This design also enables context information to be taken into account and thereby generate more fluent audio and facial movements. The network consists of five sub-networks and two output layers. Based on sampling rate of the facial feature, The lower left sub-network down-samples the acoustic feature by performing convolution three time with a stride of two. The lower right one maintains the facial feature sampling rate and reduces the feature dimension. The middle one fuses acoustic and facial information and associates them together. The upper left one up-samples and transforms the fused features into the target speaker's acoustic feature domain using transposed convolution [19] layers. The upper right one transforms the fused features into target speaker's facial feature domain. In addition, the feature maps having the same shape are connected by a residual path [20]. Batch normalization [21] is performed in each hidden layer after rectified linear unit (ReLU) [22] activation.

The training data is truncated every two seconds, so the mel-spectrum and facial feature have shapes of 400×80 and 50×4236, respectively. The L1 norm is used as the training objective, and the acoustic part loss is weighted by 10. During the test phase, the entire length of test data was input to the network.

3.2. WaveNet

The WaveNet is used to convert the transformed mel-spectrum and facial feature into the speech waveform. WaveNet is an autoregressive [23] neural-network-based waveform model that generates waveform sampling points one by one.

The WaveNet structure is the same as that of the one used in another study [24] except the condition module which takes the mel-spectrum as input. It consists of a linear projection layer, 40 dilated convolution [25] layers, a post-processing block with a softmax output layer, and a condition module. The linear projection layer takes as input a waveform value generated in the previous time step while the condition module takes the transformed mel-spectrum and facial feature as input. Given the outputs from the linear layer and the condition module, the dilated convolution layers compute hidden features, which the post-processing module uses to compute the distribution waveform sampling point for the current time step. A waveform value is generated from this distribution, and this process is repeated to generate the entire waveform.

3.3. Image reconstruction network

The image reconstruction network synthesizes an RGB image from the transformed acoustic and facial features. It contains nine stacked layers: two fully connected layers and seven convolution layers, as shown on the left in Figure 3. Eight frames of acoustic features and one frame of facial feature are concatenated as inputs. The network first transforms the concatenated feature into a fused feature with 4096 dimensions using two fully connected layers. The fused feature is then reshaped into a 2-D image ($64 \times 64 \times 1$) and sent to the next convolution layer. Finally, an image ($256 \times 256 \times 3$) is generated by performing convolutions and transposed convolutions.

Training of the image reconstruction network is based on a least squares generative adversarial network (LSGAN) [26] consisting of

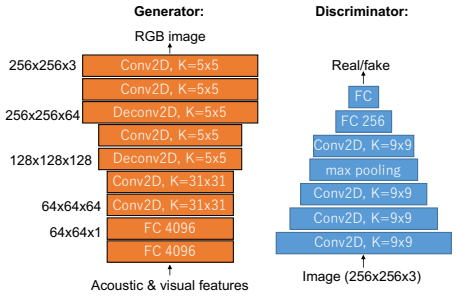

Fig. 3. Image reconstruction network. Generator is used to reconstruct image from acoustic and visual features. Discriminator is an additional network used for training generator using adversarial loss. "Conv2D" means 2-D CNN layer, "Deconv2D" means transposed 2-D CNN layer, "FC" means fully connected layer, and "K" means kernel size. The convolution layers in the generator have (from the bottom) 64, 64, 128, 128, 64, 64, and 3 channels. The Conv2D layers have a stride of one, and the Deconv2D layers have a stride of two. The Conv2D layers in the discriminator have (from the bottom) 8, 16, 32, and 32 channels. Both the Conv2D and max pooling layers have a stride of two. Activation function of all hidden layers is ReLU. Batch normalization is performed in the generator before activation and not in the discriminator.

a generator and a discriminator (Figure 3 right side). The discriminator maximizes the probability of data from training images and minimizes the probability of generated images from the generator. The generator strives to generate images similar to the training data in order to maximize the probability and fool the discriminator. The L1 norm (weighted by 10) is used to stabilize the training process.

4. EXPERIMENTAL SETUP

We compared the performance of the proposed AVSC method with a baseline method that separately transforms acoustic and facial features. We carried out an objective and a subjective experiment. The objective experiment evaluated the correlation between speech and lip movements. The subjective experiment evaluated the naturalness, quality, and speaker similarity of the converted speech and video.

4.1. Database

To accurately evaluate the correlation between audio and facial features, we created an emotional audiovisual database using input from two Japanese female actors. Seven emotions were defined: neutral, normal happiness, strong happiness, normal sadness, strong sadness, normal anger, and strong anger. For each emotion, we used 100 different sentences selected from dialogs in novels. We asked the two actors to utter each sentence while displaying the corresponding emotion. Four people monitored the recording sessions, and if any of them felt that the target emotion was not displayed in the speech or facial expression, the actor was instructed to repeat the recording of that sentence.

The recording took place in a soundproof chamber. A condenser microphone (NEUMANN87) and a video camera (Sony HDR-720V/B) were set at front of the actor. A green cloth sheet was used as the background. The audio was recorded at 96 kHz with 24-bit resolution. The video was recorded at 60 fps with 1920×1080 resolution. There were 17 recording sessions in total. A clapperboard was clapped shut at the beginning of each session to enable the audio to be synchronized with the video. Finally, the audio signal recorded by the video camera was replaced with that recorded using the condenser microphone.

Since the two actors used the same sets of sentences, the database had parallel recordings (700 per actor). The duration was approximately 1h10min for each actor. The average sentence duration was 5.9 s. We refer to the two actors as speakers F01 and F02.

4.2. Training data and test data

We designated speaker F01 as the source speaker and F02 as the target speaker. We randomly selected 90 data samples for each emotion as training data (630 samples in total) for each speaker. The remaining 70 samples were used as test data. We down-sampled the audio signal to 48 kHz with 16-bit resolution and then extracted the mel-spectrum. We down-sampled the video signal to 25 fps, centered the speaker, and resized the video to 1080×1080. We then extracted the image and resized it to 224×224 for VGG feature extraction and 256×256 for face keypoint extraction as well as for use as training data for the image reconstruction network.

4.3. Proposed method setup

For dynamic time warping (DTW)-based alignment, the distance between features from the two speakers was calculated by summing the dimension-averaged L2-norm of the acoustic feature and that of the facial feature (we tied every eight acoustic features to match length of the facial feature). The learning rate for the audiovisual transformation network was 10^{-4}, the mini-batch size was 64, and the number of epochs was 600.

The WaveNet was adapted from a pre-trained model [24] that was trained using 15 hours of Japanese speech data. We fine tuned over 199 epochs using the F02 training data. The number of epochs was set on the basis of the results of a preliminary experiment.

The image reconstruction network was trained using the F02 training data. The learning rates were set to 10^{-3} and 10^{-5} for the generator and discriminator, respectively, the mini-batch size was 64, and number of epochs was 30.

4.4. Baseline setup

For the baseline method, we removed the acoustic-related or facial-related part from the proposed method. In addition, since it was not necessary for the baseline method to down- or up-sample acoustic features, we changed the stride of the audiovisual transformation network to one. From the results of a preliminary experiment on WaveNet adaptation, we set the number of epochs to 100. The other hyper-parameters (learning rate, mini-batch size, and number of epochs) were the same as for the proposed method. As additional information, we tuned the learning rate and number of epochs using the baseline method and directly applied them to the proposed method. We did not tune mini-batch size.

4.5. Evaluation setup

We evaluated the correlation between the converted speech and lip movements using canonical correlation analysis, which calculates correlation coefficient r between two sequences. We re-extracted the mel-spectrum and lip keypoints from the converted speech and images. We set the window length and hop length for the mel-spectrum to 40 ms and 40 ms, respectively.

The naturalness and quality were evaluated on a 1-to-5 Likert mean opinion score (MOS) scale. The speaker similarity was evaluated using a preference test. The evaluation was carried out by means of a crowdsourced web-based interface. On each web page, we presented three questions about naturalness and quality for the audio-only evaluation case, the visual-only evaluation case, and the audiovisual evaluation case. We presented only audio or silent video for the audio-only evaluation case and the visual-only evaluation case. For the audiovisual evaluation case, we asked the evaluators to view

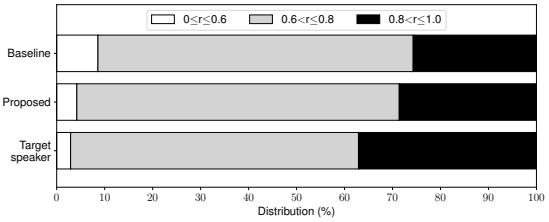

Fig. 4. Distribution of correlation coefficient r between mel-spectrum and lip movements.

a video and assess the quality of speech, image, and synchronization between speech and lip movement. We also presented additional three questions about speaker similarity for the audio-only evaluation case, visual-only evaluation case, and audiovisual evaluation case. The evaluators were limited to a maximum of 50 pages, and they had to listen/view all samples and answer all questions. There was a total of 186 valid evaluators, and they produced 4,995 page data points, which is equivalent to 35.7 evaluations per sample.

The statistical significance analysis was based on an unpaired two-tail t-test with a 99% confidence interval.

5. RESULTS

5.1. Correlation between speech and lip movements

Although our main intention is to correlate voice and facial expression, the proposed method includes synchronized generation of speech and lip movements. Figure 4 shows the distribution of correlation coefficients between the mel-spectrum and lip movements. The correlation was higher for the proposed method than for the baseline. This suggests that acoustic and facial feature are more closely associated if they are jointly and simultaneously transformed. However, there was a gap between the results for the proposed method and for the target speaker.

5.2. Subjective evaluation

As shown in Table 1, the MOS values with the proposed method were significantly better in both the audio-only and audiovisual evaluation cases achieved than with the baseline method. One reason was that the facial feature compensated for the acoustic feature and the proposed method achieved better synchronous. The slightly better performance of the proposed method in the visual-only evaluation case is attributed to the facial feature dominating the fused feature, making it difficult to take advantage of the acoustic feature. The scores for the emotional test samples were smaller than those for the neutral samples in almost case with the baseline method. It was possible to achieve a higher or similar score as the neutral one by fusing both the acoustic and facial features. e.g., sadness in the audio-only evaluation case and happiness in the audiovisual evaluation case. This indicates that facial movements and some speech characteristics might help enhance emotion transformation.

The preference test results (Figure 5) were similar. The proposed method achieved higher speaker similarity than the baseline method for the audio-only and audiovisual evaluation cases. This might be because facial identity and voice identity helped to estimate accurate parameters of the networks.

6. SUMMARY AND FUTURE WORK

Our proposed audiovisual speaker conversion method simultaneously transforms voice and facial expression of a source speaker into those of a target speaker. We implemented this method using three networks: a conversion network fuses and transforms the acoustic and facial features of the source speaker into those of the target speaker, a WaveNet synthesizes the waveform from the converted features, and an image reconstruction network generates RGB images from the converted features. Experiments using an emotional audiovisual database showed that the proposed method can achieve higher naturalness/quality and speaker similarity than a baseline method that separately transforms the acoustic and facial features.

Since the facial features may dominate the transformation, we plan to improve our method to better balance the acoustic and facial features. The use of a parallel training approach makes it necessary to align training data, so we had to carefully balance the acoustic and facial features. We will thus consider developing a non-parallel training method [27] for audiovisual speaker conversion.

Table 1. MOS values for speech, visual, and audiovisual naturalness/quality. "Pr" means proposed method, "Bs" means baseline method, and "+" indicates strong emotion.

Emotion type	Evaluation modality					
	Audio-only		Visual-only		Audiovisual	
	Pr	Bs	Pr	Bs	Pr	Bs
Neutral	2.70	2.33	3.79	3.65	3.37	2.98
Happiness	2.39	2.15	3.69	3.73	3.41	2.93
+Happiness	2.33	2.05	3.36	3.33	3.25	2.85
Sadness	2.72	2.24	3.32	3.40	3.21	2.95
+Sadness	2.24	2.01	3.46	3.45	3.31	3.16
Anger	2.46	2.08	3.42	3.35	3.22	2.99
+Anger	1.96	1.79	3.27	3.11	3.06	2.76
Average	2.40	2.09	3.47	3.43	3.26	2.95

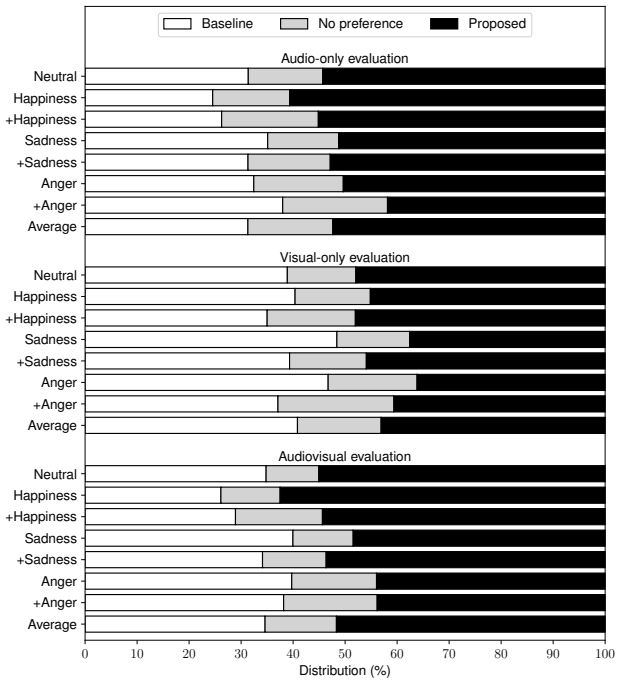

Fig. 5. Speaker similarity preference test results. "+" means strong emotion.

7. ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Numbers (16H06302, 17H04687, 18H04120, 18H04112, 18KT0051) and by JST CREST Grant Number JPMJCR18A6, Japan.

8. REFERENCES

[1] Masanobu Abe, Satoshi Nakamura, Kiyohiro Shikano, and Hisao Kuwabara, "Voice conversion through vector quantization," *Journal of the Acoustical Society of Japan (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[2] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner, "Face2face: Real-time face capture and reenactment of RGB videos," in *Proc. CVPR*, 2016, pp. 2387–2395.

[3] Maria Markaki Kristina Stankovic Aurelie Zara Levent Arslan Thierry Dutoit Igor S. Panzic Murat Saraclar Yannis Stylianou Zeynep Inanoglu, Matthieu Jottrand, "Multimodal speaker identity conversion - continued -," in *Proceedings of the eNTERFACE07 Workshop on Multimodal Interfaces*, 2007.

[4] Hanna Greige and Walid Karam, *Audio-Visual Biometrics and Forgery*, 08 2011.

[5] Gopal Ananthakrishnan, Olov Engwall, and Daniel Neiberg, "Exploring the predictability of non-unique acoustic-to-articulatory mappings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2672–2682, 2012.

[6] Satoshi Tamura, Kento Horio, Hajime Endo, Satoru Hayamizu, and Tomoki Toda, "Audio-visual voice conversion using deep canonical correlation analysis for deep bottleneck features," in *Proc. Interspeech*, 2018, pp. 2469–2473.

[7] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning*, 2013, pp. 1247–1255.

[8] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg, "Visual speech enhancement using noise-invariant training," *arXiv preprint arXiv:1711.08789*, 2017.

[9] Mandar Gogate, Ahsan Adeel, Ricard Marxer, Jon Barker, and Amir Hussain, "DNN driven speaker independent audio-visual mask estimation for speech separation," in *Proc. Interspeech*, 2018, pp. 2723–2727.

[10] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech*, 2018, pp. 3244–3248.

[11] Yaman Kumar, Mayank Aggarwal, Pratham Nawal, Shin'ichi Satoh, Rajiv Ratn Shah, and Roger Zimmerman, "Harnessing ai for speech reconstruction using multi-view silent video feed," in *2018 ACM Multimedia Conference (MM '18)*, 2018.

[12] Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brébisson, and Yoshua Bengio, "Obamanet: Photo-realistic lip-sync from text," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[13] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 95, 2017.

[14] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews, "A deep learning approach for generalized speech animation," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 93, 2017.

[15] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[16] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[17] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[18] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2D pose estimation using Part Affinity Fields," in *CVPR*, 2017.

[19] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus, "Deconvolutional networks," in *Proc. CVPR*. 2010, pp. 2528–2535, IEEE.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[21] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*. 2015, pp. 448–456, JMLR.org.

[22] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li, "Empirical evaluation of rectified activations in convolutional network," in *ICML Deep Learning Workshop*, 2015.

[23] Michael I Jordan, "Serial order: A parallel distributed processing approach," in *Advances in psychology*, vol. 121, pp. 471–495. Elsevier, 1997.

[24] Hieu-Thi Luong, Xin Wang, Junichi Yamagishi, and Nobuyuki Nishizawa, "Investigating accuracy of pitch-accent annotations in neural-network-based speech synthesis and denoising effects," in *Proc. Interspeech*, 2018, pp. 37–41.

[25] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2016.

[26] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *Proc. ICCV*. IEEE, 2017, pp. 2813–2821.

[27] Fuming Fang, Junichi Yamagishi, Isao Echizen, and Jaime Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," in *Proc. ICASSP*, 2018, pp. 5279–5283.