ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888185

VoiceGuider: Enhancing Out-of-Domain Performance in Parameter-Efficient Speaker-Adaptive Text-to-Speech via Autoguidance

Jiheum Yeom
Electrical and Computer Engineering
Seoul National University
Seoul, Republic of Korea
quilava1234@snu.ac.kr

Heeseung Kim
Electrical and Computer Engineering
Seoul National University
Seoul, Republic of Korea
gmltmd789@snu.ac.kr

Jooyoung Choi
Electrical and Computer Engineering
Seoul National University
Seoul, Republic of Korea
jy_choi@snu.ac.kr

Che Hyun Lee
Electrical and Computer Engineering
Seoul National University
Seoul, Republic of Korea
saga1214@snu.ac.kr

Nohil Park
Electrical and Computer Engineering
Seoul National University
Seoul, Republic of Korea
pnoil2588@snu.ac.kr

Sungroh Yoon*
ECE, AIIS, ASRI, INMC, ISRC, and IPAI
Seoul National University
Seoul, Republic of Korea
sryoon@snu.ac.kr

***Abstract*—When applying parameter-efficient finetuning via LoRA onto speaker adaptive text-to-speech models, adaptation performance may decline compared to full-finetuned counterparts, especially for out-of-domain speakers. Here, we propose VoiceGuider, a parameter-efficient speaker adaptive text-to-speech system reinforced with autoguidance to enhance the speaker adaptation performance, reducing the gap against full-finetuned models. We carefully explore various ways of strengthening autoguidance, ultimately finding the optimal strategy. VoiceGuider as a result shows robust adaptation performance especially on extreme out-of-domain speech data. We provide audible samples in our demo page.**

***Index Terms*—text-to-speech, speaker adaptive text-to-speech, diffusion, autoguidance, parameter-efficient text-to-speech**

I. Introduction

While deep generative models, such as diffusion models [1]–[3], have consistently demonstrated impactful advancements [4]–[6], text-to-speech (TTS) models have mirrored these achievements, advancing significantly in both quality and efficiency [7], [8]. Speaker-adaptive TTS, which aims to synthesize speech from speakers unseen during pretraining, reflects this trend of innovation. These advancements are primarily categorized into two approaches: zero-shot adaptation [9]–[14] and few-shot adaptation [9], [15]–[18]. While zero-shot adaptation eliminates the need for additional training, it typically requires substantial investments in training resources, model size, and dataset comprehensiveness.

Conversely, few-shot adaptation models are generally recognized to outperform their zero-shot counterparts while significantly reducing training costs [9], [18]. Studies have demonstrated personalization using several minutes to as little as 5$\sim$10 seconds of reference data per target speaker [9], [15], [16], [18], where the latter is also acknowledged as one-shot adaptation model. Furthermore, some researches aim to achieve speaker adaptation with minimal parameter increments and performance drops [16], [17], utilizing methods such as Low-Rank Adaptation (LoRA) [19]. Among these, VoiceTailor [20] successfully combines a pretrained diffusion-based decoder with LoRA to establish an efficient framework for training plug-in adapters for one-shot speaker-adaptive TTS.

While various parameter-efficient one-shot TTS models demonstrate strong adaptation performance, their evaluations are mostly conducted on in-domain data similar to datasets used for pretraining, often comprising well-constrained speech data such as audiobook recordings [21]–[23]. In real-world scenarios, demands for handling in-the-wild reference data grow, yet obtaining sufficient quantities of such data for pretraining is typically challenging. Therefore, achieving robustness to out-of-domain (OoD) data during speaker adaptation, particularly with in-the-wild recordings, remains a critical challenge. Although the previous baseline, VoiceTailor, shows reasonable speaker adaptation performance on in-domain data comparable to full-finetuning, significant performance degradation is observed when adapting to OoD data that substantially deviates from the pretraining distribution.

In this work, we propose VoiceGuider, a parameter-efficient one-shot TTS model designed to robustly adapt to OoD reference data. We utilize VoiceTailor [20], a parameter-efficient one-shot TTS model, as our backbone model. By incorporating autoguidance [24], which enhances conditioning by guiding generation with a degraded model, we aim to improve speaker adaptation performance for reference data significantly far

* Corresponding Author

This work was supported by Samsung Electronics Co., Ltd (IO240124-08661-01) and Korean Government (2022R1A3B1077720, 2022R1A5A708390811, RS-2022-II220959, BK21 Four Program, IITP-2024-RS-2024-00397085 & RS-2021-II211343: AI Graduate School Program)

Authorized licensed use limited to: Seoul National University. Downloaded on November 08,2025 at 01:35:08 UTC from IEEE Xplore. Restrictions apply.

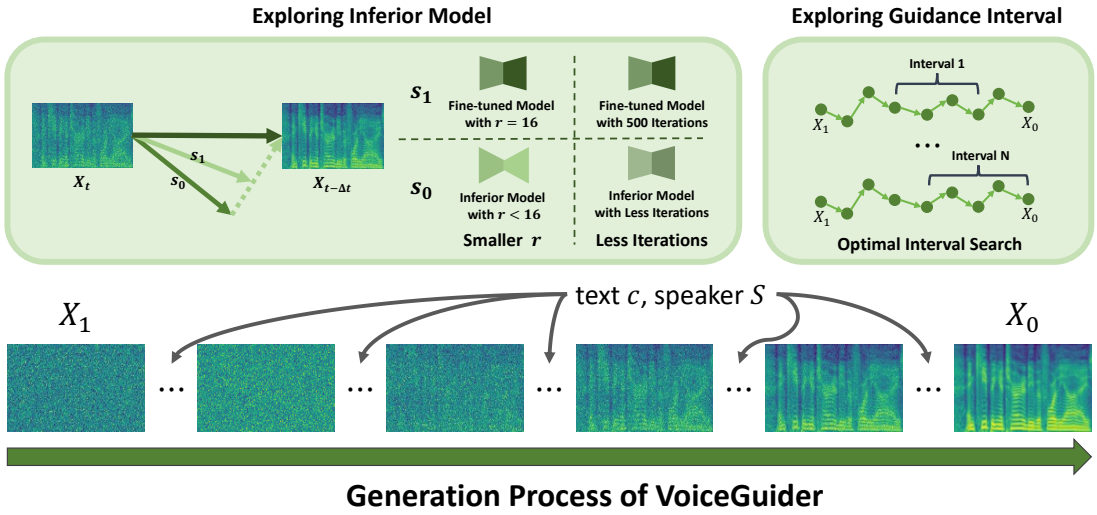

Fig. 1. **Overview of VoiceGuider**. VoiceGuider is a parameter-efficient one-shot TTS method for out-of-domain speakers. To eliminate errors caused by the parameter-efficient adapter, VoiceGuider enhances the prediction of the adapted model s_1 by pushing away from the predictions of the inferior model s_0. We explore two key ingredients of VoiceGuider: the method for obtaining the inferior model (left) and the guidance interval (right).

from the training distribution, such as 'in-the-wild' recordings. In addition to exploring degraded model candidates proposed in [24], we analyze various candidates tailored to parameter-efficient models to identify the optimal autoguidance strategy for VoiceTailor. Furthermore, by examining the integration of autoguidance at different stages of the generation process, VoiceGuider effectively incorporates target speaker information, thereby enhancing speaker adaptation for OoD data.

We demonstrate that VoiceGuider achieves comparable performance to full-finetuning baselines on GigaSpeech [25], a dataset including in-the-wild data, while surpassing parameter-efficient one-shot baseline. Additionally, we assess the effectiveness of various guidance techniques for parameter-efficient one-shot TTS by comparing different degraded model candidates proposed for autoguidance. The robustness of our model across various in-the-wild data samples, and the samples used in our evaluations, is showcased on our demo page[1].

II. Methodology

In this section, we explain the background, baseline model, and the methods used to alleviate OoD performance degradation problem in parameter-efficient speaker adaptive TTS.

A. Background

Denoising Diffusion. Diffusion models [2] are generative models that add Gaussian noise to data in multiple steps and learn a denoising process to generate data. In the case of speaker-adaptive TTS, given a text embedding c and a speaker embedding S, the diffusion models are trained to recover the noise ϵ_t added to the noisy mel-spectrogram X_t using the following objective function:

$$L(\theta) = \mathbb{E}_{t,X_0,\epsilon_t}[\|\sqrt{1-\lambda_t}s_\theta(X_t|c,S) + \epsilon_t\|_2^2], \quad (1)$$

where s_θ is a diffusion model and λ_t is a predefined noise schedule of GradTTS [7] and $t \sim [0,1]$ indicates noise level.

Diffusion Guidance. Recently, diffusion models employ classifier-free guidance (CFG) [5] to improve sample quality and the likelihood of given conditions. At each generation step, CFG modifies the model's prediction with an extrapolation between two predictions:

$$\hat{s}_\gamma(X_t|c,S) = s_\theta(X_t|c,S) + \gamma(s_\theta(X_t|c,S) - s_\theta(X_t|c,\emptyset)), \quad (2)$$

where γ is a guidance scale. CFG pushes away the unconditional distribution to *avoid* undesired speaker, thereby increasing the likelihood of speaker S.

B. VoiceTailor: Parameter-Efficient Baseline

Given a pretrained diffusion model, VoiceTailor [20] trains a parameter-efficient adapter, namely LoRA [19], to adapt to a new speaker using (1). For each linear layer W_0 in attention modules of the pretrained diffusion model, VoiceTailor trains only a new matrix $\Delta W = BA$ where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, thus $W = W_0 + \alpha \cdot BA$ with scale α. By choosing small rank r, VoiceTailor trains only 0.25% of the whole parameters. Despite the small number of parameters, VoiceTailor achieves comparable speaker adaptation performance to the full-finetuning baseline, UnitSpeech [18].

However, while VoiceTailor demonstrates strong performance for in-domain speakers closer to the pretrained domain, we observe that it falls behind Unitspeech for OoD speakers.

C. VoiceGuider: Eliminating LoRA Error with Autoguidance

We argue that the limited capacity of the LoRA in VoiceTailor [20] results in prediction errors for OoD speakers, and these errors are amplified by the iterative generation process of the diffusion model, leading to subpar performance compared to full-finetuning. Given that CFG effectively steers away

[1]Demo: https://voiceguider.github.io/

Authorized licensed use limited to: Seoul National University. Downloaded on November 08,2025 at 01:35:08 UTC from IEEE Xplore. Restrictions apply.

TABLE I
DATASET SPECIFICATIONS. HERE, THE TERM **PRETRAIN** AND **IN-THE-WILD** REFERS TO WHETHER THE DATASET WAS USED FOR PRETRAINING AND CONTAINS IN-THE-WILD OOD DATA, RESPECTIVELY.

Dataset	Pretrain	OoD	In-the-wild
LibriTTS [23]	✓	×	×
VCTK [26]	×	✓	×
GigaSpeech [25]	×	✓	✓

from undesired samples, we propose VoiceGuider, a diffusion guidance to *eliminate LoRA errors* in speaker-adaptive TTS.

Autoguidance. A concurrent study on diffusion guidance [24] shows that a strong conditional model and an unconditional model share correlated errors. The study empirically shows that since the unconditional model's error is more overemphasized, the extrapolation in CFG *eliminates* this error. [24] further proposes autoguidance, which guides a strong model $s_1(X_t|c,S)$ with an inferior model $s_0(X_t|c,S)$ instead of the unconditional one $s_1(X_t|c,\emptyset)$, and demonstrates its effectiveness in conditional image synthesis. In our case of speaker-adaptive TTS, we combine CFG and autoguidance to amplify speaker likelihood and reduce LoRA errors:

$$\hat{s}_\gamma(X_t|c,S) = s_1(X_t|c,S) + \underbrace{\gamma_S(s_1(X_t|c,S) - s_1(X_t|c,\emptyset))}_{\text{Speaker likelihood } \uparrow} + \underbrace{\gamma_a(s_1(X_t|c,S) - s_0(X_t|c,S))}_{\text{LoRA error } \downarrow}, \tag{3}$$

where γ_S and γ_a denote scales of CFG and autoguidance. Furthermore, as illustrated in Fig. 1, we aim to identify the optimal autoguidance strategy for speaker-adaptive TTS from the following candidates of the inferior model s_0:

- **Shorter training time:** We construct s_0 using a model that consumed a shorter training time than s_1. This is the simplest approach to obtain inferior model, as it can be derived from the intermediate checkpoints of s_1.
- **Smaller LoRA rank:** We construct s_0 using a smaller LoRA rank size than s_1. Our baseline, VoiceTailor, demonstrated that smaller ranks, such as 2 and 4, are inferior compared to a rank size of 16.

Limited guidance interval. Along with exploring inferior models, we later demonstrate that autoguidance can be detrimental in certain interval of the generation process. Specifically, at high noise levels, the guidance may blindly push away from the data distribution, leading to mode dropping and degraded sample quality [27]. We further demonstrate that guidance in the range of $t \in [0.6, 1.0]$ is detrimental, and disabling both classifier-free guidance and autoguidance in this range reduces errors and achieves higher speaker similarity. Thus, both guidance scales γ_S and γ_a can be written as:

$$\gamma(t) = \begin{cases} \gamma & \text{if } t \in (t_{\text{lo}}, t_{\text{hi}}] \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

In the disabled interval, the inferior model s_0 in autoguidance can be seen as equivalent to the finetuned model s_1.

TABLE II
SECS RESULTS FOR PROBLEM STATEMENT VERIFICATION RESULTS.

Method	LibriTTS	VCTK	GigaSpeech
Full-finetuned [18]	0.948	0.873	0.902
LoRA-tuned [20]	0.940	0.854	0.875
Performance Gap	0.008	0.019	0.027

III. EXPERIMENTS AND RESULTS

A. Experimental Setup

1) Datasets: To first validate the performance degradation of parameter-efficient one-shot TTS in OoD reference data, we use three different TTS datasets. Each represents the in-domain dataset utilized during pretraining, the OoD dataset, and the extreme OoD dataset further apart from the in-domain such as 'in-the-wild' data. Details of the datasets are specified in Table I.

As shown, we pretrain our model with LibriTTS, thus positioned as the in-domain dataset. VCTK is not used for training, thus becomes the OoD dataset. GigaSpeech is not used for pretraining and includes in-the-wild speech data, becoming the extreme OoD data. After using these three datasets to validate the degradation problem, we use only the extreme OoD data, GigaSpeech, for the rest of the experiments.

2) Training details: Our main model builds upon the structure of VoiceTailor, and is finetuned for 500 iterations on a single NVIDIA A40 GPU. All finetuning is processed with Adam optimizer [28] at learning rate of 10^{-4}. For LoRA settings, we use rank r of 16 and scaling factor α of 8, equivalent to VoiceTailor. The inferior model for autoguidance is trained with the same hyperparameters of the main model except for the varying factors specified in Section III-D.

3) Evaluation settings: We use 50 sentences from the test set of each dataset for evaluation. Character error rate (CER) and speaker encoder cosine similarity (SECS) are evaluated each for pronunciation and speaker adaptation quality, using the CTC-based conformer [29] of the NEMO toolkit [30] and speaker encoder of Resemblyzer [31] package, respectively. For subjective evaluation, we evaluate with 5-scale mean opinion score (MOS) for naturalness and human preference tests asking which model sample resembles the reference more, regarding speaker adaptation quality compared to baseline.

For inferior model of VoiceGuider, we set rank $r = 1$, scaling factor $\alpha = 8$, and stop finetuning at 100 iterations, while finetuned model s_1 follows VoiceTailor's hyperparameters. For inference, we set both autoguidance scale γ_a and CFG scale γ_S to 1.0, and set guidance intervals $[t_{\text{lo}}, t_{\text{hi}}]$ to $[0.1, 0.6]$.

For baselines, we use UnitSpeech for the full-finetuned adaptation model and VoiceTailor for the LoRA-tuned baseline. All model outputs are reconstructed with BigVGAN [34], using the official checkpoint. When conducting subjective evaluation and preference tests, we additionally deploy open-source pretrained zero-shot adaptation models of XTTS *v2* [32] and CosyVoice [33] to compare against. During evalua-

Authorized licensed use limited to: Seoul National University. Downloaded on November 08,2025 at 01:35:08 UTC from IEEE Xplore. Restrictions apply.

TABLE III
MODEL COMPARSION OF VOICEGUIDER AGAINST VARIOUS BASELINES. BOLD TEXT INDICATES THE WINNING SIDE WITH STATISICAL SIGNIFICANCE (WILCOXON SIGNED-RANK TEST WITH $p < 0.05$).

Method	Speaker Similarity *win / tie / loss*	MOS	CER(%)
Ground Truth	45.6% / 4.2% / 50.2%	3.92 ± 0.10	2.31%
XTTS *v2* [32]	50.7% / 4.0% / 45.3%	3.85 ± 0.10	1.53%
CosyVoice [33]	48.0% / 5.1% / 46.9%	4.04 ± 0.09	4.24%
UnitSpeech [18]	41.6% / 11.6% / 46.9%	4.10 ± 0.09	1.53%
VoiceTailor [20]	**51.1%** / 9.6% / 39.3%	4.01 ± 0.09	1.56%
VoiceGuider	–	4.00 ± 0.09	1.72%

tion, all samples are normalized to the same volume of -27dB and resampled to 16kHz for a fair comparison.

B. Problem Statement Verification

To verify the performance degradation issue on OoD speakers, we compare SECS values of full-finetuned model to LoRA-tuned model on the three datasets of in-domain, OoD, and in-the-wild OoD. Table II shows the results on our experiments. The performance gap increases as we move further away in data domain, thus validating our problem statement.

C. Model Comparison

With the degradation problem verified, we continue on with main results of VoiceGuider, identifying whether our model succeeds in alleviating the issue. Table III shows the results of VoiceGuider compared to baseline models on GigaSpeech. Since GigaSpeech is consisted heavily of in-the-wild data, it is challenging to verify naturalness through generated results, as can be inferred from ground truth showing low MOS. VoiceGuider does show MOS equivalent to VoiceTailor, which indicates that VoiceGuider retains naturalness. For preference tests on speaker similarity, VoiceGuider gains the upperhand to VoiceTailor with a p-value of 0.009, implying superior speaker adaptation on in-the-wild OoD data. Additionally, VoiceGuider also shows comparable preference to the two zero-shot models, proving comparable performance to models which use heavy amounts of data for pretraining (27k hours for XTTS *v2*, 172k hours for CosyVoice).

D. Ablation Studies

Additionally, we conduct ablation experiments on various factors that may affect autoguidance used in VoiceGuider in Table IV. Among the hyperparameters of autoguidance explained in Section III-A3, we experiment on different values of rank r and number of finetuning iterations of the inferior model, analyzing on what setting of inferior model leads to the optimal results. We also perform ablation studies on the effects of autoguidance scale γ_a. We found best results at 100 iterations in terms of training iterations. Inferior models with smaller LoRA rank, on the other hand, showed no meaningful variance. Meanwhile, increasing the intensity of guidance through different gradient scale values, leads to increase of both CER and SECS. However, while SECS value converges at a certain saturation point, CER value continues to escalate.

TABLE IV
CER, SECS RESULTS ON VARIATIONS OF VOICEGUIDER. FULL-FINETUNING IMPLIES RESULTS OF APPLYING THE OPTIMAL AUTOGUIDANCE OF VOICEGUIDER INTO FULL-FINETUNED MODEL. BOLD VALUES INDICATE HYPERPARAMETERS USED FOR OUR FINAL MODEL.

		Text-to-Speech	
		CER(%) (↓)	**SECS** (↑)
# of Iterations	0	7.98	0.847
	100	1.71	0.891
	200	1.78	0.883
	300	1.78	0.879
	400	1.67	0.875
	500	1.69	0.872
Rank r	**1**	1.71	0.891
	2	1.83	0.891
	4	1.88	0.890
	8	1.69	0.891
Autoguidance Scale γ_a	0.00	1.63	0.877
	0.33	1.61	0.886
	0.66	1.69	0.890
	1.00	1.71	0.891
	1.33	2.00	0.892
Interval Upper t_{hi}	0.9	1.94	0.893
	0.8	1.84	0.892
	0.7	1.73	0.897
	0.6	1.72	0.902
	0.5	1.92	0.903
Interval Lower t_{lo}	0.5	2.09	0.901
	0.4	2.59	0.898
	0.3	1.68	0.900
	0.2	1.88	0.901
	0.1	1.72	0.902
Full-finetuning	-	1.89	0.904

We conducted tests not only on how to construct the weak model, but also on whether the guidance truly provides only positive influence across the whole generation process by controlling the upper boundary t_{hi} and lower boundary t_{lo} of the guidance interval. While the overall adaptation performance increases as the upper bound tightens, increasing the lower bound rather inflicts degradation of CER. When we applied the guidance methods of VoiceGuider onto the full-finetuned model, only minor amounts of performance increase were found, whereas VoiceGuider showed overall performance boosts, showing SECS equivalent to that of the full-finetuned model. These results indicate that VoiceGuider eliminates LoRA errors, reaching the performance of full-finetuning.

IV. CONCLUSION

We proposed VoiceGuider, a parameter-efficient speaker adaptive TTS model with robust adaptation performance on in-the-wild OoD speakers. We first verified the performance degradation issue of LoRA-tuned models with OoD in speaker adaptive TTS experimentally. VoiceGuider, reinforced with autoguidance, showed enhanced performance on OoD data, successfully alleviating the performance gap to full-finetuned counterparts. Based on our results, we expect VoiceGuider to provide effective ways of building efficient personalization TTS models for broad ranges of real-world speakers.

Authorized licensed use limited to: Seoul National University. Downloaded on November 08,2025 at 01:35:08 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 2256–2265.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851.

[3] 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*, 2021.

[4] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794.

[5] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[6] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 32 211–32 252.

[7] 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.

[8] H. Kim, S. Kim, and S. Yoon, "Guided-TTS: A diffusion model for text-to-speech via classifier guidance," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 11 119–11 133.

[9] 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.

[10] 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 *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 2709–2720.

[11] C. Wang, S. Chen, Y. Wu, Z.-H. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers," *ArXiv*, vol. abs/2301.02111, 2023.

[12] S. Kim, K. J. Shih, R. Badlani, J. F. Santos, E. Bakhturina, M. T. Desta, R. Valle, S. Yoon, and B. Catanzaro, "P-flow: A fast and data-efficient zero-shot TTS through speech prompting," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[13] K. Shen, Z. Ju, X. Tan, E. Liu, Y. Leng, L. He, T. Qin, sheng zhao, and J. Bian, "Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," in *The Twelfth International Conference on Learning Representations*, 2024.

[14] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, and W.-N. Hsu, "Voicebox: Text-guided multilingual universal speech generation at scale," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[15] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, sheng zhao, and T.-Y. Liu, "Adaspeech: Adaptive text to speech for custom voice," in *International Conference on Learning Representations*, 2021.

[16] Y. Yan, X. Tan, B. Li, T. Qin, S. Zhao, Y. Shen, and T.-Y. Liu, "Adaspeech 2: Adaptive text to speech with untranscribed data," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6613–6617.

[17] C.-P. Hsieh, S. Ghosh, and B. Ginsburg, "Adapter-based extension of multi-speaker text-to-speech model for new speakers," in *Interspeech*, 2023, pp. 3028–3032.

[18] H. Kim, S. Kim, J. Yeom, and S. Yoon, "Unitspeech: Speaker-adaptive speech synthesis with untranscribed data," in *Interspeech*, 2023, pp. 3038–3042.

[19] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[20] H. Kim, S. gil Lee, J. Yeom, C. H. Lee, S. Kim, and S. Yoon, "Voicetailor: Lightweight plug-in adapter for diffusion-based personalized text-to-speech," in *Interspeech*, 2024, pp. 4413–4417.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

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

[23] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Proc. Interspeech*, 2019, pp. 1526–1530.

[24] T. Karras, M. Aittala, T. Kynkäänniemi, J. Lehtinen, T. Aila, and S. Laine, "Guiding a diffusion model with a bad version of itself," *arXiv preprint arXiv:2406.02507*, 2024.

[25] G. Chen, S. Chai, G.-B. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Z. You, and Z. Yan, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," in *Interspeech*, 2021, pp. 3670–3674.

[26] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[27] T. Kynkäänniemi, M. Aittala, T. Karras, S. Laine, T. Aila, and J. Lehtinen, "Applying guidance in a limited interval improves sample and distribution quality in diffusion models," *arXiv preprint arXiv:2404.07724*, 2024.

[28] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

[29] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[30] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook *et al.*, "Nemo: a toolkit for building ai applications using neural modules," *arXiv preprint arXiv:1909.09577*, 2019.

[31] G. Louppe, "Resemblyzer," https://github.com/resemble-ai/Resemblyzer, 2019.

[32] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, and J. Weber, "Xtts: a massively multilingual zero-shot text-to-speech model," in *Interspeech*, 2024, pp. 4978–4982.

[33] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma, Z. Gao, and Z. Yan, "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," 2024.

[34] S. gil Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVGAN: A universal neural vocoder with large-scale training," in *The Eleventh International Conference on Learning Representations*, 2023.

Authorized licensed use limited to: Seoul National University. Downloaded on November 08,2025 at 01:35:08 UTC from IEEE Xplore. Restrictions apply.