# IMAGINARY VOICE: FACE-STYLED DIFFUSION MODEL FOR TEXT-TO-SPEECH

*Jiyoung Lee*[1], *Joon Son Chung*[2], *Soo-Whan Chung*[3]

[1]NAVER AI Lab, South Korea
[2]Korea Advanced Institute of Science and Technology, South Korea
[3]NAVER Cloud, South Korea

## ABSTRACT

The goal of this work is zero-shot text-to-speech synthesis, with speaking styles and voices learnt from facial characteristics. Inspired by the natural fact that people can imagine the voice of someone when they look at his or her face, we introduce a face-styled diffusion text-to-speech (TTS) model within a unified framework learnt from visible attributes, called FACE-TTS. This is the first time that face images are used as a condition to train a TTS model.

We jointly train cross-model biometrics and TTS models to preserve speaker identity between face images and generated speech segments. We also propose a speaker feature binding loss to enforce the similarity of the generated and the ground truth speech segments in speaker embedding space. Since the biometric information is extracted directly from the face image, our method does not require extra fine-tuning steps to generate speech from unseen and unheard speakers. We train and evaluate the model on the LRS3 dataset, an in-the-wild audio-visual corpus containing background noise and diverse speaking styles. The project page is https://facetts.github.io.

***Index Terms***— Multi-speaker text-to-speech (TTS), Audio-visual biometrics, Diffusion model

## 1. INTRODUCTION

Text-to-speech (TTS) is one of the core tasks in speech processing that generates speech waveform from a given text transcription. Deep generative models have been introduced to produce high-quality spectral features from text sequences [1, 2, 3]. They have brought remarkable improvements in the quality of synthetic speech signals, compared to traditional parametric synthesis methods.

Recent works on diffusion models [4, 5, 6] have provided excellent generation results with outputs of high quality in various research fields such as image generation, video generation, and natural language processing. For example, diffusion methods have achieved noteworthy results in image generation models; *e.g.* DALLE-2 [7], Stable Diffusion [8]. Likewise, diffusion methods have shown impressive results in TTS compared to the previous generative methods, both in acoustic modeling [9, 10, 11] and in the vocoder [12, 13].

However, there are several unresolved challenges in the field of TTS. One problem we address in this paper is expanding single speaker TTS model to multi-speaker TTS. Since every person has different speaking styles, tones or accents, it is very challenging for the TTS model to learn various speaker styles. The second and related problem is that a significant amount of target speakers' speech samples are required to generate voices of unseen speakers, even for multi-speaker TTS. The variability of speaking styles means that the model must have access to significant amount of enrollment data to learn about each speaker. Since it is difficult to obtain clean enrollment utterances for each speaker, this raises the question "what if face images can be used for enrollment instead of clean speech?"

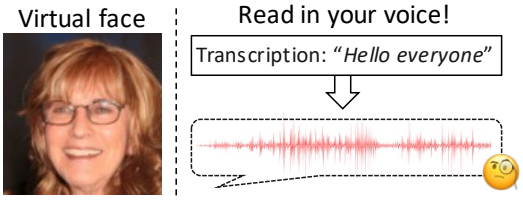



**Fig. 1**: FACE-TTS generates speech from a given text, conditioned on a face image. The face image is sampled from [8].

In [14, 15], the authors propose to leverage face images to control speaker characteristics of synthesised speech. They train the face identity encoder to share a joint embedding space with the voice encoder, independently from the TTS model. This approach enables generation of speech for unseen speakers without extra speaker adaptation. However, these works do not use the face images as inputs when training the TTS models. Instead, the models are trained using speaker embeddings as the input, and the embeddings are swapped to face images only during inference.

In this paper, we propose a novel speech synthesis model, FACE-TTS, which leverages face images to provide a robust characteristic of speakers. In [16, 17], the authors have explored cross-modal biometrics and demonstrated that there is a strong correlation between voices and face appearances. Inspired by this, we design a multi-speaker TTS model, where speaking styles are conditioned on face attributes. While it is difficult to collect speech segments for the enrollment of every speaker, it is much easier to obtain face images. We enforce the matching of the identity of the face and the identity of the synthesised speech to train a robust cross-modal representation of speaking style. Our approach is capable of generating speech signals without speaker enrollment, which is advantageous in the zero-shot or few-shot TTS modeling. Our backbone structure for the TTS model is derived from Grad-TTS [11], which learns acoustic features using the diffusion method. Unlike other face-to-speech synthesis methods [14, 15], FACE-TTS is trained end-to-end from the face encoder to the acoustic model, using in-the-wild datasets. To the best of our knowledge, this is the first time that face images are used as a condition to train a TTS model. We perform qualitative and quantitative tests to assess the speaker representations as well as the perceptual quality of the synthesised speech. In addition, we verify through subject measures whether the synthesised speech fits well with the appearances of virtual humans who do not have their own voices as illustrated in Fig. 1.

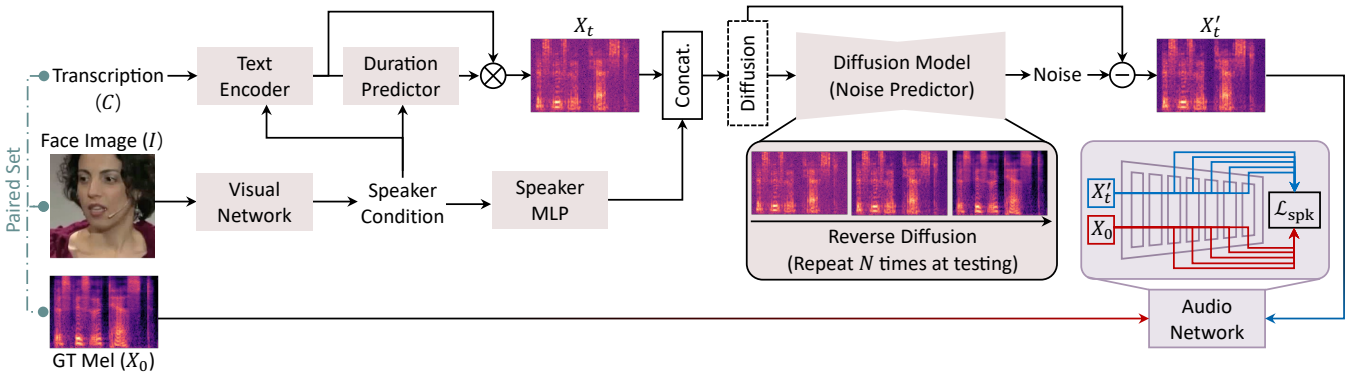



**Fig. 2**: The overall configuration of FACE-TTS. Given a text transcription and a face image, our method generates a speech sample using a diffusion model conditioned on face images to model speaker characteristics. The whole network except for audio network is trained end-to-end using the LRS3 dataset. Notice that the audio network are used only during training.

## 2. RELATED WORK

**Text-to-speech.** With the success of deep neural networks, the perceptual quality of synthesised speech is dramatically improved compared to the previous statistical parametric speech synthesis [18]. In general, TTS models are composed of two modules; an acoustic model and a vocoder. The acoustic model generates speech features (commonly mel-spectrogram) from text sequences, and the vocoder takes the features to generate speech waveform. There have been many approaches using generative modeling methods [1, 2, 19], and Tacotron-based models [3, 20] incorporate a sequence-to-sequence model to transform the text sequences into the acoustic representations. GAN-based models [2, 19] have brought innovative contributions to TTS in the last decade using adversarial training strategy. Recently, another successful generative approach, diffusion-based methods [11, 12, 13, 9], have been proposed in speech synthesis, as the diffusion methods have proved their effectiveness in various generation tasks [21, 8, 7]. Compared to GAN-based models, diffusion methods have advantages in impressive results as well as distribution coverage, a fixed training objective, and scalability.

**Audio-visual biometrics.** People instinctively co-relate others' facial appearances and their voices by learning through experiences, because face and voice provide related identity information [22]. In order to learn this correlation between faces and voices, several prior works [16, 17] have tried to use self-supervised methods in the way that people learn from experience. They have leveraged the fact that a face image and speech segment from a single-speaker video should have a common identity. In [23, 24], the authors have shown that visual identity has a strong correlation with the speaker identity by separating input signals using face images. Various self-supervised losses have been considered to learn robust cross-modal embeddings for biometrics matching, such as cross-entropy loss [25], contrastive loss [16] and disentanglement-based loss [26]. Motivated by these previous works, we leverage the cross-modal biometrics matching to foster conditions that reflect speaker-dependent characteristics for multi-speaker TTS model.

## 3. FACE-TTS

### 3.1. Score-based Diffusion Model

FACE-TTS is based on a score-based diffusion model, specifically Grad-TTS [11], which consists of three main parts; (1) text encoder, (2) duration predictor, (3) diffusion model. Formally, given a text transcription $C$ and a corresponding mel-spectrogram $X_0$ for training, the forward process progressively adds standard Gaussian noise to satisfy the following continuous stochastic differential equation (SDE) [27]:

$$dX_t = -\frac{1}{2}X_t\beta_t dt + \sqrt{\beta_t}dW_t, \tag{1}$$

where $W_t$ is the standard Brownian motion, and $\beta_t$ is a noise schedule. In the reverse diffusion process, $X_0$ can be obtained from $X_t$ corresponding to the text as follows:

$$dX_t = -\Big(\frac{1}{2}X_t + \mathcal{S}(X_t, t)\Big)\beta_t dt + \sqrt{\beta_t}d\tilde{W}_t, \tag{2}$$

where $\tilde{W}_t$ is the reverse-time Brownian motion and $\mathcal{S}(X_t, t)$ is a diffusion model that estimates the gradient of the log-density of noisy data $\nabla_{X_t} \log p_t(X_t)$. Namely, we infer the speech $X_0$ from noisy data $X_t$ with $N$ steps by solving the SDE:

$$X_{t-\frac{1}{N}} = X_t + \frac{\beta_t}{N}\Big(\frac{1}{2}X_t + \mathcal{S}(X_t, t)\Big) + \sqrt{\beta_t}\tilde{W}_t, \tag{3}$$

where $t \in \{\frac{1}{N}, \frac{2}{N}, \dots, 1\}$. We note that $N$ is the number of steps to the discretised reverse process, and $t$ indexes a subsequence of time steps in the reverse process. We follow most parts similar to the original methodology [11] and explain the different points in the below sections. The overall architecture is illustrated in Fig. 2.

### 3.2. Speaker Conditioning with Cross-modal Biometrics

In [11, 28], the authors do not utilise a speaker model for learning speaking styles in their TTS models, but prepare a pre-defined speaker codebook for each identity. Thus, it is difficult to present a new speaker in their models, and it requires a challenging adaptation procedure to resolve this problem. In [29, 30], they prove that the speaker embedding precisely adjusts speaking styles in synthesised speech. However, there still remains a problem. Speaker embeddings usually represent excessive details of speakers, and it yields unstable training in the acoustic modeling of TTS. Therefore, speaker embeddings should be generalised to represent speakers' voices in synthesised speech.

In this paper, we provide identity embedding from a face image as a conditioning feature on the TTS model for multi-speaker modelling. Since the face embedding from the cross-modal biometric model represents the identity related to the voice, it is suitable to generate speech that matches face attributes. Such face embedding does not contain a complex distribution of speakers, but only associative representations from voice and face, and it naturally generalises the speaker embedding and allows efficient multi-speaker

modelling. Given a mel-spectrogram $X$=$X_0$ and a face image $I$, the network is pre-trained to associate the same speaker identity from the different modalities, where the overall network consists of audio network $\mathcal{F}(X)$ and visual network $\mathcal{G}(I)$.

The visual network ingests a face image of the target speaker to produce a speaker representation. Then the text encoder and the duration predictor estimate the statistics of acoustic features from given a text transcription and a face image. In details, the text encoder generates acoustic features fit to text sequences, and the duration predictor colourises features with predicted speaking duration of the target speaker for the natural pronunciation. During training, the diffusion process adds Gaussian noise on colourised features to make noisy data, and the diffusion model estimates the gradient of data distribution in noisy data to obtain the target audio. Specifically, the speaker representation guides the diffusion model to estimate gradients optimal to generate synthesised speech in the speaker's voice. We note that the network configuration follows [11].

However, to learn various speakers' characteristics for the multi-speaker TTS, the TTS model requires sufficient length of recorded speech for each person. Previous works [9, 14, 15] trained their models using audiobook dataset read by several speakers with enough lengths of utterances, where it is difficult to generalise models for unseen speakers. To solve this problem, we suggest an effective strategy, a speaker feature binding loss, maintaining speaker characteristics of target voices in synthesised speech. It allows FACE-TTS to learn face-voice association from audio segments even with a short length. Formally, latent embeddings from convolution layers of the audio network trained in cross-modal biometrics are extracted from synthesised speech and target voices, respectively. The speaker feature binding loss $\mathcal{L}_{\text{spk}}$ train our FACE-TTS model by minimising distances of two latent embedding sets as follows:

$$\mathcal{L}_{\text{spk}} = \sum\nolimits_{B} |\mathcal{F}_b(X_0) - \mathcal{F}_b(X'_t)|, \tag{4}$$

where $X_0$ is a mel-spectrogram from a target speaker's utterance and $X'_t$ is a denoised output from the network, and $B$ indicates the number of convolution blocks of audio network except for the first two convolution blocks. We freeze the audio network not to be updated with this loss. This training strategy enforces to form a speaker-related latent distribution of synthesised speech similar to that of the target speech.

### 3.3. Training & Inference

In training session, FACE-TTS learns multi-speaker speech synthesis through multiple training criteria. To train text and duration encoders, we exploit the prior loss to estimate the mean from a normal distribution and the duration loss [28] to control the duration of pronunciation using a monotonic alignment between speech and text sequences. Diffusion loss trains the diffusion model to estimate the gradient of data distribution as in [11]. Our final training objective is described as:

$$\mathcal{L} = \mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{duration}} + \mathcal{L}_{\text{diff}} + \gamma\mathcal{L}_{\text{spk}}, \tag{5}$$

where $\gamma$ is empirically set to $1e$-2. We emphasise that the whole framework is trained end-to-end on LRS3 dataset obtained from in-the-wild environments. Thanks to video in LRS3 with various angles and facial expressions, our FACE-TTS is more robust to real-world face images than previous works [14, 15] that only used the front view of a face image.

For inference, the trained FACE-TTS samples a mel-spectrogram of utterance $X_0$ from the noisy data $X_t$ that is estimated by transcription with speaker condition by target speaker's face. The reverse diffusion process is repeatedly processed to estimate step-by-step noise gradually. Finally, we used a pretrained vocoder to transform the estimated mel-spectrogram to a raw waveform.

## 4. EXPERIMENTS

### 4.1. Experimental Settings

**Datasets.** LRS3 [31] is an audio-visual dataset collated from TED videos, which has audio-visual pairs with corresponding text transcriptions. We use the *trainval* split for the training and the *test* split for the evaluation, excluding speech samples shorter than 1.3 seconds. Also, we pick out speech samples of speakers who have at least 10 seconds of audio in total. A total of 14,114 utterances and 2,007 speakers is used for training, 50 utterances for validation, and test set includes 412 speakers. The widely used multi-speaker TTS dataset, such as LibriTTS [32], has 550 seconds per speaker in average from well-recorded audio books, whereas LRS3 [31] has a length of about 34 seconds extracted from real-world environments. Therefore, it is extremely challenging to use LRS3 data to train TTS models. We use the test split (448 samples) of LJSpeech [33] to obtain text descriptions in the out-of-distribution for a fair comparison with previous works [11, 28]. The cross-modal biometric model [34] is re-implemented following to the same configuration of mel-spectrogram with vocoder [19]. It is trained on VoxCeleb2 [35] dataset which contains 5,994 speakers in audio-visual pairs.

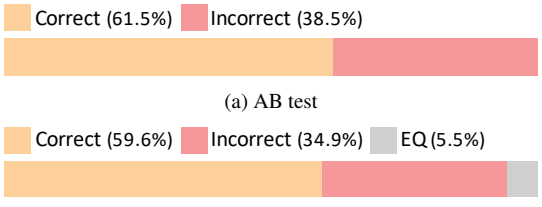
**Audio and image representation.** The inputs to the network, including cross-modal biometric model, TTS model and vocoder, are the 128-dimensional mel-spectrogram extracted at every 10ms with 62.5ms frame length in 16kHz sampling rate. For the image input, the face image is randomly sampled from each video and resized into 224×224 pixels, same as in [17]. The cross-modal biometric model (*i.e.* audio and visual networks) embeds audio and face images onto 512-dimensional vectors.

**Evaluation protocols.** In our experiments, the generated mel-spectrogram is synthesised into an audio waveform using HiFi-GAN as the vocoder. We first report 'Mel.+HiFi-GAN' to inform the degradation amount caused by the vocoder. In this case, mel-spectrogram of target speech is transformed into the waveform without synthesis process. It is natural that it shows a little lower scores with the 'Ground-Truth' result, and it can be the upper-bound score of synthesis results. We perform mean opinion score (MOS) test, which is a common metric to measure subjective perceptual quality of synthesised speech. A total of 17 participants are asked to judge the quality about the synthesis results in 5-scale: 1=Bad; 2=Poor; 3=Fair; 4=Good; 5=Excellent. In the test, 10 utterances are randomly selected from the test set and synthesised using each model. Additionally, we conduct two preference tests, 1) AB forced matching test; synthesised speech and two face images, 2) ABX preference test; two synthesised speech signals and one face image. For the validity of our model for the virtual human speech generation, we perform the MOS test whether the synthesis outputs are harmonised with the face images generated from the recent image generation model [8]. Here, we provide choice options from 1 to 4, where the higher score means the synthesised speech is harmonised with the face image. For the objective evaluation, we establish the 5-way cross-modal forced matching test through the cross-modal biometric model, which has to select the matching identity from synthesised speech and 5 face images. In this matching test, we verify the synthesised speech represents similar identity appeared in a face image.

**Implementation details.** For the fair comparison, we train Grad-TTS with LibriTTS and LRS3 datasets, respectively. Also, our

| Method | Spk. ID | 5-scale MOS |
|---|---|---|
| Ground Truth | - | 4.865±.001 |
| Mel.+HiFi-GAN [19] (Upper bound) | - | 4.653±.035 |
| Grad-TTS [11]† (Seen) | Embed | 3.718±.318 |
| FACE-TTS (Seen) | Audio | 3.547±.331 |
| FACE-TTS (Seen) | Face | 3.706±.154 |
| FACE-TTS (Unseen) | Audio | 3.218±.249 |
| FACE-TTS (Unseen) | Face | 3.282±.219 |

**Table 1**: Subjective evaluation for comparison of audio quality with mean opinion score (MOS) metric. Grad-TTS† is trained on LibriTTS, and FACE-TTS are trained on LRS3.




(a) AB test

(b) ABX test

**Fig. 3**: Results of preference tests. (a) Preference for a face matching two synthesised utterances. (b) Preference for a synthesised utterance matching two face appearances.

FACE-TTS is trained using identity embeddings from audio inputs and face inputs, where both embeddings are obtained from cross-modal biometrics. We follow the most of training configuration of Grad-TTS [11]. Since the visual network are initialised with pre-trained weights for the biometric matching task on VoxCeleb2, we give a smaller value ($1e$-6) as a initial learning rate for those networks. We notice that, except for audio and visual networks, the other networks are trained from scratch. The computation time and flops are increased linearly, while more denoising step increases the audio quality. Thus, we equally use 10 denoising or sampling steps to generate speech signals for the inference.

### 4.2. Results

**Audio quality.** We brought pre-trained parameters of Grad-TTS from the author for the comparison, and it had been trained in the LibriTTS dataset for multi-speaker TTS. In our preliminary experiment, we empirically found that Grad-TTS trained on LRS3 dataset showed competitive perceptual quality. Therefore, we evaluated the Grad-TTS trained on LibriTTS as a comparison following authors' official implementation, and we re-sampled generated audio from Grad-TTS from 22.05kHz to 16kHz. FACE-TTS with audio speaker ID was fully trained with the audio network in cross-modal biometrics instead of the visual network. In Table 1, the result indicates that FACE-TTS using face images shows competitive audio quality to Grad-TTS trained on clean speech dataset under the seen speaker condition. We observed that our FACE-TTS can generate audio of fine quality (*i.e.* above 3 score) even for unseen speakers. Furthermore, there is a little difference in the performance between the models using face or audio as conditions. Compared audio-conditioned models, face conditioning has brought more fine-grained audio quality, because the face represents robust identity compared to the speech influenced by recording environments.

**Speaker verification.** We further evaluate the speaker verification task with generated utterances and face images. First, AB and ABX preference tests are performed on human evaluators. To evaluate under more challenging conditions, we conducted the experiment with gender unified. That is, the face or audio in the two cases to be selected were selected from samples of the same gender. The evaluators selected a correct answer rate of about 60% as reported in Fig. 3. Furthermore, Table 2 shows 5-way cross-modal speaker matching accuracy for objective evaluations on the LRS3 dataset. Following their official implementation, we train Grad-TTS [11] on the LRS3 dataset for this experiment. Although the Grad-TTS trained on the LRS3 shows competitive audio quality with ours, capturing the speakers' characteristics in the sound with the settled speaker embedding seems challenging in the Grad-TTS. Moreover, our speaker loss improves the matching performance 2.6% than FACE-TTS without the loss, training the diffusion model to sample the utterance, which is more proper to the target face. However, it still has room to improve the performance up to the result in the first row (Mel.+HiFi-GAN). We remain it as future work.

| Method | Spk. ID | Acc. (%) |
|---|---|---|
| Mel.+HiFi-GAN [19] (Upper bound) | - | 48.6 |
| Grad-TTS [11] | Embed | 19.4 |
| FACE-TTS (w/o. $\mathcal{L}_{\text{spk}}$) | Face | 35.4 |
| FACE-TTS | Face | 38.0 |

**Table 2**: Speaker identification matching accuracy. Since Grad-TTS uses speaker id embedding, its model is evaluated with *seen* speakers and our model is evaluated with *unseen* speakers. Random accuracy is 20%.

| Test sample | 4-scale MOS |
|---|---|
| LRS3 (Real) | 3.471±.291 |
| Stable Diffusion [8]+FACE-TTS (Fake) | 2.941±.462 |

**Table 3**: Matching preference between virtual face images from Stable Diffusion [8] and generated utterances with MOS.

**Virtual speech generation.** To demonstrate the utility of our FACE-TTS, we synthesised speech with virtual face images generated from [8]. Table 3 reports the subjective evaluation of 4 points Likert-scale measurement: 1=Bad; 2=Neutral; 3=Good; 4=Excellent. We had assessors evaluate virtual faces without knowing they were mixed. As the baseline, we also evaluate the preference of ground-truth face-voice pairs, which are randomly selected on the LRS3 dataset. Surprisingly, people gave 'Good' score on average, in that utterance from our FACE-TTS is well matched with virtual face images.

## 5. CONCLUSION

In this work, we proposed FACE-TTS for multi-speaker text-to-speech synthesis with speaker identity conditioned by a face image. For this goal, we leveraged the cross-modal biometric to specify the speaker characteristics from the face for the diffusion-based TTS model, instead of enrolled speech. To jointly train the two modules for enhancing generation performance, we introduced the speaker feature binding loss to maintain speaker consistency between synthesised speech and reference speech. Both quantitative and qualitative evaluations confirmed the high-quality generation of FACE-TTS, particularly containing good representations of target speakers' voices. Moreover, we demonstrated that FACE-TTS is effective for using realistic-sounding voices of virtual humans, which introduces an interesting application to the emerging field.

**Acknowledgments.** The NAVER Smart Machine Learning (NSML) platform [36] has been used in the experiments.

## 6. REFERENCES


[1] A. v. d. 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 preprint arXiv:1609.03499*, 2016.

[2] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP*, 2020.

[3] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, et al., "Tacotron: Towards end-to-end speech synthesis," in *INTERSPEECH*, 2017.

[4] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *ICML*, 2015.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.

[6] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *NeurIPS*, 2021.

[7] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[9] H. Kim, S. Kim, and S. Yoon, "Guided-tts: A diffusion model for text-to-speech via classifier guidance," in *ICML*, 2022.

[10] S. Kim, H. Kim, and S. Yoon, "Guided-tts 2: A diffusion model for high-quality adaptive text-to-speech with untranscribed data," *arXiv preprint arXiv:2205.15370*, 2022.

[11] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *ICML*, 2021.

[12] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," in *ICLR*, 2020.

[13] Y. Koizumi, H. Zen, K. Yatabe, N. Chen, and M. Bacchiani, "Specgrad: Diffusion probabilistic model based neural vocoder with adaptive noise spectral shaping," in *INTERSPEECH*, 2022.

[14] 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 *INTERSPEECH*, 2020.

[15] J. Wang, Z. Wang, X. Hu, X. Li, Q. Fang, and L. Liu, "Residual-guided personalized speech synthesis based on face image," in *ICASSP*, 2022.

[16] A. Nagrani, S. Albanie, and A. Zisserman, "Learnable pins: Cross-modal embeddings for person identity," in *ECCV*, 2018.

[17] S.-W. Chung, H. G. Kang, and J. S. Chung, "Seeing voices and hearing voices: learning discriminative embeddings using cross-modal self-supervision," in *INTERSPEECH*, 2020.

[18] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, 2009.

[19] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, 2020.

[20] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP*, 2018.

[21] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *ICML*, 2022.

[22] 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.

[23] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang, "Facefilter: Audio-visual speech separation using still images," in *INTERSPEECH*, 2020.

[24] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," in *CVPR*, 2021.

[25] S.-W. Chung, J. S. Chung, and H.-G. Kang, "Perfect match: Self-supervised embeddings for cross-modal retrieval," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 568–576, 2020.

[26] A. Nagrani, J. S. Chung, S. Albanie, and A. Zisserman, "Disentangled speech embeddings using cross-modal self-supervision," in *ICASSP*, 2020.

[27] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *ICLR*, 2021.

[28] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," in *NeurIPS*, 2020.

[29] E. Nachmani, A. Polyak, Y. Taigman, and L. Wolf, "Fitting new speakers based on a short untranscribed sample," in *ICML*, 2018.

[30] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, S. Zhao, and T.-Y. Liu, "Adaspeech: Adaptive text to speech for custom voice," in *ICLR*, 2021.

[31] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8717–8727, 2018.

[32] 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 *INTERSPEECH*, 2019.

[33] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[34] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *CVPR*, 2018.

[35] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[36] H. Kim, M. Kim, D. Seo, J. Kim, H. Park, S. Park, H. Jo, K. Kim, Y. Yang, Y. Kim, et al., "Nsml: Meet the mlaas platform with a real-world case study," *arXiv preprint arXiv:1810.09957*, 2018.