In: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), December 2021

HEARING FACES: TARGET SPEAKER TEXT-TO-SPEECH SYNTHESIS FROM A FACE

Björn Plüster, Cornelius Weber, Leyuan Qu, Stefan Wermter

Knowledge Technology, Department of Informatics, University of Hamburg
bjoern.pluester@studium.uni-hamburg.de, {weber, qu, wermter}@informatik.uni-hamburg.de

ABSTRACT

The existence of a learnable cross-modal association between a person's face and their voice is recently becoming more and more evident. This provides the basis for the task of target speaker text-to-speech (TTS) synthesis from face reference. In this paper, we approach this task by proposing a cross-modal model architecture combining existing unimodal models. We use *Tacotron 2* multi-speaker TTS with auditory speaker embeddings based on *Global Style Tokens*. We transfer learn a *FaceNet* face encoder to predict these embeddings from a static face image reference instead of a voice reference and thus predict a speaker's voice and speaking characteristics from their face. Compared to Face2Speech, the only existing work on this task, we use a more modular architecture that allows the use of openly available and pretrained model components. This approach enables high-quality speech synthesis and allows for an easily extensible model architecture. Experimental results show good matching ability while retaining better voice naturalness than *Face2Speech*. We examine the limitations of our model and discuss multiple possible avenues of improvement for future work.

Index Terms: multi-speaker text-to-speech synthesis, speaker embedding, cross-modal learning, transfer learning

1. INTRODUCTION

Recent multi-speaker TTS models such as *Tacotron 2* [1], *Deep Voice 2* [2] and more recently *FastSpeech2* [3] have shown the ability to generate speech with near-human naturalness and fidelity. Some models have also shown the ability to model different vocal characteristics which reflect speaker identities by using voice embeddings [2, 4]. These embeddings are extracted from voice reference. In this paper, we propose a method of approximating such an embedding from face reference, building on recent findings in cross-modal studies.

When shown an image of a person's face, humans tend to have a preconception of the sound of their voice. Early studies on cross-modal association found that humans could identify the corresponding voice to dynamic facial data with greater than chance accuracy [5, 6] but failed, however, to show the same ability on static facial data. More recently, though, static face images have been shown in human studies to carry significant cross-modal information [7, 8] and the ability of machines to work with this information has been demonstrated. Nagrani *et al.*[9] present models performing well on cross-modal matching tasks from face to voice and from voice to face and Kim *et al.*[10] present a learned cross-modal representation exhibiting a matching capability similar to that of humans.

In this work, we aim to use models widely used and proven in their performance in their field and train them in a way that enables the aforementioned cross-modal information transfer. We use existing unimodal models as components of our multimodal architecture, aiming to benefit from the advances in their respective fields. Our approach uses the *Global Style Token* architecture [4] for representing speaker style from auditory inputs as a teacher model to transfer learn a deep convolutional neural network to represent speaker style from a face image input. The predicted *style embedding* is used by a TTS system, in our case *Tacotron 2* [1] to synthesize speech in a target speakers voice.

We propose an architecture that has a modular structure and we demonstrate its functionality with a specific choice of components. Due to the modular nature of this architecture and the use of the *ESPNet-TTS* toolkit [11, 12], each of these components can be readily exchanged for others. More specifically, the face encoder, the TTS model, and the vocoder can be exchanged. We provide a non-exhaustive list of explicit examples of other options but limit our experiments to one specific setup as a proof of concept.

- The face encoder can be replaced by any method that accepts a face image input and outputs an embedding vector such as FaceNet [13] (InceptionResNetV1 [14]), VarGFaceNet [15], and LightFace [16].

- Tacotron 2 with GST (the TTS model) can be exchanged with any that are available in the ESPnet toolkit in a configuration that uses GSTs. The list of available models [17] includes Tacotron 2 [1], FastSpeech2 [3], and Transformer TTS [18] among others.

- The choice of vocoder can also affect the performance. Some suggested options are ParallelWaveGAN [19], MelGAN [20], and WaveGlow [21].

Target speaker speech synthesis from face reference has interesting potential applications, such as the approximation of voices of historical figures of whom no recordings exist, which could be applicable in museums or other learning environments. It may also be interesting within the context of animation to fit a voice to the created character. Further, it may be a valuable method for controlling TTS systems' voice or perhaps even emotion.

Training models for such a task requires a large and diverse dataset of triplets of voice recordings, face images, and word-level text transcriptions. Voice audio and transcription tuples are required for finetuning the TTS system, and voice audio and face image tuples are necessary for training and transfer learning a face recognition model. We generate an appropriate dataset from the Lip Reading Sentences 3 (LRS3) dataset [22]. It consists of short, transcribed, and face-aligned videos from many different TED and TEDx talks and was initially created for a lip-reading task.

Experimental results show our model's ability to predict the voice to a given face while retaining good perceptual speech quality and faster than real-time inference speeds. We discuss the limitations of our proposed architecture and suggest possible improvements that could be achieved by extending our work.

2. RELATED WORK

Cross-modal learning on face and voice data is a recently emerging topic. Visual generation from auditory inputs has shown impressive results with *WAV2PIX* by Duarte *et al.*[23] and *Speech2Face* by Oh *et al.*[24] both using intermediate embedding representations of the voice for face generation. Fang *et al.*[25] propose a method for simultaneous audio and image generation from audio-visual inputs, allowing generated facial expressions and voice to be highly correlated.

Cross-modal information gain has also been successfully employed in target speaker speech separation with Afouras *et al.* [26] initially using data from lip movement to performance and Ephrat *et al.*[27] exploring the use of a full face embedding. More recently, Qu *et al.*[28] show that using a static, pre-enrolled face embedding also benefits speech separation.

Recently, Goto *et al.*[29] first addressed target speaker speech synthesis from static face reference. They propose *Face2Speech*, a deep neural network consisting of a speech encoder, a face encoder, and a multi-speaker TTS system. They employ an DNN-based statistical parametric speech synthesis (SPSS)[30] approach for their multi-speaker TTS and use the WORLD vocoder [31] for waveform synthesis. While SPSS is robust and has good inference performance, its main drawbacks lie in the quality of the produced speech and in that it requires expert knowledge in modeling [30]. We aim to improve perceptual speech quality in this task by using more recent methods for multi-speaker TTS and waveform synthesis. We compare our experimental results to *Face2Speech* in section 4. At the time of writing, no other previous works cover this specific task.

3. PROPOSED MODEL

Figure 1 details our proposed architecture. We finetune a face encoder in a supervised manner to predict the target *style embedding* from face reference. The *Global Style Token* [4] module acts as a teacher during this step. During inference, this predicted style embedding is passed to *Tacotron 2* [1] for Mel spectrogram synthesis. ParallelWaveGAN [19] finally converts this spectrogram to a waveform.

3.1. Speech Encoding and Synthesis

The *Tacotron 2* model with *Global Style Tokens* (GSTs) is a network trained for multi-speaker text-to-speech synthesis with voice reference [4]. It uses a 512-dimensional intermediate representation (*style embedding*) to model speaker voice and prosody, which is produced by the *style token layer*. This embedding is passed to the decoder along with the encoded text features to predict a Mel spectrogram finally used for waveform synthesis by the vocoder. The model we use is a PyTorch implementation in the *ESPNet-TTS* toolkit [11, 12], in particular a configuration pretrained on the VCTK Corpus [32].

Tacotron 2 is a neural network architecture composed of a recurrent sequence-to-sequence network for Mel spectrogram prediction from a sequence of character embeddings consisting of a text encoder and a decoder, and a modified WaveNet [33] vocoder predicting a waveform from the spectrogram. In this work, we replace the vocoder with a Pytorch implementation of ParallelWaveGAN [19, 34] pretrained on the VCTK Corpus [32]. It is a state-of-the-art generative adversarial network-based waveform prediction network offering faster inference speeds and good perceptual quality.

We choose Tacotron 2 with GST as a multi-speaker TTS architecture in contrast to other, more recent, architectures because the use of GSTs with Tacotron 2 is well documented. As stated earlier, replacing Tacotron 2 with another TTS system that implements the GST architecture is possible and straightforward. This poses a significant advantage of the proposed architecture and enables the simple extension of this architecture in future work. This is further discussed in section 5.

3.1.1. Text Encoder

First, text input is preprocessed. We remove unwanted symbols, expand known abbreviations, and convert to phonetic tokens representing pronunciation using g2pE [35]. The text encoder maps the sequence of tokens to a sequence of embeddings. These are then passed through 3 convolutional layers

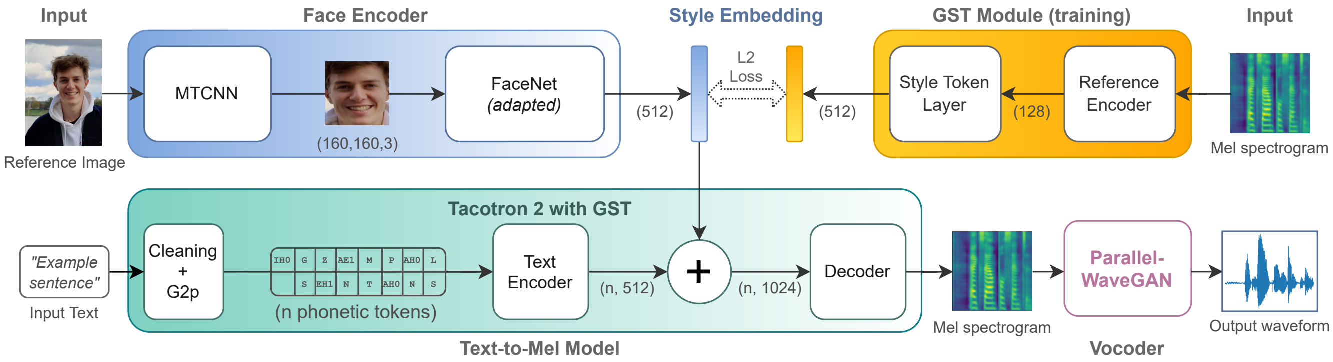

Fig. 1: Overview of the proposed architecture. The face encoder receives an image and predicts a style embedding for the Tacotron 2 with GST model to predict a Mel spectrogram in the target voice. The vocoder predicts a waveform from the spectrogram. During training, the GST module from voice reference acts as a teacher to the face encoder.

of 512 filters of shape 5 × 1, each with batch normalization and ReLU activation, and then into a 512-unit bidirectional LSTM layer.

3.1.2. *Style Embedding*

The Global Style Token model as proposed by Wang *et al.*[4], is a network which predicts a *style embedding* from acoustic input to represent the reference speaker's vocal characteristics. It is jointly trained with Tacotron 2 based on its reconstruction loss.

The first part of the GST module is a reference encoder that takes a log-Mel filterbank as input and extracts a 128-d reference embedding. The reference encoder consists of six stacked 2-d convolutional layers with batch normalization [36] and ReLU activation, followed by a single-layer 128 unit unidirectional GRU [37]. The style token layer which follows is a bank of 128 randomly initialized but fixed 512-d style tokens and a multi-head attention mechanism [38] with 8 attention heads. It takes the reference embedding as input and outputs a 512-d style embedding, a combination of the tokens most relevant to different aspects of the reference, effectively giving a weighted average of the style tokens in the bank. Each token captures some speech attributes such as speaking rate or pitch. We use the style embedding as a target vector to train the face encoder.

The log-Mel filterbank is extracted from the 24 kHz input waveform using first a Short Time Fourier Transform (STFT) with 300 sample hop length (12.5 ms) and 1200 sample window length (50ms) zero-padded to length 2048 (85.33ms). This STFT is then projected to a Mel filterbank with 80 Mel bands and frequencies from 80 to 7600 Hz and then scaled logarithmically with base 10. Finally, the extracted filterbank is normalized by global mean and variance normalization based on the training data.

3.1.3. *Decoder and Vocoder*

The style embedding is appended to each encoded character of the input sequence. Then, this sequence is passed to the decoder. The decoder is an autoregressive RNN for Mel spectrogram prediction and is implemented following Shen *et al.*[1]. It uses a location-sensitive attention mechanism [39] to attend to the input sequence as well as previous time steps. The final Mel spectrogram then is passed to ParallelWaveGAN, which generates a 24 kHz waveform from the input.

3.2. Face Encoder

The face encoder we use here is FaceNet [40, 13], a Pytorch implementation of the Inception-ResNet-v1 architecture [14]. The model is pretrained on the VGGFace2 [41] dataset. It is prefaced by a multi-task cascaded neural network (MTCNN) [42] for face detection and alignment. We crop input images to the face region with a margin of 10 px and resize to 160 × 160 px.

We first train the model on a face recognition task on our dataset and then finetune with the style embeddings extracted from the images' corresponding utterances as target values. After this transfer learning step, the adapted FaceNet replaces the GST module.

4. EXPERIMENTS

The dataset we use is generated based on the Lip Reading Sentences 3 (LRS3) [22] dataset, which consists of thousands of videos from TED and TEDx talks, each transcribed at the word level. This gives the basis for an audio-visual dataset with word transcription, satisfying Tacotron 2 finetuning, FaceNet training, and FaceNet transfer learning requirements. The dataset has a training, validation, and a test set. The training and validation set have some speaker identity overlaps, and the test set is entirely disjoint. We do not use the same dataset as Goto *et al.*[29] because text transcriptions, which are not available in their proposed dataset, are required for Tacotron 2 training.

After audio extraction and upsampling from 16 kHz to 24 kHz, we split the audio files into equal-length segments. All segments are around five seconds in length with a minimum of three seconds and a maximum of eight seconds. We make sure not to split spoken words by only splitting before or after each word to avoid misaligning the transcriptions. Then a random frame is extracted from the corresponding video for each audio file and is preprocessed with MTCNN. Audio and image extraction, as well as audio splitting and upsampling, was done using FFmpeg.

The final training set has $\sim$260k triplets of utterances, transcriptions, and images by 5089 speakers. The validation set has $\sim$32k such triplets by 4004 speakers and the test set has 1321 triplets by 412 speakers. We perform all training on 2x NVIDIA GTX 1070 GPUs with 8G memory each.

4.1. Tacotron 2 Finetuning

The model we use was pretrained on the VCTK Corpus [32] which consists of high-quality, professionally recorded utterances from 110 English speakers with various accents. Conversely, the audio of our generated dataset contains noise and reverberation. The voice generated by the pretrained model when using audio from this dataset as reference did not match the reference voice well, meaning that finetuning was necessary. Finetuning the full model led to a degeneration of speech quality but tuning only the GST module while keeping encoder and decoder frozen showed better results. We discuss these issues and resulting limitations in section 5.

We train Tacotron 2 on pairs of voice audio and text transcription. The loss function is an equally weighted sum of L1 and MSE loss on the generated Mel spectrogram. We use the Adam optimizer with a learning rate of 0.001. We train for two epochs of 16000 iterations with a batch size of 3.8M elements, where each element is a single Mel filterbank frame of the preprocessed input audio.

4.2. FaceNet Training

We train FaceNet on a facial recognition classification task with a linear layer appended to the final latent layer, making the output dimensional equal to the number of classes or in this case speakers. We start with a pretrained model to aid in convergence. We train for 15 epochs with 256 batch size using cross-entropy loss and the Adam optimizer with a learning rate of 0.001, reduced by a factor of 0.1 at epoch 5.

As the goal is to predict the style embedding corresponding to a face image, we prepare data tuples of face image and style embedding produced by the finetuned GST module given the corresponding utterance, effectively using the GST model as a teacher model to FaceNet. We then freeze all parameters of FaceNet except for the final linear layer producing the embedding and train for a total of 15 epochs. We use MSE Loss, a batch size of 512 and Adam optimizer with a learning rate of 0.01, reduced by a factor of 0.1 at epochs 5 and 10.

Table 1: Evaluation results comparing matching scores and naturalness of our model with *Face2Speech*, each with 95% confidence intervals. Matching scores are on a scale of 1 to 4 and naturalness is on a scale of 1 to 5.

Model	Matching ↓	Naturalness (MOS) ↑
Face2Speech	$\mathbf{2.01 \pm 0.07}$ [2]	3.50 ± 0.08
Ours	2.35 ± 0.08	$\mathbf{3.69 \pm 0.07}$

4.3. Evaluation

We evaluate the performance of the presented method by conducting two surveys and comparing the results with those presented by Goto *et al.*[29]. The first is a measure of how well the synthesized voice matches the reference face and the second survey is a mean opinion score (MOS) evaluation of voice naturalness. We also look at the style embeddings generated from face reference compared to those from voice reference. We conducted all surveys using Amazon Mechanical Turk. Samples shown to participants are generated from the test set and are available in our web demo[1].

It should be noted that this comparison is to be taken with a grain of salt, as the models were trained and evaluated on different datasets. While our model is trained on a dataset derived from the LRS3 dataset, Face2Speech was trained on a mixture of VoxCeleb and VGGFace2 datasets. We did not attempt to train Face2Speech on our dataset because the code was not published.

4.3.1. Matching Evaluation

To evaluate how well our method performs, we conduct a survey with 30 participants following [29]. Each participant is shown 20 samples of a face and the corresponding voice synthesized from that face reference. They are then asked to rate how well the speech matches the corresponding face image on a scale of 1-4: *1: Match well, 2: Match moderately, 3: Match slightly, and 4: Do not match.*

The evaluation results are visible in the first column of Table 1. We present the matching score of Face2Speech from [29]. With a mean score of 2.38, our model fails to outperform the *Face2Speech* model but shows its ability to produce a voice matching to a face moderately well. We suspect this ability can be significantly improved within our proposed architecture and discuss possibilities in section 5.

[1] `https://bjoernpl.github.io/FaceTTS/`

[2] Matching score taken from evaluation by Goto *et al.* in [29]

4.3.2. *Naturalness Evaluation*

We evaluate the naturalness of speech produced by conducting a survey asking participants to rate how natural (human-like) a sample sounds. Goto *et al.*[29] perform a naturalness evaluation between their systems using a preference AB test. While this gives an idea of their systems' relative naturalness, it does not allow us to compare. To facilitate comparison, we conduct a mean opinion score (MOS) evaluation of naturalness for both our model and theirs using samples from their demo. Thirty participants are shown 20 samples each of synthesized speech and are tasked with rating the naturalness of the voice on a scale of 1 to 5, from which we calculate the MOS.

The results, visible in the second column of Table 1, show that speech synthesized by our proposed model is more natural than by *Face2Speech* by a significant margin of about 0.2. This indicates that a cross-modal information transfer is possible while keeping the qualitative benefits of the Tacotron architecture.

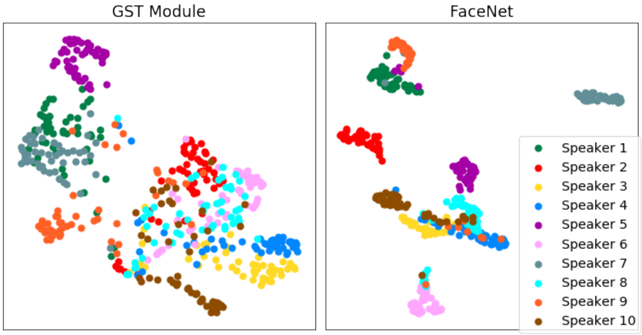

Fig. 2: UMAP [43] dimensionality reduction showing style embeddings generated from voice reference with GST module (left) and embeddings from face reference with FaceNet (right).

4.3.3. *Style Embedding Space*

Figure 2 shows style embeddings with their dimensionality reduced using UMAP [43] extracted from voice reference (left) and face reference (right) for ten different speakers. We use the validation set here since each speaker in the test set only has a few samples making visualization less clear.

The figure indicates a disparity between the two modalities. For speakers 3, 4, and 10, auditory similarity translates well to facial similarity, visible by the low distances between each speaker's embeddings. Speakers 1 and 7, for example, do not show this property. While their embeddings are very close to each other from auditory reference, they are distant when generated from face reference. The left plot also hints at the GST model's limitations, where for some speakers or samples it fails to produce an embedding distinctly characterizing the speaker. This is most evident for speaker 8 whose embeddings are scattered throughout the embedding space.

5. DISCUSSION AND FUTURE WORK

This work demonstrates that predicting a style embedding from face reference is possible and can result in a generated voice of higher naturalness than achieved in previous work. The match to an individuals' vocal characteristics however, is less pronounced than with Face2Speech. The cause may lie in how the Tacotron 2 with GST model was trained and in its configuration. The VCTK Corpus[32] used for training consists of only 110 speakers, all reading similar texts. While there is variation in the speakers' accents and the sample size is relatively large, a higher number of more varied speakers could lead to a more diverse set of voices learned by the system. Wang *et al.*[4] discuss this idea, showing that it is possible to learn from a more diverse dataset but that a larger bank of style tokens is necessary to capture the greater variance. Performing a complete training run on the dataset with a larger bank of style tokens would most likely lead to much better results but was infeasible to us due to limitations in computational resources. We suspect that the bank of 128 style tokens was not large enough and estimate that a bank of at least 1024 tokens may be more appropriate. The consequence of this is visible in Figure 2 where voice embeddings are not clustered well for each speaker, hinting that the GST module is not correctly predicting a proper style embedding. Wang *et al.*[4] show an apparent clustering for style embeddings from the same speaker.

Another main factor limiting the performance of this approach also lies within the dataset. While Tacotron 2 and its GST module are trained on clean, professionally recorded audio, the audio of the LRS3 dataset is very noisy. Each recording is recorded in different settings, with different equipment and varying levels of environmental noise. First indicated by Wang *et al.*[4], when the GST module is confronted with noise in the training dataset, it learns to assign some of the style tokens in the bank to model this noise. This leaves a smaller range of style tokens to model a diverse range of voices, leading to a less diverse output. A countermeasure to this problem may be to denoise the audio before training. This could be done based on the audio (e.g. [44, 45, 46]) or in the case of LRS3 also by methods that denoise using audiovisual input (e.g. [47, 48]).

As mentioned in the introduction, our proposed architecture is intentionally modular. Significant performance and efficiency improvements may be possible by employing different models for face encoder, TTS, and vocoder. Also, we hypothesize that introducing the notion of style tokens in form of an adapted style token module to the face encoder may aid performance. If this facial GST module shares the same bank of style tokens as its auditory counterpart, it will explicitly model the same embedding space leading to greater accuracy in voice style approximation.

Finally, it may also be of interest to investigate a model similar to the proposed but to use an x-vector [49] intermediate representation instead of the GSTs to focus more on speaker identity and less on prosody and speaking style.

6. CONCLUSION

We propose a modular architecture for target speaker text-to-speech synthesis with face image reference, reusing an available pretrained Tacotron 2 with GST model in a cross-modal transfer learning task. Experimental results show that prediction of style embeddings from a face reference is possible and that we achieve unprecedented naturalness in this task. Comparison with Face2Speech shows a slightly reduced matching ability but greater speech naturalness. We discuss ideas for future improvements which may allow more accurate matching and better capturing of vocal identity. These include performing a complete training run with the Tacotron 2 and GST model on our dataset for improved style embedding extraction or making use of the proposed architecture's modularity by exchanging components.

7. ACKNOWLEDGEMENTS

The authors gratefully acknowledge partial support from the DFG under project CML (TRR 169).

8. REFERENCES

[1] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) - Proceedings*, vol. 2018-April, pp. 4779–4783, 2018.

[2] Sercan O. Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou, "Deep Voice 2: Multi-speaker neural text-to-speech," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, pp. 2963–2971, 2017.

[3] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," *arXiv preprint arXiv:2006.04558*, pp. 1–11, 2020.

[4] Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *International Conference on Machine Learning*, vol. 12, pp. 8229–8238, 2018.

[5] Miyuki Kamachi, Harold Hill, Karen Lander, and Eric Vatikiotis-Bateson, "'Putting the Face to the Voice': Matching Identity across Modality," *Current Biology*, vol. 13, no. 19, pp. 1709–1714, 2003.

[6] Lorin Lachs and David B. Pisoni, "Crossmodal Source Identification in Speech Perception," *Ecological Psychology*, vol. 16, no. 3, pp. 159–187, 2004.

[7] Lauren W. Mavica and Elan Barenholtz, "Matching Voice and Face Identity from Static Images," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 39, no. 2, pp. 307–312, 2013.

[8] Harriet M.J. Smith, Andrew K. Dunn, Thom Baguley, and Paula C. Stacey, "Matching novel face and voice identity using static and dynamic facial images," *Attention, Perception, and Psychophysics*, vol. 78, no. 3, pp. 868–879, 2016.

[9] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman, "Seeing Voices and Hearing Faces: Cross-Modal Biometric Matching," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 8427–8436, 2018.

[10] Changil Kim, Hijung Valentina Shin, Tae Hyun Oh, Alexandre Kaspar, Mohamed Elgharib, and Wojciech Matusik, "On Learning Associations of Faces and Voices," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11365 LNCS, no. 1, pp. 276–292, 2019.

[11] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, "ESPNet: End-to-end speech processing toolkit," in *Proceedings of the Annual Conference of the International Speech Communication Association*, vol. 2018-Septe, pp. 2207–2211.

[12] Tomoki Hayashi, Ryuichi Yamamoto, Katsuki Inoue, Takenori Yoshimura, Shinji Watanabe, Tomoki Toda, Kazuya Takeda, Yu Zhang, and Xu Tan, "Espnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) - Proceedings*, 2020, pp. 7654–7658.

[13] Tim Esler, "FaceNet-Pytorch," [Online]. Available: `https://github.com/timesler/facenet-pytorch`, 2020.

[14] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," *31st AAAI Conference on Artificial Intelligence*, pp. 4278–4284, 2017.

[15] Mengjia Yan, Mengao Zhao, Zining Xu, Qian Zhang, Guoli Wang, and Zhizhong Su, "VarGFaceNet: An efficient variable group convolutional neural network for lightweight face recognition," *Proceedings - International Conference on Computer Vision Workshop*, pp. 2647–2654, 2019.

[16] Sefik Ilkin Serengil and Alper Ozpinar, "LightFace: A Hybrid Deep Face Recognition Framework," in *Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2020, pp. 23–27.

[17] Tomoki Hayashi, "ESPNet Model Zoo," [Online]. Available: `https://github.com/espnet/espnet_model_zoo`, 2020.

[18] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu, "Neural Speech Synthesis with Transformer Network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6706–6713, 2019.

[19] Ryuichi Yamamoto, Eunwoo Song, and Jae Min Kim, "Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) - Proceedings*, vol. 2020-May, pp. 6199–6203, 2020.

[20] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," in *Advances in Neural Information Processing Systems*, H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox, and R Garnett, Eds. 2019, vol. 32, Curran Associates, Inc.

[21] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "Waveglow: A Flow-based Generative Network for Speech Synthesis," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) - Proceedings*, vol. 2019-May, pp. 3617–3621, 2019.

[22] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[23] Amanda Duarte, Francisco Roldan, Miquel Tubau, Janna Escur, Santiago Pascual, Amaia Salvador, Eva Mohedano, Kevin McGuinness, Jordi Torres, and Xavier Giro-I-Nieto, "Wav2Pix: Speech-conditioned Face Generation Using Generative Adversarial Networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) - Proceedings*, vol. 2019-May, pp. 8633–8637, 2019.

[24] Tae Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T. Freeman, Michael Rubinstein, and Wojciech Matusik, "Speech2Face: Learning the Face Behind a Voice," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 7531–7540, 2019.

[25] Fuming Fang, Xin Wang, Junichi Yamagishi, and Isao Echizen, "Audiovisual Speaker Conversion: Jointly and Simultaneously Transforming Facial Expression and Acoustic Characteristics," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) - Proceedings*, vol. 2019-May, pp. 6795–6799, 2019.

[26] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "The conversation: Deep Audio-Visual Speech Enhancement," in *Proceedings of the Annual Conference of the International Speech Communication Association*, vol. 2018-Septe, pp. 3244–3248.

[27] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein, "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation," *ACM Transactions on Graphics*, vol. 37, no. 4, 2018.

[28] Leyuan Qu, Cornelius Weber, and Stefan Wermter, "Multimodal Target Speech Separation with Voice and Face References," in *Proceedings of the Annual Conference of the International Speech Communication Association*, vol. 2020-Octob, pp. 1416–1420.

[29] Shunsuke Goto, Kotaro Onishi, Yuki Saito, Kentaro Tachibana, and Koichiro Mori, "Face2Speech: Towards Multi-Speaker Text-to-Speech Synthesis Using an Embedding Vector Predicted from a Face Image," *Proceedings of the Annual Conference of the International Speech Communication Association*, vol. 2020-Octob, pp. 1321–1325.

[30] Heiga Zen, Keiichi Tokuda, and Alan W Black, "Statistical Parametric Speech Synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[31] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[32] Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, and Others, "CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)," 2019.

[33] 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.

[34] Tomoki Hayashi, "Parallel WaveGAN implementation with Pytorch," [Online]. Available: `https://github.com/kan-bayashi/ParallelWaveGAN`, 2020.

[35] Park Kyubyong and Jongseok Kim, "g2pE," [Online]. Available: `https://github.com/Kyubyong/g2p`, 2019.

[36] Sergey Ioffe and Christian Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *32nd International Conference on Machine Learning*, vol. 1, pp. 448–456, 2015.

[37] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, pp. 5999–6009.

[39] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," *Advances in Neural Information Processing Systems*, vol. 2015-Janua, pp. 577–585.

[40] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, pp. 815–823, 2015.

[41] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 67–74, 2018.

[42] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[43] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger, "UMAP: Uniform Manifold Approximation and Projection," *The Journal of Open Source Software*, vol. 3, no. 29, pp. 861, 2018.

[44] François G. Germain, Qifeng Chen, and Vladlen Koltun, "Speech denoising with deep feature losses," in *Proceedings of the Annual Conference of the International Speech Communication Association*, vol. 2019-Septe, pp. 2723–2727.

[45] Ruilin Xu, Rundi Wu, Yuko Ishiwaka, Carl Vondrick, and Changxi Zheng, "Listening to Sounds of Silence for Speech Denoising," in *Advances in Neural Information Processing Systems*, H Larochelle, M Ranzato, R Hadsell, M F Balcan, and H Lin, Eds. 2020, vol. 33, pp. 9633–9648, Curran Associates, Inc.

[46] Shang Yi Chuang, Yu Tsao, Chen Chou Lo, and Hsin Min Wang, "Lite Audio-Visual Speech Enhancement," in *Proceedings of the Annual Conference of the International Speech Communication Association*, vol. 2020-Octob, pp. 1131–1135.

[47] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg, "Visual Speech Enhancement," in *Proceedings of the Annual Conference of the International Speech Communication Association*, vol. 2018-Septe, pp. 1170–1174.

[48] Madhav Mahesh Kashyap, Anuj Tambwekar, Krishnamoorthy Manohara, and S Natarajan, "Speech Denoising without Clean Training Data: a Noise2Noise Approach," *arXiv preprint arXiv:2104.03838*, 2021.

[49] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) - Proceedings*, vol. 2018-April, pp. 5329–5333, 2018.