INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Face2Speech: Towards Multi-Speaker Text-to-Speech Synthesis Using an Embedding Vector Predicted from a Face Image

Shunsuke Goto[1,2], *Kotaro Onishi*[1,3], *Yuki Saito*[2], *Kentaro Tachibana*[1], *and Koichiro Mori*[1]

[1]DeNA Co., Ltd., Tokyo, Japan
[2]The University of Tokyo, Japan
[3]The University of Electro-Communications, Tokyo, Japan

goto@gavo.t.u-tokyo.ac.jp, koichiro.mori@dena.com

Abstract

We are quite able to imagine voice characteristics of a speaker from his/her appearance, especially a face. In this paper, we propose Face2Speech, which generates speech with its characteristics predicted from a face image. This framework consists of three separately trained modules: a speech encoder, a multi-speaker text-to-speech (TTS), and a face encoder. The speech encoder outputs an embedding vector which is distinguishable from other speakers. The multi-speaker TTS synthesizes speech by using the embedding vector, and then the face encoder outputs the embedding vector of a speaker from the speaker's face image. Experimental results of matching and naturalness tests demonstrate that synthetic speech generated with the face-derived embedding vector is comparable to one with the speech-derived embedding vector.

Index Terms: cross-modal face/voice generation, text-to-speech synthesis, multi-speaker modeling, speaker embedding

1. Introduction

Humans are, to a certain degree, able to imagine voice characteristics from a person's face, and vice versa. Namely, we can recognize some characteristics such as gender, age, and ethnicity from his/her voice and face, and there is a relationship between characteristics identified by vocal features and characteristics identified by facial features [1]. This paper tries to offer a better understanding of the audio/visual cross-modality of a human's perception through a machine learning framework, which would be beneficial for both the fields of speech processing and computer vision.

This paper especially focuses on text-to-speech (TTS) synthesis [2], which is a technique for synthesizing a natural-sounding and easily-controllable human voice from a given text. Recent developments in TTS based on deep neural networks (DNNs) [3, 4] have made the quality of the synthesized speech so high that it is almost indistinguishable from natural speech in reading-style TTS [5]. In addition, there are some studies for developing more controllable DNN-based TTS systems that can easily modify characteristics of the synthesized speech such as the speaker identity and emotion. Akuzawa *et al.* proposed the VAE-Loop [6], which introduces variational autoencoders (VAEs) [7] to its TTS model that can learn a data-driven embedding vector of speech. Similar ideas have been introduced to more powerful frameworks based on end-to-end TTS [8, 9]. Also, it has been reported that embedding vectors derived from a speaker recognition model (e.g., d-vector [10] and x-vector [11]) can be used for controlling the speaker identity of the synthesized speech [12].

From the viewpoint of the above-mentioned cross-modality of a human's perception, we can also utilize the facial features of a person for identifying the person, as well as vocal features. Hence, in this paper we aim to introduce facial features to a DNN-based multi-speaker TTS framework that can synthesize any arbitrary speakers' voices using the embedding vector of a speaker. The use of facial features would have many advantages in the practical application of TTS. For instance, we can intuitively identify the speaker for a synthesized voice based on visual information about the speaker, rather than using vocal features that are hard to visualize and taking time to listen to the speaker's voice samples. Besides, facial features can offer a natural means to control speaking style within a single speaker (e.g., emotional TTS [13]). Moreover, once the relationships between vocal and facial features are learned, we can introduce them to other media-related applications such as the voice searching systems [14].

In this paper, we propose a new DNN-based multi-speaker TTS framework named *Face2Speech*, which uses a face image to control the voice characteristics of the synthesized speech. The most difficult point for building such a TTS framework is how we collect training datasets, since there is no one that has a sufficiently large amount of triplets of (text, speech, face image), and constructing such a dataset would be extremely difficult. Therefore, we utilize two kinds of datasets; pairs of (text, speech) and (face image, speech) for separately training three modules in our framework: *a speech encoder*, *a multi-speaker TTS*, and *a face encoder*, as shown in Fig. 1. Firstly, the speech encoder is trained to extract an embedding vector from speech by minimizing the loss function derived from speaker verification. Secondly, the multi-speaker TTS is trained with the pairs of (text, speech) so that the module synthesizes speech from a given text and embedding vector generated from the pre-trained speech encoder. Thirdly, the face encoder is trained with the pairs of (face image, speech) for making an embedding vector extracted from a face image of a speaker closer to one derived from his/her speech. Finally, the face encoder and multi-speaker TTS are concatenated for building our Face2Speech model.

In experimental evaluation, we build our Face2Speech model by using four datasets: VoxCeleb2 [15] and VGGFace2 [16] as pairs of (face image, speech), and VCTK [17] and LibriTTS [18] as pairs of (text, speech). Experimental results demonstrate that our Face2Speech can synthesize speech which is comparable to speech synthesized with an embedding vector extracted from speech corresponding to the given face image.

2. Related Work

Recently, with the emergence of social networking services (e.g., YouTube, Twitter, and Facebook), several attempts have been made to correlate different modalities such as video, images, audio, and text. We especially focus on cross-modal audio/visual learning, especially speech/face generation or con-

http://dx.doi.org/10.21437/Interspeech.2020-2136

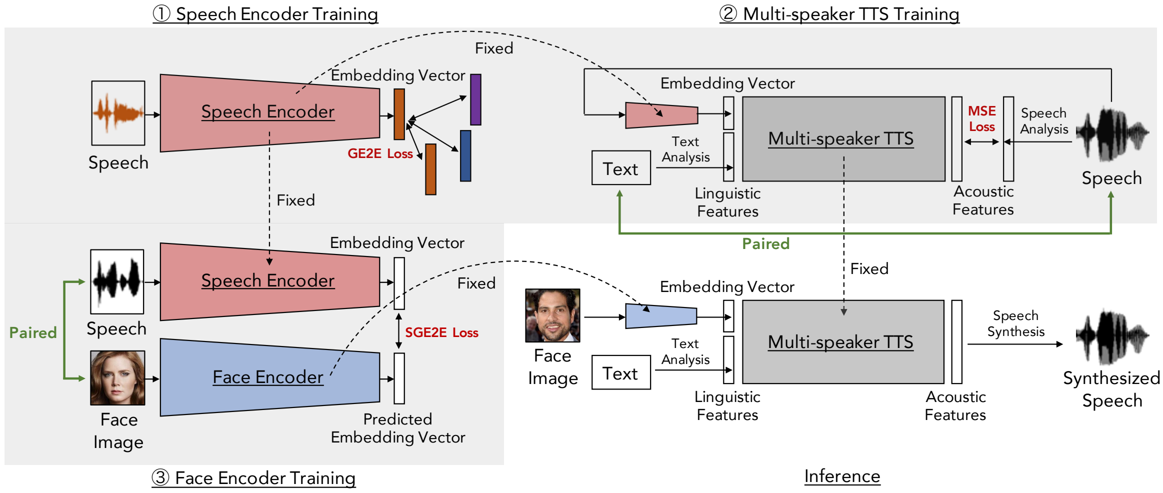

Figure 1: *Overview of Face2Speech. This framework consists of three separately trained modules: 1) speech encoder, 2) multi-speaker TTS, and 3) face encoder. After training, speech can be synthesized from a given text and a face image.*

version. Oh *et al.* [19] proposed a model that maps speech to face by predicting facial features from spectrogram. Ohsugi *et al.* [20] developed a method that converts face to eigenvoice [21] based on subjective impressions. Not only generating one type of data from the other type of data, but conversion from the two types of data (e.g., facial features and acoustic features) of a source speaker to those of a target speaker has been proposed [22].

There have been a few approaches of multi-modal TTS synthesis using text, speech and face. Székely *et al.* [23, 24] developed a TTS system that synthesizes speech in a voice style predicted from facial expressions. The TTS system proposed by Schroeter *et al.* [25] encompasses speech synthesis and face visualization with lips synchronized with the speech. Our work differs from these works in that the Face2Speech model synthesizes speech with characteristics predicted from a face image.

3. Face2Speech Model

As shown in Fig. 1, the three modules in our Face2Speech model are separately trained. Each of the modules: the speech encoder, the multi-speaker TTS, and the face encoder, is trained with speech, (text, speech), and (face image, speech), respectively.

3.1. Speech Encoder

Embedding vectors, which capture the characteristics of speakers, have often been used for speaker verification [26, 27]. The speech encoder in our framework follows a method proposed by Wan *et al* [26]. A log-Mel spectrogram is fed to the speech encoder for extracting an embedding vector. In training the speech encoder, each mini-batch consists of $M \times N$ utterances; each of the N different speakers has M utterances. Let the L2-normalized embedding vector of the jth speaker's ith utterance be $\mathbf{e}_{ji}$ $(1 \leq j \leq N, 1 \leq i \leq M)$. The centroid of the embedding vectors from the jth speaker is defined as $\mathbf{c}_j = \frac{1}{M}\sum_{m=1}^{M} \mathbf{e}_{jm}$. The element of the similarity matrix $\mathbf{S} = (S_{ji,k})_{(N \cdot M) \times N}$ is then defined as a cosine similarity:

$$S_{ji,k} = w \cdot \cos(\mathbf{e}_{ji}, \mathbf{c}_k) + b, \tag{1}$$

where w and b are trainable scalar parameters. By using the similarity matrix $\mathbf{S}$, the generalized end-to-end loss (**GE2E Loss**) function for training the speech encoder $L(\cdot)$ is defined as

$$L(\mathbf{S}) = \sum_{i,j}\left(-\log \frac{\exp(S_{ji,j})}{\sum_{k=1}^{N} \exp(S_{ji,k})}\right). \tag{2}$$

This loss function has an effect of making the cosine similarity of embedding vectors of the same speaker larger and those of the other speakers smaller.

3.2. Multi-speaker TTS

We employ statistical parametric speech synthesis [28] for a multi-speaker TTS model that consists of duration and acoustic models. The duration model outputs the number of frames from the joint vector of linguistic features per phoneme and embedding vectors per utterance, which are obtained by the speech encoder. On the other hand, the acoustic model generates acoustic features (i.e., Mel-cepstral coefficients (MCEPs), log F_0, and an aperiodicity measure) from the joint vector of linguistic features and the embedding vectors frame by frame. Both of the models are trained to minimize the mean squared error (**MSE Loss**) between the target features (i.e., the duration or acoustic feature) and output vectors of the models.

3.3. Face Encoder

We prepare pairs of a face image and speech for training the face encoder. The input for the face encoder is a face image of a speaker and the output is the centroid of the embedding vector extracted from utterances of the speaker. A possible choice for the loss function of the face encoder would be the MSE Loss. However, since the embedding vector of the speech encoder is learned by minimizing the cosine similarities, it is not guaranteed that the face encoder trained to minimize the MSE Loss will output the embedding vector that minimizes the cosine similarities to the embedding vector of a target speaker. Therefore, we employ the loss function for the face encoder similar to Eq. (2). Let M be the number of utterances of the speaker in a mini-batch, and $\tilde{M}$ be the number of all utterances of the speaker. While in the speech encoder, the centroid of the jth speaker is calculated in each mini-batch as $\mathbf{c}_j = \frac{1}{M}\sum_{m=1}^{M} \mathbf{e}_{jm}$, the centroid in the face encoder is calculated from the output of the speech encoder as

$\tilde{\mathbf{c}}_j = \frac{1}{\tilde{M}} \sum_{m=1}^{\tilde{M}} \mathbf{e}_{jm}$, and it is constant in each mini-batch. Therefore, the similarity matrix of the face encoder $\tilde{S}_{ji,k}$ is defined as

$$\tilde{S}_{ji,k} = w \cdot \cos\left(\mathbf{e}_{ji}, \tilde{\mathbf{c}}_k\right) + b. \tag{3}$$

Once given the similarity matrix, the loss is calculated by Eq. (2). Since training a face encoder can be considered as supervised learning, we name the loss function supervised GE2E loss (**SGE2E Loss**).

4. Experiments

4.1. Dataset

VoxCeleb2 [15] and VGGFace2 [16] were used as pairs of face image and speech. VoxCeleb2 is a dataset that contains videos of more than 6,000 celebrities taken from YouTube. VGGFace2 is a large image dataset whose identities overlap with VoxCeleb2. Since we required pairs of a face image and speech to train the face encoder, it would be possible to train the face encoder using pairs of a face image and a corresponding speech which are extracted from a video dataset, i.e., VoxCeleb2. However, because many of the videos we used had low resolution, cropping errors frequently occurred. In addition, since cropped face images lacked fine details of the original face, they were not suitable for feature extraction. We find that the face images from VGGFace2 have higher resolution compared to ones from VoxCeleb2. Therefore, we used face images from VGGFace2, and the speech from VoxCeleb2. We used images and speech of 5,993 speakers for training, and those of 118 speakers for evaluation.

VCTK [17] and LibriTTS [18] were used as pairs of text and speech. We used utterances given by 108 speakers from VCTK and 805 speakers from LibriTTS. 847 speakers were used for training, and 66 speakers for evaluation. Both datasets have cleaner speech than VoxCeleb2. All speech samples were downsampled to 16 kHz.

4.2. Experimental Conditions

4.2.1. Speech Encoder

The length of window, hop size, and FFT were set to 400 (25 ms), 160 (10 ms), and 512 (64 ms) samples, respectively. We used 40-dimensional log Mel-spectrograms as an input and 256-dimensional embedding vectors as an output. The speech encoder consisted of three long short-term memory (LSTM) layers with 768 hidden units and one linear output layer with 256 units. The activation function for hidden layers was tanh. The number of speaker N and that of utterances M in each mini-batch were set to 32 and 4, respectively. The number of training epochs was 500, and the Adam optimizer [29] with its learning rate setting to 10^{-5} was used for the training.

4.2.2. Multi-speaker TTS

The joint vector of 420-dimensional phoneme-wise linguistic features (e.g., phoneme identity and accent type) and 256-dimensional embedding vectors obtained from the speech encoder was fed into the duration model. We used WORLD vocoder [30] (D4C edition [31]) for the acoustic feature extraction. The acoustic features consisted of 40-dimensional MCEPs, log F_0, an aperiodic measure, their dynamic features (i.e., Δ and $\Delta\Delta$), and voiced/unvoiced flag every 5 ms. Namely, the dimensionality of the acoustic features was 127. The joint vector of the linguistic features and the embedding vectors per frame was fed into the acoustic model. The acoustic sequence transitions were generated by utilizing the explicit relations between the static and dynamic features [32]. The WORLD vocoder was used for synthesizing a speech waveform from the generated acoustic features. Inputs of the duration and acoustic models were normalized to be within $[0.01, 0.99]$, and outputs of those were normalized to have zero mean and unit variance using the training data. The embedding vectors were averaged over each utterance. Both of the duration and acoustic models consisted of three bi-directional LSTM layers with 512 hidden units. The activation function for hidden layers was tanh. The number of training epochs was 40, and the Adam optimizer with its learning rate setting to 10^{-4} was used for the training.

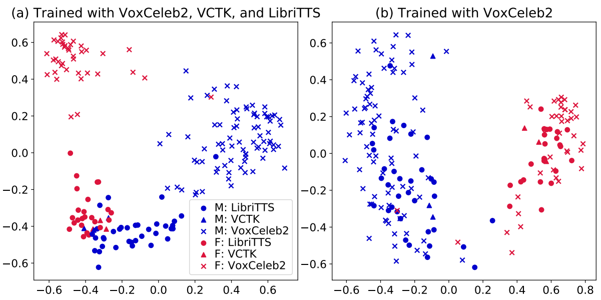

Figure 2: *PCA visualization of embedding vectors extracted from speech encoders trained with (a) VoxCeleb2, VCTK, and LibriTTS or (b) VoxCeleb2.*

4.2.3. Face Encoder

The input size of a face image was scaled to 160 × 160. The value of each pixel is normalized to be within $[-1.0, 1.0]$. Flip transformation was used for data augmentation. We used a pre-trained face detection model[1] for extracting face images from VGGFace2. Images in which the recognition model failed to detect a face were not used. The embedding vectors obtained from the speech encoder were used as an output, which were averaged over all utterances from the same speaker. VGG19 [33] was used as the network architecture. The number of training epochs was 124, and the Adam optimizer with its learning rate setting to 2.0×10^{-3} was used for the training.

4.3. Investigation into effects of datasets on training speech encoder

The speech encoder is a model which maps speech to speaker identity without text. It has been reported that differences of environment among datasets did not unfavorably affect the quality of synthesized speech and similarity between synthesized speech and original speech, even if multiple datasets were used for training the speech encoder [34]. However, in the proposed method, we should pay attention to the use of multiple datasets because the datasets used for training the TTS and the face encoder models are different. In a preliminary experiment, we compared two speech encoders: one was trained with VoxCeleb2, VCTK, and LibriTTS, and the other was trained with VoxCeleb2 only.

Figure 2 shows embedding vectors extracted from speech encoders trained with (a) the above-mentioned three datasets and (b) VoxCeleb2 only. We can observe that embedding vectors extracted by the two encoders have a tendency to construct

[1] https://github.com/ipazc/mtcnn

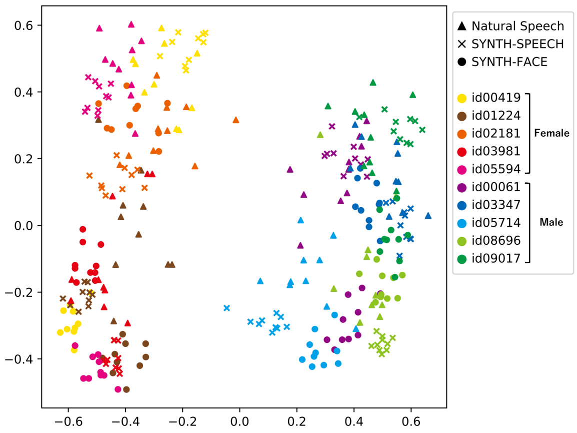

Figure 3: *PCA visualization of the embedding vectors extracted from natural speech, SYNTH-FACE's speech and SYNTH-SPEECH's speech.*

a cluster differentiated by speaker gender. Also, we can infer from Fig. 2(a) that the dataset differences contribute to the cluster construction in the embedding space, in addition to the gender difference. Meanwhile, this tendency does not appear in the embedding space learned by the speech encoder trained with VoxCeleb2 only, as shown in Fig. 2(b). These results indicate that the speech encoder tends to capture not only the speaker identity, but also the dataset identity, which is unfavorable for building the Face2Speech model whose multi-speaker TTS is trained with datasets different from the speech encoder training. Therefore, in the following evaluations, we used VoxCeleb2 for training the speech encoder.

4.4. Evaluation

Since there is no conventional method that can be compared to the proposed method, we prepared two systems in this framework for synthesizing speech: SYNTH-SPEECH and SYNTH-FACE. For evaluation, we used a dataset of test speakers. The difference between the two systems was how to generate embedding vectors. While the embedding vector of SYNTH-SPEECH was made by applying the speech encoder to the utterances of the speaker, that of SYNTH-FACE was generated by applying the face encoder to the face images of the speaker. Since the face encoder was trained to minimize the SGE2E loss between the embedding vector from the speech encoder and output vector from the face encoder, SYNTH-SPEECH can be considered as the upper bound of this framework.

4.4.1. Visualization of embedding space

In this evaluation, we made a PCA visualization of the embedding vectors. Figure 3 shows the visualization of the embedding vectors extracted from natural speech, SYNTH-FACE's speech and SYNTH-SPEECH's speech. Although the points of each speaker's SYNTH-FACE are located not necessarily close to the speaker's SYNTH-SPEECH, the points of SYNTH-FACE are located close to the points of the natural speech or SYNTH-SPEECH of the same gender. In addition, from a qualitative point of view, SYNTH-FACE's synthesized speech is as various as that of SYNTH-SPEECH.

Table 1: *Matching scores on a four-point scale and preference scores of naturalness with 95% confidence intervals. Note that **the lower matching score is the better**, while the higher preference score is the better.*

System	Matching Score	Preference score
SYNTH-FACE	2.01 ± 0.07	0.548 ± 0.049
SYNTH-SPEECH	1.91 ± 0.06	0.452 ± 0.049

4.4.2. Matching evaluation

We evaluated how well the synthesized speech using an embedding vector extracted from a face image of the speaker matches the face image. We conducted a matching evaluation test. We employed the same systems for synthesizing speech as described in Section 4.4, and 30 subjects participated in each evaluation. We prepared pairs of a person's face image and a speech sample with an embedding vector extracted from the face image or speech utterances of the person. Each subject was given 20 pairs of samples, and asked to score how well the synthesized speech matches the corresponding face image. The score is on a scale of 1 to 4: 1) Match well, 2) Match moderately, 3) Match slightly, and 4) Not match.[2] Following [35], we employed this scoring metric.

The second column of Table 1 shows the result of matching evaluation. We observed that the score of SYNTH-FACE was slightly higher than SYNTH-SPEECH, but there was no significant difference between the two systems. This result indicates that as far as humans can recognize, synthesized speech made with the face-derived embedding vectors matches the presented face as well as speech made with the speech-derived embedding vectors.

4.4.3. Naturalness evaluation

We conducted a preference AB test to evaluate the naturalness of the synthesized speech. 30 listeners participated in the evaluation. Each listener evaluated 10 pairs of speech samples generated by the two systems, and chose the natural-sounding one.

The third column of Table 1 shows the result of the naturalness evaluation. We found that the score of SYNTH-FACE was not significantly different from that of SYNTH-SPEECH. The results demonstrate that by using only one face image, we can generate speech as natural as the speech generated from voice samples, which will lead to practical application of TTS.

5. Conclusion

We proposed a DNN-based multi-speaker TTS framework named Face2Speech, which used a single face image to control the speaker identity of synthesized speech. Experimental results of matching and naturalness tests demonstrated that the Face2Speech model was comparable to a multi-speaker TTS model using a speech-derived embedding vector. In future work, we need to make clear how much subjective matching evaluation and speaker characteristics are correlated, and make detailed investigation of objective evaluation. In addition, although we used a WORLD vocoder for speech synthesis, we need to investigate the effect of incorporating a neural vocoder, because it is expected to produce better quality speech and to make it easier to recognize differences between speakers.

[2] Our samples are provided at https://dena.github.io/Face2Speech/

6. References

[1] H. M. J. Smith, A. K. Dunn, T. Baguley, and P. C. Stacey, "Concordant cues in faces and voices: Testing the backup signal hypothesis," *Evolutionary Psychology*, vol. 14, no. 1, 2016.

[2] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," in *Proc. ICASSP*, New York, U.S.A., Apr. 1988, pp. 679–682.

[3] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, Vancouver, Canada, May 2013, pp. 7962–7966.

[4] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: a generative model for raw audio," *arXiv*, vol. abs/1609.03499, 2016.

[5] J. Shen, R. Pang, R. J. Weiss, M. S. annd N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 4779–4783.

[6] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 3067–3071.

[7] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. ICLR*, Banff, Canada, Apr. 2014.

[8] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R.-A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *arXiv*, vol. abs/1803.09017, 2018.

[9] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang, "Hierarchical generative modeling for controllable speech synthesis," in *Proc. ICLR*, New Orleans, U.S.A., May 2019.

[10] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 4080–4084.

[11] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 5329–5333.

[12] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," in *Proc. SSW*, Vienna, Austria, Sep. 2019, pp. 155–160.

[13] M. Schröder, "Emotional speech synthesis: a review," in *Proc. Eurospeech*, Aalborg, Denmark, Sep. 2001, pp. 561–564.

[14] V. Vestman, B. Soomro, A. Kanervisto, V. Hautamäki, and T. Kinnunen, "Who do I sound like? showcasing speaker recognition technology by YouTube voice search," in *Proc. ICASSP*, Brighton, U.K., May 2019, pp. 5871–5785.

[15] J.-S. Chung, A. agrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 1086–1090.

[16] Q. Cao, L. Shen, W. Xie, O.-M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. FG*, Xi'an, China, 2018, pp. 67–74.

[17] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[18] H. Zen, V. Dang, R. Clark, Y. Zhang, R.-J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: a corpus derived from LibriSpeech for text-to-speech," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 1526–1530.

[19] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, "Speech2Face: Learning the face behind a voice," in *Proc. CVPR*, Long Beach, U.S.A., June 2019, pp. 7539–7548.

[20] Y. Ohsugi, D. Saito, and N. Minematsu, "A comparative study of statistical conversion of face to voice based on their subjective impressions." in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 1001–1005.

[21] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, Nov. 2000.

[22] F. Fang, X. Wang, J. Yamagishi, and I. Echizen, "Audiovisual speaker conversion: jointly and simultaneously transforming facial expression and acoustic characteristics," in *Proc. ICASSP*, Brighton, U.K., May 2019, pp. 6795–6799.

[23] É. Székely, Z. Ahmed, S. Hennig, J.-P. Cabral, and J. Carson-Berndsen, "Predicting synthetic voice style from facial expressions. an application for augmented conversations," *Speech Communication*, vol. 57, pp. 63–75, Feb. 2014.

[24] É. Székely, Z. Ahmed, J.-P. Cabral, and J. Carson-Berndsen, "WinkTalk: a demonstration of a multimodal speech synthesis platform linking facial expressions to expressive synthetic voices," in *Proc. SLAPT*, Montréal, Canada, June 2012, pp. 5–8.

[25] J. Schroeter, J. Ostermann, H.-P. Graf, M. Beutnagel, E. Cosatto, A. Syrdal, A. Conkie, and Y. Stylianou, "Multimodal speech synthesis," in *Proc. ICME*, New York, U.S.A., July 2000, pp. 571–574.

[26] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 4879–4883.

[27] L. Chao, M. Xiaokong, J. Bing, L. Xiangang, Z. Xuewei, L. Xiao, C. Ying, K. Ajay, and Z. Zhenyao, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv*, vol. abs/1705.02304, 2017.

[28] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[29] D.-P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proc. ICLR*, San Diego, U.S.A., May 2015.

[30] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, July 2016.

[31] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, Nov. 2016.

[32] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, Istanbul, Turkey, June 2000, pp. 1315–1318.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv*, vol. abs/1409.1556, 2014.

[34] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proc. NIPS*, Montréal, Canada, Dec. 2018, pp. 4480–4490.

[35] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proc. ACMMM*, Mountain View, U.S.A., Oct. 2017, pp. 349–357.