



MakeSinger: A Semi-Supervised Training Method for Data-Efficient Singing Voice Synthesis via Classifier-free Diffusion Guidance

Semin Kim¹, Myeonghun Jeong¹, Hyeonseung Lee¹, Minchan Kim¹, Byoung Jin Choi¹, Nam Soo Kim¹

¹Department of Electrical and Computer Engineering and INMC,
Seoul National University, Seoul, South Korea

{smkim21, mhjeong, hslee, bjchoi, mckim}@hi.snu.ac.kr, nkim@snu.ac.kr

Abstract

In this paper, we propose MakeSinger, a semi-supervised training method for singing voice synthesis (SVS) via classifier-free diffusion guidance. The challenge in SVS lies in the costly process of gathering aligned sets of text, pitch, and audio data. MakeSinger enables the training of the diffusion-based SVS model from any speech and singing voice data regardless of its labeling, thereby enhancing the quality of generated voices with large amount of unlabeled data. At inference, our novel dual guiding mechanism gives text and pitch guidance on the reverse diffusion step by estimating the score of masked input. Experimental results show that the model trained in a semi-supervised manner outperforms other baselines trained only on the labeled data in terms of pronunciation, pitch accuracy and overall quality. Furthermore, we demonstrate that by adding Text-to-Speech (TTS) data in training, the model can synthesize the singing voices of TTS speakers even without their singing voices.

Index Terms: singing voice synthesis, semi-supervised training, diffusion generative model

1. Introduction

Singing voice synthesis (SVS) is a rapidly growing field of research adopting various deep learning techniques. Similar to Text-to-Speech (TTS), a typical two-stage pipeline of neural SVS consists of an acoustic model and a vocoder. The acoustic model [1, 2, 3, 4] generates acoustic features such as mel-spectrogram from the musical score, including text, tempo, and pitch. The vocoder [5, 6, 7, 8] then converts these acoustic features to a waveform. Recent advances include end-to-end models that aim to improve the sound quality and naturalness such that the generated audio results in a more human-like voice [9, 10, 11, 12].

However, training such models demands a high-quality dataset with precisely aligned (*text*, *pitch*, *audio*) triplets. This makes collecting data for SVS challenging, as it involves professional singers who can accurately sing to the pitch and tempo of the musical score. Semi-supervised approaches have emerged to address this issue, leveraging extracted pseudo labels from unlabeled data. For instance, [13] employed Harvest [14] to generate pitch labels, and phoneme classifier for text labels. DeepSinger [15] gathered unlabeled data from publicly available music websites while extracting the pitch using Parselmouth¹. Despite their effectiveness, these approaches come with a significant limitation: they are not entirely free from data labeling. They still require (*text*, *audio*) pair for the training data, or require an auxiliary classifier specifically tailored for

singing voices, which eventually demands a large amount of labeled data to prevent potential error propagation from inaccurate labeling.

Meanwhile, diffusion generative models [16, 17, 18] have shown significant breakthroughs in many areas of deep learning. Notably, studies such as [19, 20, 21] have demonstrated the remarkable effectiveness of diffusion models in TTS and SVS. Additionally, [22, 23] introduced a training method that capitalizes on the strengths of diffusion models, incorporating the training of a separate phoneme classifier to provide text guidance. To avoid building the classifier, [24] introduced classifier-free guidance, enabling a model to learn from both conditional and unconditional data and generate class-conditional samples by a single joint model. These guiding approaches facilitate efficient learning from unlabeled datasets.

Inspired by these works, we propose MakeSinger, a semi-supervised training method for data-efficient singing voice synthesis that leverages any speech or singing voice, regardless of its labeling. MakeSinger employs classifier-free guidance, taking advantage of the joint model's capability to learn from both labeled and unlabeled data. We also present a novel dual guiding mechanism that adjusts the balance between text and pitch guidance during reverse diffusion steps. Our experimental results indicate that incorporating a substantial amount of unlabeled singing voice data into training enhances the pronunciation and overall quality of the synthesized voices, outperforming baselines supervised with more labeled songs. Furthermore, by leveraging the multi-speaker TTS dataset, we demonstrate MakeSinger's ability to generate the singing voices of diverse TTS speakers, even without specific singing voice data for those speakers. We present our samples generated by MakeSinger on our demo page².

2. Proposed Method

In this section, we describe MakeSinger, our proposed semi-supervised training method for SVS. As a base architecture, we use Grad-TTS [19], a score-based TTS model, with several modifications. In training procedure, our diffusion model learns from any audio without relying on its labeling, contributing the robustness of the system in the low-resource scenario with a limited availability of paired data. During inference, we apply a dual guiding mechanism to provide a balanced and appropriate level of pitch and text guidance. We explain detailed architecture of MakeSinger in Section 2.1, and our dual guiding mechanism in Section 2.2. The comprehensive overview of the MakeSinger is illustrated in Figure 1.

¹<https://github.com/YannickJadoul/Parselmouth>

²<https://makesinger.github.io/MakeSinger-demo>

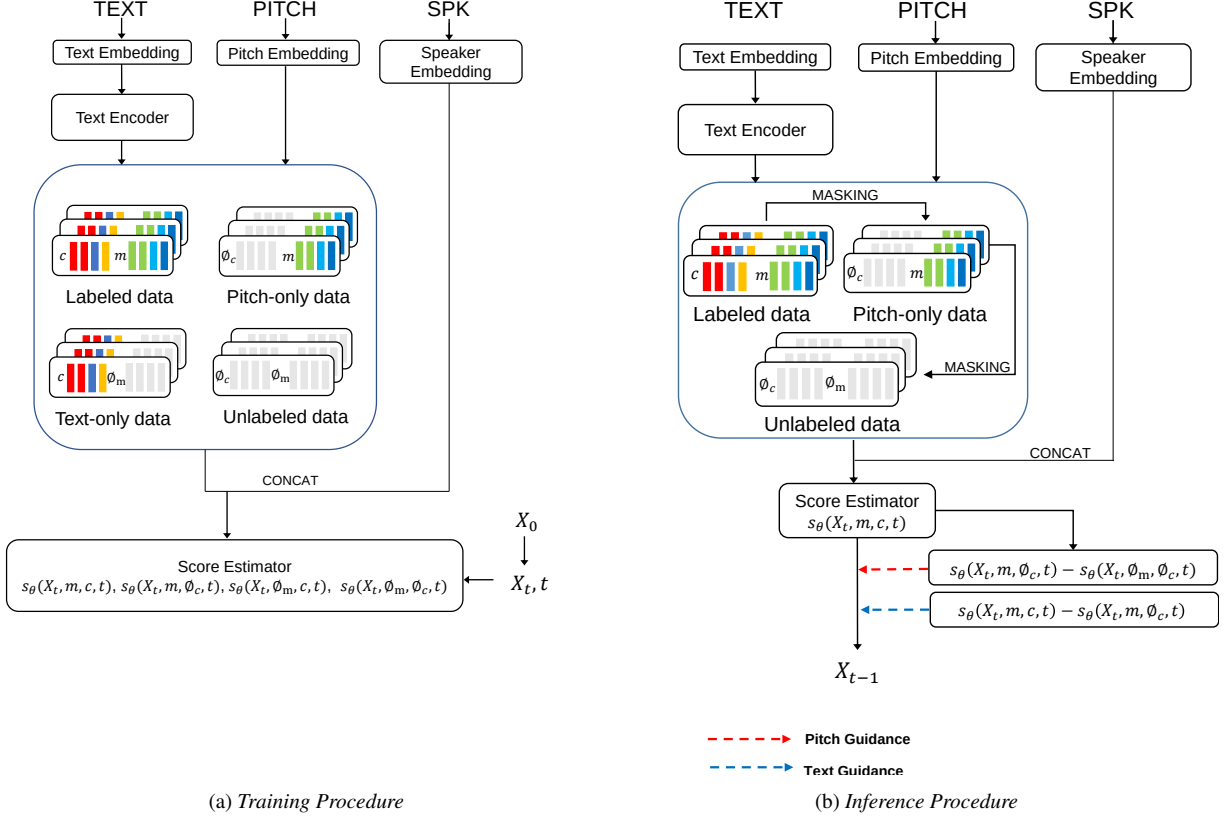


Figure 1: Overall procedure of MakeSinger.

2.1. Architecture

MakeSinger is designed to generate a mel-spectrogram corresponding to the given text, pitch, and speaker. When the model learns from labeled data, temporally aligned text and pitch sequence along with its mel-spectrogram becomes an input. In case where neither text nor pitch label exists, we use $\langle unknown \rangle$ token sequence as substituting label. MakeSinger consists of two main parts: encoder and score estimator.

Encoder: The encoder part converts text, pitch, and speaker information into the feature sequence of mel-spectrogram frame length. Text encoder consists of a pre-net, 6 Transformers [25], and a linear projection layer. All the feature sequences are concatenated together to give conditional information to the score estimator. As in Figure 1, we denote the feature sequences extracted from the text and pitch as c and m . We additionally denote the feature sequence extracted from $\langle unknown \rangle$ text token as $\emptyset_c$ and $\langle unknown \rangle$ pitch token as $\emptyset_m$. Unlike in GradTTS, our encoder solely focuses on encoding the conditional information without estimating the mean and variance of the prior noise distribution, as the strong prior information can diminish the effect of the guiding process in Section 2.2.

Score estimator: The score estimator s_θ , which has U-Net [26] architecture, predicts the score at the given diffused mel-spectrogram X_t and the concatenated feature sequence, where t stands for the diffusion timestep. When training the model, we first pick uniformly random t between 0 and 1. Then, we transform the ground-truth mel-spectrogram X_0 to the diffused mel-spectrogram X_t through a forward diffusion process. We leverage score-matching loss, an L2 loss between the predicted and computed scores of X_t to train the score estimator.

2.2. Dual guiding in diffusion generative model

Original diffusion guiding mechanism [18] employs an additional classifier that demands enough labeled data to train the classifier as in Guided-TTS [22]. To address the low resource problem of the singing voice, we adopt a classifier-free guiding from [24] in a semi-supervised manner. This method involves providing guidance at each reverse diffusion step, facilitating class-conditional generation.

In our guiding scheme for generating singing voices, we recognize the necessity to manage underlying textual and pitch information. A simple approach is modeling the guidance for text and pitch with a single term. However, this approach hinders separate control over text and pitch, limiting the model’s flexibility. To ensure finer control, we opt for dual guidance, applying distinct guidance terms for text and pitch. We formulate the dual guidance for diffused mel-spectrogram X_t , diffusion timestep t , and conditional information y . We start from a general diffusion guiding formulation from [18]:

$$\nabla_{X_t} \log p_t(X_t|y) = \nabla_{X_t} \log p_t(X_t) + \nabla_{X_t} \log p_t(y|X_t). \quad (1)$$

We split y to pitch condition m and text condition c . Then, applying Bayes’ rule, we can get classifier-free guidance for both labels:

$$\begin{aligned} & \nabla_{X_t} \log p_t(m, c|X_t) \\ &= \nabla_{X_t} \log p_t(X_t|m, c) - \nabla_{X_t} \log p_t(X_t) \\ &= (\nabla_{X_t} \log p_t(X_t|m, c) - \nabla_{X_t} \log p_t(X_t|m)) \\ &+ (\nabla_{X_t} \log p_t(X_t|m) - \nabla_{X_t} \log p_t(X_t)). \end{aligned} \quad (2)$$

Finally, we can derive the score using dual guiding with scaling factors:

$$\begin{aligned} s_t &= \nabla_{X_t} \log p_t(X_t|m, c) \\ &+ w_1(\nabla_{X_t} \log p_t(X_t|m, c) - \nabla_{X_t} \log p_t(X_t|m)) \\ &+ w_2(\nabla_{X_t} \log p_t(X_t|m) - \nabla_{X_t} \log p_t(X_t)), \end{aligned} \quad (3)$$

where s_t is a time-dependent total score for the reverse time process, w_1 and w_2 are scaling factors for text and pitch guidance. By a similar approach, we can also derive the following:

$$\begin{aligned} s_t &= \nabla_{X_t} \log p_t(X_t|m, c) \\ &+ w_1(\nabla_{X_t} \log p_t(X_t|m, c) - \nabla_{X_t} \log p_t(X_t|c)) \\ &+ w_2(\nabla_{X_t} \log p_t(X_t|c) - \nabla_{X_t} \log p_t(X_t)). \end{aligned} \quad (4)$$

Note that we provide guidance to the score estimated from conditional input since it generates better samples compared to giving guidance to the unconditional score. Actual implementation for Eq. 3 is expressed as follows:

$$\begin{aligned} s_t &= s_\theta(X_t, m, c, t) + \alpha_t^1 w_1(s_\theta(X_t, m, c, t) \\ &- s_\theta(X_t, m, \emptyset_c, t)) + \alpha_t^2 w_2(s_\theta(X_t, m, \emptyset_c, t) \\ &- s_\theta(X_t, \emptyset_m, \emptyset_c, t)). \end{aligned} \quad (5)$$

We emphasize that a single joint model trained with both conditional and unconditional data estimates all the score terms. The model masks the original feature sequence to pitch-only and unlabeled data sequences to estimate each guidance, as depicted in Figure 1b. By employing norm-based guidance from [22], we optimize the control by scaling each norm of the guiding gradient in proportion to the norm of the score of the joint model in every diffusion timestep t . α_t^1 and α_t^2 in Eq. 5 are normalizing factors of the norm-based guidance.

3. Experimental Settings

3.1. Datasets

We conducted multi-speaker SVS experiments on the open-sourced Korean multi-speaker singing voice dataset³. We employed 1,210 ballad songs sung by 12 singers. We spitted it into 973 songs for training and 277 songs for the test. We also conducted SVS experiments utilizing TTS data from open-sourced large-scale Korean multi-speaker dataset⁴, aiming to synthesis singing voices of TTS speakers. All audio files were resampled to 22,050 Hz and we computed the mel-spectrograms using a 1,024-point short-time Fourier transform with 256-point hop size and mel-scale filterbanks featuring 80 bins.

3.2. Input Processing

Our dataset is paired with musical instrument digital interface (MIDI) files that serve as labels. As MIDI file contains aligned pitch and text expressed in syllables according to the tempo, we extracted m , c , f^s , and f^e of length I from each song. m , c denotes sequence of pitches and Korean syllables respectively, and f_i^s , f_i^e correspond to start and end frame index of i th pitch and syllable. f_i^s , f_i^e is computed by considering the frequency of waves and mel-spectrogram frame length. Since a Korean syllable can be divided into three phonemes: an onset, a nucleus, and an optional coda, we allocated frames for each

phoneme. In Korean, most vocalization time spends on pronouncing vowels, which is always the nucleus. Following the approach of [4, 9, 27], we allocated three frames for the onset and coda, the rest for the nucleus while dividing the frames of a syllable.

For TTS data, We used Montreal Forced Aligner (MFA) [28], which automatically extracts speech-phoneme alignments, to extend syllables to mel-spectrogram length. In order to standardize the phoneme set for both the SVS and TTS, we applied the same allocation of onset, nucleus, and coda as we did for the singing voice data.

3.3. Model Configuration

We trained the following models for comparison in the multi-speaker SVS experiments.

- **MLP-Singer**: MLP-Mixer [29] based Korean SVS model.
- **VISinger**: End-to-end SVS model based on VITS [30]. Given that our experimental data were pre-aligned, we removed the duration predictor and length regulator from the original model. Additionally, we substituted its encoder to ours for fair comparison.
- **MakeSinger without semi-supervision**: Our proposed model only supervised with labeled data.
- **MakeSinger with semi-supervision**: Our proposed model with semi-supervised training. Dual guiding was applied in the inference procedure.

To demonstrate the effectiveness of our proposed method and its correlation with the amount of labeled data, we varied the number of labeled songs across three different settings, maintaining constant 12 singers: 12, 36, and 108 songs. MLP-Singer, VISinger, and MakeSinger without semi-supervision exclusively utilized labeled songs for training.

For the MakeSinger with semi-supervision, we extended the training set by including the remaining unlabeled songs. Specifically, we used [12, 961], [36, 937], [108, 865] labeled and unlabeled songs. During inference, we employed Eq. 3 from Section 2 for reverse diffusion steps. To improve the gradient estimation of $\nabla_x \log p_t(X_t|m)$, we introduced pitch-only data in the training data created by masking text labels of the labeled data. We empirically configured the scaling factor for text guidance $w_1 = 0.2$ and pitch guidance $w_2 = 0.02$ in the dual guiding process.

Both MakeSinger models used 200 reverse diffusion steps and shared hyperparameter settings with Grad-TTS. Unlike [18, 19] that used an ordinary differential equation (ODE) without the stochastic term in the inference process, we opted for a stochastic differential equation (SDE) to better capture the vibrato commonly found at the end of sentences in ballad songs.

As MLP-Singer and MakeSinger generate mel-spectrogram, we used Hifi-GAN [7] to generate the waveform. We fine-tuned its universal pre-trained version⁵ to our singing voice dataset since Hifi-GAN is originally trained to synthesis speech. MLP-Singer is trained on a single 2080Ti GPU. All the other models were trained up to 60 epochs on 4 2080Ti GPUs.

4. Results

4.1. Main Results

To verify above four systems in three different settings, we performed a subjective evaluation using the mean opinion

³<https://bit.ly/3uZc7WM>

⁴<https://bit.ly/4c46sz6>

⁵<https://github.com/jik876/hifi-gan>

score (MOS) test, and for objective evaluation, we utilized F0 root mean square error (F0-RMSE) and semitone accuracy (S-ACC). Results are presented in Table. 1

Subjective Evaluation Metrics: The MOS test involved 14 participants and was conducted with 95 % confidence intervals. We evaluated 24 randomly sampled song segments each 4 to 6 seconds long, focusing on pronunciation accuracy (MOS-P) and overall quality (MOS-Q). For pronunciation accuracy, we instructed testers to evaluate how naturally the singer pronounced the text in terms of ballad singing. For overall quality, we instructed testers to evaluate the comprehensive quality of singing, including sound quality, pronunciation, pitch accuracy, and naturalness.

Objective Evaluation Metrics: We computed F0-RMSE and S-ACC against reference audios reconstructed by a vocoder. S-ACC measures how accurately a system can determine the pitch of a musical note in semitones. Randomly selected 350 segments of songs between 4 to 6 seconds were used to compute average score.

Results: As shown in Table. 1, even MakeSinger without semi-supervision outperformed other baseline systems for every metric when trained with 36 and 108 songs. Furthermore, MakeSinger with semi-supervision achieved best scores across all evaluations, demonstrating the effectiveness of dual guiding mechanism, especially in pitch accuracy. Remarkably, with just 12 or 36 labeled songs, the semi-supervised MakeSinger showed comparable or superior performance to other baseline models trained on 108 labeled songs in every metric, highlighting our model’s data efficiency. This efficiency was most pronounced with the smallest labeled dataset, where the semi-supervised model showed significant improvement over its non-semi-supervised counterpart. The performance gap between the two proposed models becomes smaller as the quantity of labeled songs increased, which aligns with the expectation that the importance of guiding decreases as more labeled data becomes available for the model to learn from.

4.2. Ablation study

To validate our dual guiding mechanism, we carried out an ablation study on the same semi-supervised MakeSinger model, comparing three different approaches during inference: dual guiding, single guiding, and no guiding. Single guiding approach models text and pitch guidance in a single term, which is expressed as follows:

$$s_t = s_\theta(X_t, m, c, t) + \alpha_t w(s_\theta(X_t, m, c, t) - s_\theta(X_t, \emptyset_m, \emptyset_c, t)). \quad (6)$$

We set $w = 0.2$ in the single guiding to match the scale of guidance. The result in Table 2 shows that the single guiding mechanism degraded MakeSinger’s pronunciation and overall quality, whereas completely omitting the guiding mechanism led to further degradation.

4.3. Singing Voice Synthesis based on TTS data

Utilizing TTS data helps MakeSinger to generate singing voices in TTS speakers’ voices while maintaining the characteristics and timbre of the TTS speaker. Exploiting the high quality and high coverage of multi-speaker TTS data can also be a breakthrough to the insufficient number of singers and songs in the SVS dataset. Eq. 4 is more suitable for the task, as the model learns to estimate $\nabla_x \log p_t(X_t|c)$ from the TTS data, which only has text labels.

Table 1: *Experimental results in terms of MOS-P, MOS-Q, F0-RMSE, and S-ACC.*

Method	MOS-P	MOS-Q	F0-RMSE (↓)	S-ACC (↑)
Reconstruction	4.64±0.07	4.47±0.08	-	-
12 labeled songs				
MLP-singer	2.79±0.11	2.52±0.10	0.2468	41.76
VISinger	2.20±0.11	2.19±0.09	0.3005	37.14
MakeSinger				
- w/o semi-sup.	2.37±0.11	2.54±0.11	0.2543	48.98
- w/ semi-sup.	3.30±0.12	3.48±0.12	0.1736	54.29
36 labeled songs				
MLP-singer	3.14±0.11	2.71±0.09	0.2419	42.95
VISinger	2.74±0.11	2.59±0.09	0.2121	45.91
MakeSinger				
- w/o semi-sup.	3.41±0.11	3.36±0.11	0.1942	52.51
- w/ semi-sup.	3.61±0.12	3.66±0.12	0.1720	54.25
108 labeled songs				
MLP-singer	3.45±0.11	3.01±0.09	0.2233	46.06
VISinger	3.51±0.11	3.33±0.10	0.2059	50.98
MakeSinger				
- w/o semi-sup.	3.95±0.10	3.76±0.09	0.1971	52.57
- w/ semi-sup.	4.03±0.09	4.08±0.09	0.1702	56.05

Table 2: *Ablation study with MOS-P and MOS-Q.*

Method	MOS-P	MOS-Q
Reconstruction	4.54±0.10	4.49±0.10
MakeSinger		
- no guiding	3.80±0.10	3.77±0.07
- single guiding	3.87±0.10	3.84±0.07
- dual guiding	3.95±0.10	3.94±0.06

To show the feasibility of generating singing voices of TTS speakers, we incorporated multi-speaker TTS data into the training process. This involved using 487 labeled songs along with the remaining unlabeled songs and 3 hours of TTS data uttered by 80 speakers. Results show that MakeSinger is capable of generating singing voices from TTS speakers with dependable quality, as presented in our demo page.

5. Conclusion

We presented MakeSinger, a diffusion-based semi-supervised training method for singing voice synthesis. Exploiting the advantage of the classifier-free guidance, we could utilize any human speech and singing voice data regardless of its labeling. Furthermore, during the reverse diffusion steps, our novel dual guiding mechanism provides better controllability to the model regarding text and pitch. To the best of our knowledge, MakeSinger is the first SVS model to leverage classifier-free diffusion guiding in a semi-supervised manner. By applying our method, we showed that the model can generate singing voices with better pronunciation and sound quality. Also, we demonstrated that MakeSinger generates songs in the voice of TTS speakers by utilizing multi-speaker TTS data in training.

6. Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-00456, Development of Ultra-high Speech Quality Technology for Remote Multi-speaker Conference System)

7. References

- [1] M. Blaauw and J. Bonada, "Sequence-to-sequence singing synthesis using the feed-forward transformer," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7229–7233.
- [2] G.-H. Lee, T.-W. Kim, H. Bae, M.-J. Lee, Y.-I. Kim, and H.-Y. Cho, "N-singer: A non-autoregressive korean singing voice synthesis system for pronunciation enhancement," *arXiv preprint arXiv:2106.15205*, 2021.
- [3] Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Sinsy: A deep neural network-based singing voice synthesis system," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2803–2815, 2021.
- [4] J. Tae, H. Kim, and Y. Lee, "Mlp singer: Towards rapid parallel korean singing voice synthesis," in *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2021, pp. 1–6.
- [5] A. van den 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," in *Arxiv*, 2016. [Online]. Available: <https://arxiv.org/abs/1609.03499>
- [6] 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 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.
- [7] 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. 17 022–17 033, 2020.
- [8] R. Huang, F. Chen, Y. Ren, J. Liu, C. Cui, and Z. Zhao, "Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3945–3954.
- [9] J. Lee, H.-S. Choi, C.-B. Jeon, J. Koo, and K. Lee, "Adversarially trained end-to-end korean singing voice synthesis system," in *Interspeech*, 2019.
- [10] J. Chen, X. Tan, J. Luan, T. Qin, and T.-Y. Liu, "Hifisinger: Towards high-fidelity neural singing voice synthesis," *arXiv preprint arXiv:2009.01776*, 2020.
- [11] Y. Zhang, J. Cong, H. Xue, L. Xie, P. Zhu, and M. Bi, "Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7237–7241.
- [12] Y. Zhang, H. Xue, H. Li, L. Xie, T. Guo, R. Zhang, and C. Gong, "Visinger 2: High-fidelity end-to-end singing voice synthesis enhanced by digital signal processing synthesizer," *arXiv preprint arXiv:2211.02903*, 2022.
- [13] S. Choi and J. Nam, "A melody-unsupervision model for singing voice synthesis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7242–7246.
- [14] M. Morise *et al.*, "Harvest: A high-performance fundamental frequency estimator from speech signals," in *INTERSPEECH*, 2017, pp. 2321–2325.
- [15] Y. Ren, X. Tan, T. Qin, J. Luan, Z. Zhao, and T.-Y. Liu, "DeepSinger: Singing voice synthesis with data mined from the web," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1979–1989.
- [16] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.
- [17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
- [18] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2020.
- [19] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8599–8608.
- [20] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "DiffSinger: Singing voice synthesis via shallow diffusion mechanism," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 020–11 028.
- [21] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim, "Diff-tts: A denoising diffusion model for text-to-speech," *arXiv preprint arXiv:2104.01409*, 2021.
- [22] H. Kim, S. Kim, and S. Yoon, "Guided-tts: A diffusion model for text-to-speech via classifier guidance," in *International Conference on Machine Learning*. PMLR, 2022, pp. 11 119–11 133.
- [23] 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.
- [24] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.
- [25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention-MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.
- [27] S. Choi, W. Kim, S. Park, S. Yong, and J. Nam, "Korean singing voice synthesis based on auto-regressive boundary equilibrium gan," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7234–7238.
- [28] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kald," in *Interspeech*, vol. 2017, 2017, pp. 498–502.
- [29] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit *et al.*, "Mlp-mixer: An all-mlp architecture for vision," *Advances in neural information processing systems*, vol. 34, pp. 24 261–24 272, 2021.
- [30] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.