



FVTTS: Face Based Voice Synthesis for Text-to-Speech

Minyoung Lee¹, Eunil Park^{1,2,*}, Sungeun Hong²

¹Department of Applied Artificial Intelligence, Sungkyunkwan University, Seoul 03063, Korea

²Department of Immersive Media Engineering, Sungkyunkwan University, Seoul 03063, Korea

myleeda2@gmail.com, eunilpark@skku.edu, csehong@skku.edu

Abstract

A face is expressive of individual identity and used in various studies such as identification, authentication, and personalization. Similarly, a voice is a means of expressing individuals, and personalized voice synthesis based on voice reference is active. However, the voice-based method confronts voice sample dependency limitations. We propose Face-based Voice synthesis for Text-To-Speech (FVTTS) to synthesize voice from face images that are more expressive of personal identity than voice samples. A major challenge in face-based TTS methods is extracting distinct voice features highly related to voice from the face image. Our face encoder is designed to tackle this by integrating global facial attributes with voice-related features to represent personalized characteristics. FVTTS has shown superiority in various metrics and adaptability across different data domains. We establish a new standard in face-based TTS, leading the way in personalized voice synthesis.

Index Terms: face-based TTS, face voice conversion, face to speech, end-to-end TTS

1. Introduction

A personal face is expressive of individual identity encompassing gender, age, and ethnicity. Face images are widely used in various studies such as identification, authentication, and personalization. Similarly, voice is a means of expressing individuals, and personalized voice synthesis has been developed. In personalized voice synthesis, previous studies achieved remarkable performance with short reference voice samples [1, 2, 3, 4]. YourTTS [3] introduced a multilingual approach for multi-speaker TTS and synthesized a new voice based on the input sample voice. MegaTTS [4] proposed a large-scale TTS model with inductive bias. They achieved state-of-the-art performance by generating new voices and maintaining unique voice attributes such as content, tone, and phase by introducing the prosody large language model. However, these approaches require a reference sample of the target voice to synthesize the new voice. There could be situations where obtaining the target speaker's voice is problematic, or synthesizing voices for fictional characters, such as animation characters and virtual reality avatars, is required. In addition, the reference voice-based TTS relies solely on the features presented in a given voice sample and has difficulty representing diverse characteristics.

We can consider face images as conditions of personalization text-to-speech (TTS), instead of conventional widely used voice samples. As mentioned, the face images are expressive of personal identity and represent personal characteristics not contained in voice samples. Motivated by the correlation between

the face and the voice [5, 6], we generate the personalized voice using the face image. Previous studies pointed out that articulatory structures like vocal cords, facial muscles, and facial bones are responsible for producing personal voices [7, 8]. From the perspective of neuroscience, humans can match the unknown face to the voice and vice versa [9, 10]. This voice-face correlation led to the research into the voice synthesis from the face image [11, 12, 13]. Face2Speech [11] was an early model to utilize face images in TTS systems, training a face encoder to align face embeddings and speech embeddings. Subsequent developments focused on extracting more natural and realistic voice features from the speaker's face images as aligning extracted facial features with pre-existing speech embeddings [12, 13]. However, as extracting personalized voice features from face images poses challenges due to their diverse features, including non-voice-related ones, the voice-related face features were not learned directly from the face image in the previous methods. FaceTTS [14] extracted the personalized voice-related features directly from the speaker's face image to synthesize voices. However, they required an additional network to minimize the difference between target and synthetic voices.

In this paper, we propose Face-based Voice synthesis for TTS (FVTTS) model, which synthesizes the new voice from the face image with end-to-end structure. Our proposed model extracts the characteristics of the personal voice directly from the face image, not utilizing any voice samples. FVTTS marks a significant advancement by synthesizing more natural voices without additional network learning. Given the text, speech, and face image, our model extracts each feature by learning the encoders simultaneously. We propose the face encoder structure to extract two face features. One is the global face image features and another is the vocal-related features. With the proposed face encoder, FVTTS extracts distinct personalized voice characteristics directly from the face image. From the multi-modal features, our model learns the proper distribution of speech given the text and face for personalized voice synthesis.

To sum up, our contributions are as follows:

- We propose FVTTS, the face-based end-to-end TTS model. FVTTS generates more natural personalized voices utilizing the unique features from the face without any voice samples.
- We propose integrating a face encoder into the TTS structure, designed to extract speech directly from facial images by combining global image attributes and specific facial features that represent personalized voice characteristics.
- With the experiments, we demonstrate the potential of face-based zero-shot TTS and establish our model as a new benchmark for face-based TTS systems. Supplementary demo voice samples and codes are presented at the project page¹.

*Corresponding author

¹<https://dxlabsskku.github.io/FVTTS/>

2. Method

FVTTS generates a personalized voice based on the speaker’s face image. To solve the difficulty of extracting voice features directly from the face image, we introduce the face encoder that fuses two different image features. With the fused face features, FVTTS synthesizes the personalized voice based on VITS [15] structure. Figure 1 shows the overall architecture of FVTTS and our proposed face encoder structure.

2.1. Encoder

Text Encoder. The text encoder extracts text features from the text phonemes. We use a transformer encoder that is used in VITS. The outputs of the text encoder are used to compute the distribution of the latent variable through the projection layer.

Face Encoder. The face encoder is proposed to get speaker features, h_{sid} , that express the personalized voice feature from the face image. Since we use the face image, we do not need to predefine the number of speaker embeddings to be learned. As shown in Figure 1b, the face encoder is composed of a global face feature extraction (ENC_g) and a personalized voice-related feature extraction (ENC_p). The same face image is input in both ENC_g and ENC_p .

The ENC_g is the network that learns the global features of the image. We use the two stacked convolution blocks and a linear layer to extract image embedding emb_{img} . For personalized features, we utilize the ENC_p . We use the pre-trained FaceNet model [16] that is generally adopted for face identification [17, 18] to extract individual characteristics. The facial embedding is calculated through the frozen FaceNet and passes the linear layer to get the final face embedding emb_{face} . The extracted features are fused to make the speaker feature, h_{sid} . emb_{img} and emb_{face} are conditioned in our model as the following equation. We set the weights of each feature, w_{img} and w_{face} , as learnable parameters updated during training. h_{sid} is the combination of two learnable embeddings as follows and we show the effectiveness through the ablation studies.

$$h_{sid} = w_{img}emb_{img} + w_{face}emb_{face} \quad (1)$$

Posterior Encoder The posterior encoder calculates the speech latent representation z with the spectrogram as input. This module consists of the non-causal WaveNet residual blocks [19]. To generate various voices conditioned on individual speaker’s characteristics, the speaker feature from the face image, h_{sid} , adds in every residual block using global conditioning.

2.2. Flow Model

Normalizing flow receives the spectrogram features from the Posterior Encoder and learns the distribution of the latent representation to synthesize voice without the intermediate representation. This allows for generating human-like rhythms of speech with end-to-end structure [15, 3]. As shown in Fig. 1a, during training process, normalizing flow is trained with the redefined Monotonic Alignment Search (MAS) [15], and the stochastic duration predictor is used for diverse rhythms from input text. On the other hand, in the inference process, we do not use MAS. We use the inverse flow network based on the text encoder and predict the duration through the stochastic duration predictor. The inference process is shown in Fig. 1c.

2.3. Vocoder

Given the speech’s latent representation, the Vocoder synthesizes the personalized voice. We use the generator of HiFi-GAN v1 [20] as the vocoder and modified discriminator referred [15]. To generate personalized voice, face embedding h_{sid} is added to the input z from the posterior encoder.

3. Experiment

3.1. Dataset

We utilize Lip Reading Sentence 3 (LRS3) [21], the largest English audio-visual dataset, as our primary training set. Frontal face images are obtained using the shape predictor 68 face landmarks model, with five images sampled per video. Our training set comprises 10,282 utterances from 3,995 distinct speakers.

We conduct experiments using two additional datasets to assess the versatility of face-based Text-to-Speech (TTS) across different datasets. First, we employ the GRID dataset [22], the controlled laboratory dataset. We extract three face images per video for performance evaluation. Second, we explore the application of face-based TTS on animated characters by collecting images from various publicly available animation movies. This dataset includes eleven characters, six female and five male, each designed to closely resemble human features.

3.2. Implementation Details

During training, we randomly sample the speaker’s face image to pair the different images of the same speaker with the same utterances for every step to prevent overfitting on specific image and utterance pairs. The model is trained using an NVIDIA GeForce RTX 3080 10GB with a batch size of 20. We use the AdamW optimizer with a learning rate of 0.0002 and set the sampling rate of voice as 44.1kHz. We train FVTTS 1.5M steps.

To evaluate the performance, we compare the results of FVTTS with the publicly available models. We use YourTTS [3] constructed on VITS [15] as same as ours for voice-based TTS model. YourTTS is specialized to synthesize the voice given the target speaker’s voice samples. Another one is FaceTTS [14], a face-styled diffusion TTS model.

3.3. Evaluation Metric

We use speaker encoder cosine similarity (SECS), word error rate (WER), mel cepstral distortion (MCD), gender accuracy (GA), and mean opinion score (MOS) as evaluation metrics. SECS [23, 3] is calculated as the cosine similarity between the speech of the same speaker with different face images. Through WER, the intelligibility of synthesized speech consistent with the input text can be measured. MCD measures the difference between the synthesis speech and the target speech. GA is used to confirm speaker diversity and distinguishability because gender is the primary feature of personal characteristics.

For subjective evaluation, the raters evaluate two to four synthesis voices per speaker for consistency (MOS-C), intelligibility (MOS-T), naturalness (MOS-N), the similarity of synthesis voice and source voice (MOS-S), and the synchronization between face and voice (MOS-M). The raters consist of nine males and six females. Each rater listens to the voice samples at least twice independently and rates in 1 to 5 points.

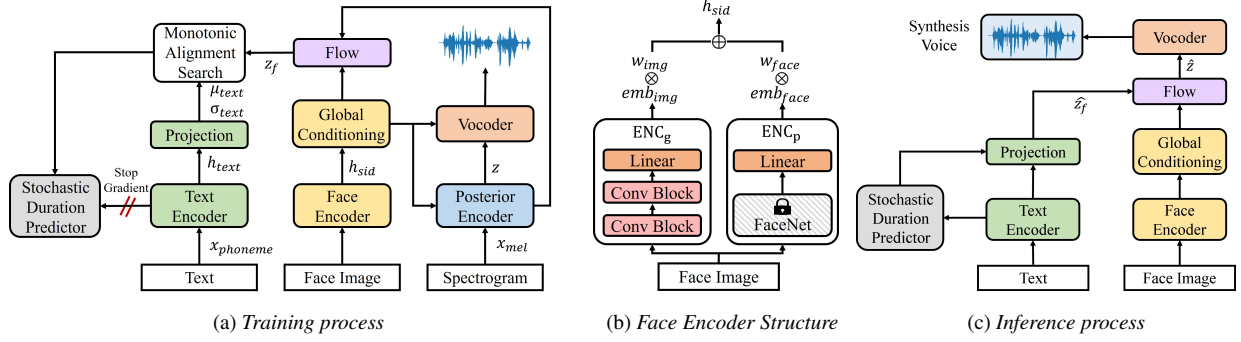


Figure 1: Overview of FVTTS. FVTTS receives text, face image, and spectrogram of speech as input, and sends them into the three encoders (2.1). The flow network (2.2) represents the distribution of speech and the vocoder (2.3) synthesizes the personalized voice

		Consistency		Intelligibility		Naturalness	Similarity		Matching	Diversity
		SECS \uparrow	MOS-C \uparrow	WER \downarrow	MOS-T \uparrow	MOS-N \uparrow	MCD \downarrow	MOS-S \uparrow	MOS-M \uparrow	GA \uparrow
Voice-based	YourTTS [3]	0.779	3.778 \pm 0.87	0.207	4.137 \pm 0.77	3.692 \pm 1.07	19.699	2.769 \pm 1.13	3.658 \pm 1.09	0.606
Face-based	FaceTTS [14]	0.748	3.231 \pm 1.22	0.265	2.846 \pm 1.03	1.752 \pm 0.77	19.832	1.462 \pm 0.69	2.051 \pm 0.92	0.576
	FVTTS (w/o ENC_p)	0.708	2.923 \pm 1.11	0.524	3.419 \pm 1.16	3.103 \pm 1.17	17.173	2.308 \pm 1.02	3.248 \pm 1.14	0.756
	FVTTS (w/o ENC_g)	0.691	2.803 \pm 0.99	0.544	3.085 \pm 1.06	2.701 \pm 1.04	16.561	2.274 \pm 0.98	3.385 \pm 1.27	0.485
	FVTTS (full)	0.754	3.897 \pm 0.95	0.306	4.231 \pm 0.76	3.906 \pm 0.97	16.792	3.000 \pm 1.10	3.838 \pm 1.10	0.788

Table 1: Result of speech synthesis on LRS3 dataset.

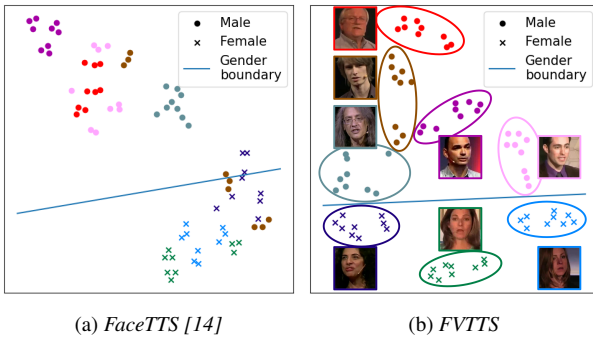


Figure 2: Visualization of synthesis speech on LRS3 Dataset. The ‘x’ and ‘o’ marks represent the female and the male speakers. The blue line divides speaker embeddings as gender.

	MOS-C \uparrow	MOS-T \uparrow	MOS-N \uparrow	MOS-M \uparrow
FaceTTS [14]	2.983 \pm 1.36	3.150 \pm 0.95	2.583 \pm 1.00	1.625 \pm 0.99
FVTTS	3.950 \pm 0.97	3.983 \pm 0.87	3.817 \pm 0.88	3.732 \pm 1.06

Table 2: Evaluation on GRID dataset.

	MOS-T \uparrow	MOS-N \uparrow	MOS-M \uparrow	GA \uparrow	Prefer (%)
FaceTTS [14]	3.475 \pm 1.24	2.524 \pm 1.00	1.848 \pm 0.84	0.60	1.07
FVTTS	4.230 \pm 0.86	4.137 \pm 0.87	3.995 \pm 0.88	0.80	98.93

Table 3: Evaluation on animation images.

3.4. Results on LRS3 Dataset

Table 1 shows the results on LRS3 dataset. It can be seen that FVTTS improves on almost metrics for face-based TTS. Our model achieves SECS and MOS-C scores similar to voice-based TTS implying that FVTTS extracts similar personalized features from different images of the same speaker. In the intelligibility, FVTTS achieves the best score at MOS-T although the WER score is lower than FaceTTS. These results are caused by the difference between machine and human intelligibility, but our goal is voice synthesis for humans and our model seems

more intelligible for humans. With the MOS-N score, our model has the best performance showing the perceived naturalness are better. With MOS-S result and MCD score, our model has demonstrated the ability of voice synthesis similar to the target speaker’s voice based on the face image. In addition, we confirm the synchronization of the face image and the synthesis voice with the best MOS-M results.

With the best GA score surpassing the voice-based TTS model, we confirm that FVTTS is good at catching the gender feature from the face image. In addition, we visualize the speaker embeddings that converted from the synthetic speech using *Resemblyzer*. Figure 2 shows that our model not only generates various voices but also captures the personalized feature and synthesizes speaker-distinguishable voices.

As for the ablation study, we show the effectiveness of our model. The full model improves the SECS score to 0.879 from 0.825 of the baseline without ENC_p . In human evaluation, the proposed structure shows more natural and better face-matching results than a single network.

3.5. Results on Cross-Data

3.5.1. Results on GRID dataset

Table 2 shows the results on GRID dataset. Especially on the face matching, FVTTS surpasses the baseline. These results represent that FVTTS generates a voice synchronized with the target speaker’s image even out of the distribution of training data, and it would be helpful to generate a new voice for a new speaker. These results mean that synthetic speech using FVTTS is easy to understand and natural regardless of the data domain and imply that our model can also be applied to other domains.

3.5.2. Results on Animation Images

We integrate speech synthesis with animation character images to demonstrate the adaptability of our model across various contexts. Table 3 presents the results on animation characters. Across measures of intelligibility, naturalness, and face matching, human evaluators consistently rate the voices generated by

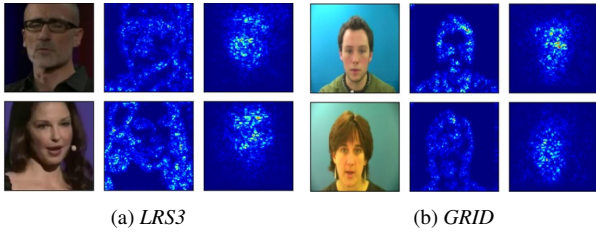


Figure 3: Visualization of face encoder. The left images show the source face image, while the middle and right images present the focus regions of ENC_g and ENC_p .

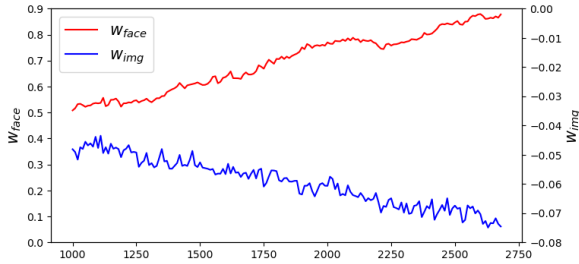


Figure 4: The variation of learnable weight in face encoder. The red and blue lines show the variation of w_{face} and w_{img} .

	Fixed w_{face}		Fixed w_{img}	
	SECS \uparrow	GA \uparrow	SECS \uparrow	GA \uparrow
0.1	0.735	0.765	0.734	0.735
0.2	0.737	0.571	0.735	0.600
0.3	0.732	0.686	0.730	0.611
0.4	0.732	0.636	0.735	0.514
0.5	0.733	0.639	0.732	0.588
0.6	0.733	0.667	0.733	0.722
0.7	0.727	0.657	0.733	0.629
0.8	0.729	0.559	0.734	0.545
0.9	0.731	0.667	0.733	0.676
1.0	0.735	0.686	0.733	0.778
FVTTS	0.754	0.788	0.754	0.788

Table 4: Results with various w_{img} and w_{face} .

our model higher than those produced by FaceTTS. Particularly notable are the substantial differences in MOS-Naturalness (MOS-N) and MOS-face Matching (MOS-M), underscoring the superior performance of our model, even within the realm of animation. Furthermore, our model demonstrates a significantly higher gender attribution score (GA 0.8) compared to FaceTTS (GA 0.6), indicating a more accurate portrayal of gender in synthetic speech. Preference ratings for synthetic voices reveal a strong inclination towards our model, with a preference rate of 98.93% over FaceTTS.

3.6. Efficiency of Face Encoder

3.6.1. Visualization of Face Encoder

We visualize each encoder’s focal area in representing personalized features to show the efficacy of the face encoder. Figure 3a and Figure 3b show the visualization results for the LRS3 and GRID datasets. ENC_g , which is designed to extract overall facial features, concentrates on delineating the face outline and identifying facial components within the images. In contrast, ENC_p is tasked with capturing more personalized features. Vi-



Figure 5: Spectrogram visualization of synthetic voices.

ualization results indicate that ENC_p predominantly emphasizes the central facial regions, particularly the nose and lips. These findings substantiate the hypothesis that ENC_g acquires global image features, while ENC_p plays a crucial role in extracting personalized voice characteristics.

3.6.2. The effectiveness of the learnable weights

Figure 4 shows the variations of learnable weights after 500,000 iterations. The orange line presents the value of w_{face} , increasing with each iteration, while the blue line signifies the value of w_{img} , steadily decreasing. This trend suggests the significance of ENC_p derived facial features in representing personalized voice characteristics. The red dashed line shows the weights for the selected model, with w_{img} at -0.0735, and w_{face} at 0.8746.

We conducted experiments to assess the validity of these weights by examining their impact. We held one weight fixed at the learned value while varying the other from 0.1 to 1.0. The evaluation outcomes are detailed in Table 4. Comparing the performance achieved by our proposed model, none of the alternative approaches surpassed our performance that utilized the learned weights. The results verify that the learnable weights of the Face Encoder during the training process significantly enhance the overall voice synthesis performance of FVTTS.

3.7. Qualitative Result

We visualize the synthetic voices spectrogram and the face images of several speakers reading the same text; ‘The question is who will have it.’ In Figure 5, the first row shows the synthetic voices of male speakers and the second row presents the synthesis results of female speakers. The left male speaker on the first row exhibits a more varied frequency range compared to the left female speaker on the second row. Further, upon comparing the voices of two male speakers, noticeable differences in attributes such as pitch and phase are discernible. The synthetic voice of each speaker demonstrates a similar pattern while retaining distinct characteristics unique to them.

4. Conclusion

We introduce FVTTS, an innovative face-based voice synthesis model that leverages facial images to generate personalized voices. Extensive evaluations have demonstrated FVTTS’s effectiveness in delivering voices with consistency, intelligibility, and diversity, achieving notable performance. Our ablation study further confirms the efficiency of our unique face encoder. While FVTTS exhibits promising results, there exist opportunities for further refinement. Our model aims to enhance its ability to accurately generate voices across a broader spectrum of gender presentations, particularly for individuals with unique hairstyles. Additionally, efforts are underway to capture a wider range of emotions reflected in facial expressions. Future enhancements will involve integrating gender-aware and emotion-aware modules, thereby enhancing the model’s capacity to produce even more nuanced and representative voices.

5. Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2021-0-00358, AI-Big data based Cyber Security Orchestration and Automated Response Technology Development). This research was also supported by the MSIT, Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) program(IITP-2024-2020-0-01816) supervised by the IITP. This research was supported by the Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(RS-2023-00271314).

6. References

- [1] Y. Wu, X. Tan, B. Li, L. He, S. Zhao, R. Song, T. Qin, and T.-Y. Liu, "AdaSpeech 4: Adaptive Text to Speech in Zero-Shot Scenarios," in *Proc. of Interspeech '22*, 2022, pp. 2568–2572.
- [2] B. Zhao, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "nnspeech: Speaker-guided conditional variational autoencoder for zero-shot multi-speaker text-to-speech," in *Proc. of ICASSP '22*. IEEE, 2022, pp. 4293–4297.
- [3] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *Proc. of ICML '22*. PMLR, 2022, pp. 2709–2720.
- [4] Z. Jiang, Y. Ren, Z. Ye, J. Liu, C. Zhang, Q. Yang, S. Ji, R. Huang, C. Wang, X. Yin *et al.*, "Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias," <https://arxiv.org/abs/2306.03509>, 2023.
- [5] Y. Wen, B. Raj, and R. Singh, "Face reconstruction from voice using generative adversarial networks," *Advances in neural information processing systems*, vol. 32, pp. 1–10, 2019.
- [6] H.-H. Lu, S.-E. Weng, Y.-F. Yen, H.-H. Shuai, and W.-H. Cheng, "Face-based voice conversion: Learning the voice behind a face," in *Proc. of ACM MM '21*, 2021, pp. 496–505.
- [7] H. Ning, X. Zheng, X. Lu, and Y. Yuan, "Disentangled representation learning for cross-modal biometric matching," *IEEE Transactions on Multimedia*, vol. 24, pp. 1763–1774, 2021.
- [8] C.-Y. Wu, C.-C. Hsu, and U. Neumann, "Cross-modal perceptionist: Can face geometry be gleaned from voices?" in *Proc. of CVPR '22*, 2022, pp. 10452–10461.
- [9] L. W. Mavica and E. 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.
- [10] H. M. Smith, A. K. Dunn, T. Baguley, and P. C. Stacey, "Matching novel face and voice identity using static and dynamic facial images," *Attention, Perception, & Psychophysics*, vol. 78, pp. 868–879, 2016.
- [11] S. Goto, K. Onishi, Y. Saito, K. Tachibana, and K. Mori, "Face2speech: Towards multi-speaker text-to-speech synthesis using an embedding vector predicted from a face image," in *Proc. of Interspeech '20*, 2020, pp. 1321–1325.
- [12] B. Plüster, C. Weber, L. Qu, and S. Wermter, "Hearing faces: Target speaker text-to-speech synthesis from a face," in *Proc. of ASRU '21*. IEEE, 2021, pp. 757–764.
- [13] J. Wang, Z. Wang, X. Hu, X. Li, Q. Fang, and L. Liu, "Residual-guided personalized speech synthesis based on face image," in *Proc. of ICASSP '22*. IEEE, 2022, pp. 4743–4747.
- [14] J. Lee, J. S. Chung, and S.-W. Chung, "Imaginary voice: Face-styled diffusion model for text-to-speech," in *Proc. of ICASSP '23*. IEEE, 2023, pp. 1–5.
- [15] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. of ICML '21*. PMLR, 2021, pp. 5530–5540.
- [16] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. of CVPR '15*, 2015, pp. 815–823.
- [17] S. Kangwanwatana and T. Sucontphunt, "Improve face verification rate using image pre-processing and facenet," in *Proc. of ICBIR '22*. IEEE, 2022, pp. 426–429.
- [18] B. Alharbi and H. S. Alshanbari, "Face-voice based multimodal biometric authentication system via facenet and gmm," *PeerJ Computer Science*, vol. 9, p. e1468, 2023.
- [19] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. of ICASSP '19*. IEEE, 2019, pp. 3617–3621.
- [20] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17022–17033, 2020.
- [21] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," <https://arxiv.org/abs/1809.00496>, 2018.
- [22] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.
- [23] E. Casanova, C. Shulby, E. Gölge, N. M. Müller, F. S. de Oliveira, A. Candido Jr., A. da Silva Soares, S. M. Aluisio, and M. A. Ponti, "SC-GlowTTS: An Efficient Zero-Shot Multi-Speaker Text-To-Speech Model," in *Proc of Interspeech '21*, 2021, pp. 3645–3649.