



Parameter-Efficient Fine-Tuning for Low-Resource Text-to-Speech via Cross-Lingual Continual Learning

Ki-Joong Kwon¹, Jun-Ho So², Sang-Hoon Lee^{1†}

¹Department of Artificial Intelligence, Ajou University, South Korea

²Department of Mathematics, Ajou University, South Korea

{kijoongkwon, jhso, sanghoonlee}@ajou.ac.kr

Abstract

As generative models gain attention, it is crucial to adapt these models efficiently even with limited high-quality data and computational resources. In this work, we investigate a parameter-efficient fine-tuning (PEFT) for low-resource text-to-speech to transfer pre-trained knowledge to a new language leveraging only a single-speaker dataset and a single NVIDIA TITAN RTX GPU. We propose three types of adapters: *Conditioning Adapter*, *Prompt Adapter*, and *DiT LoRA Adapter*, where Conditioning Adapter enhances text embeddings, Prompt Adapter refines input representations, and DiT LoRA Adapter enables speech generation efficiency. We further explore the respective optimal configuration of adapters for single-speaker and multi-speaker scenarios. Consequently, under resource constraints, we successfully achieve effective adaptation to a new language using only 1.72% of the total parameters. Audio samples, source code and checkpoints will be available at <https://peft-tts.github.io/demo/>.

Index Terms: text-to-speech, adapter, TTS, PEFT, LoRA

1. Introduction

Neural text-to-speech (TTS) models [1, 2] have significantly advanced in recent years, leveraging deep learning techniques to achieve high-quality speech synthesis [3, 4, 5, 6, 7]. However, despite these advancements, adapting TTS models to new languages remains a fundamental challenge due to the substantial data requirements and computational costs associated with training.

To effectively adapt TTS models to a new language, a large and diverse dataset containing recordings from multiple speakers in that language is typically required. This requirement significantly increases the scale and cost of TTS model adaptation, as fine-tuning often demands a substantial amount of multi-speaker data to ensure robust generalization across different voices and recording conditions. Moreover, full fine-tuning of the entire TTS model is computationally expensive and parameter-inefficient, as it requires learning a new set of weights for each newly adapted language. Additionally, this process risks catastrophic forgetting, where the model may lose its previously learned language capabilities, reducing its effectiveness in multilingual applications.

In this paper, we show that a TTS model can be successfully adapted to a new language using **only 12 hours of a single-speaker training dataset and a single NVIDIA TITAN RTX GPU** with a significant reduction in computational cost compared to full fine-tuning. Our approach demonstrates that effective language adaptation can be achieved within a single-

speaker framework while also showing the potential for extension to multi-speaker TTS, providing a foundation for scalable and efficient cross-lingual adaptation. We build upon pre-trained F5-TTS [8], a flow-based multilingual zero-shot TTS model.

To achieve this, we introduce three adapter modules: **Conditioning Adapter**, **Prompt Adapter**, and **DiT LoRA Adapter**, which enable parameter-efficient fine-tuning (PEFT) while keeping the pre-trained model frozen. By leveraging these adapters, we efficiently integrate new linguistic features while minimizing the risk of catastrophic forgetting.

Furthermore, we observe that training with a single-speaker dataset introduces a trade-off between speaker similarity and linguistic accuracy, where improving one aspect may degrade the other. This trade-off becomes even more pronounced in multi-speaker scenarios, where the model must generalize across diverse speaker identities while maintaining linguistic consistency. To address this, we leverage Prompt Adapter to adjust this balance and further refine adaptation performance. Additionally, we incorporate **DropPath** [9] as a regularization mechanism to control this trade-off dynamically, thereby maintaining audio quality. The contributions of this work are as follows:

- We demonstrate that **unseen language adaptation is feasible** with only **12 hours of a single-speaker dataset and a single GPU**, challenging the need for large-scale multi-speaker data.
- We propose an adapter-based framework for cross-lingual TTS adaptation, preserving original language capabilities while enabling multi-speaker scalability.
- We analyze the impact of different adapter modules (Conditioning, Prompt, and DiT LoRA) through a comprehensive ablation study.
- We address the trade-off between speaker similarity and linguistic accuracy, using **Prompt Adapter and DropPath** to balance adaptation performance.

2. Related Work

Recent advances in parameter-efficient transfer learning (PETL) have played a crucial role in adapting large pre-trained models to new tasks with minimal computational cost. In natural language processing (NLP), methods such as LoRA (Low-Rank Adaptation) [10, 11] and Adapter Layers [12] enable fine-tuning by injecting lightweight trainable modules into transformer layers while keeping the base model frozen. These methods achieve performance comparable to full fine-tuning while significantly reducing computational costs.

In speech processing, PETL techniques have been applied

[†]Corresponding author

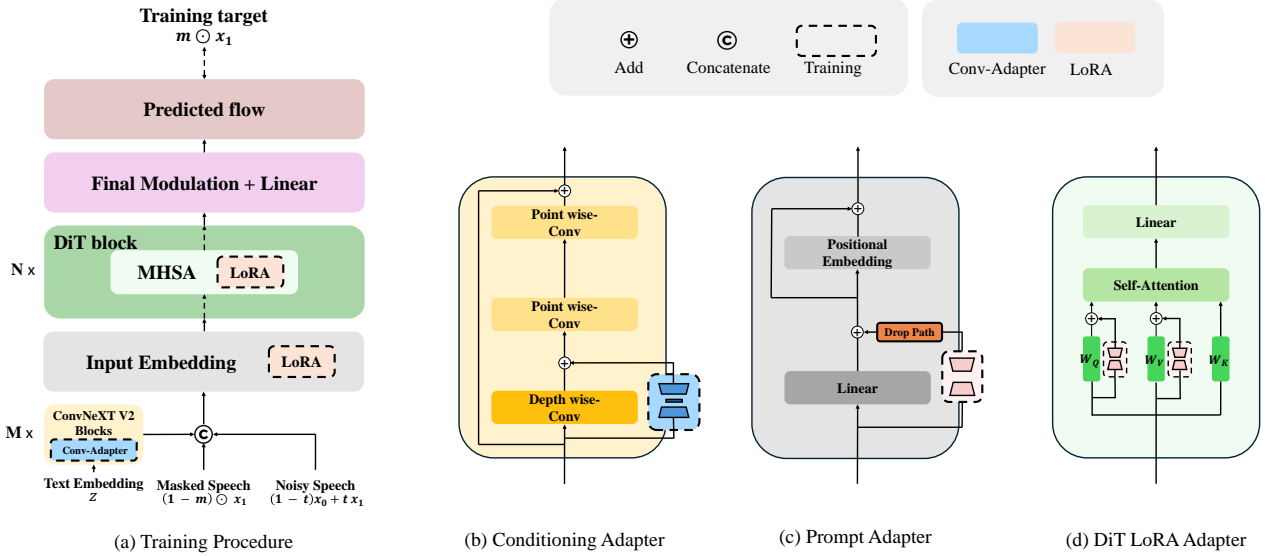


Figure 1: Overall frameworks of PEFT-TTS. (a) Based on the pre-trained F5-TTS, we added three adapters for cross-lingual continual learning. (b) Conditioning Adapter for ConvNeXt. (c) Prompt Adapter with DropPath for prompt tuning. (d) LoRA adapter for DiT.

to automatic speech recognition (ASR) [13, 14] and speaker adaptation [15]. Adapter-based approaches have been explored for domain adaptation, accent modeling, and low-resource language adaptation, demonstrating their effectiveness in adapting transformer-based speech encoders. Notably, ELP-Adapter [16] integrates adapter modules into transformer encoder layers, facilitating efficient fine-tuning for tasks such as speech recognition, speaker verification, and emotion recognition. Furthermore, Morioka et al. [17] introduced Residual Adapters for few-shot speaker adaptation, demonstrating that a model can be adapted to a new speaker with minimal data while requiring only 0.1% of additional trainable parameters. Mehrish et al. [18] explored ADAPTERMIX, a mixture of adapters designed for low-resource speaker adaptation, enabling effective speaker adaptation with less than one minute of training data. However, the application of PETL techniques in text-to-speech (TTS) remains relatively limited, particularly in cross-lingual adaptation scenarios, where effectively disentangling linguistic and speaker-specific features is crucial. Li et al. [19] proposed an efficient approach that integrates adapters and hypernetworks into a TTS architecture for multilingual speech synthesis, achieving comparable or superior performance to full fine-tuning while only updating approximately 2.5% of the model parameters. These studies highlight the potential of continual learning techniques in enhancing TTS systems for efficient adaptation to diverse languages and speakers, particularly in a low-resource environment where collecting large-scale multilingual datasets is impractical.

3. PEFT-TTS

3.1. F5-TTS: Flow-Based Multilingual TTS Architecture

F5-TTS [8] is a fully non-autoregressive (NAR) text-to-speech (TTS) model based on flow matching [20] with a Diffusion Transformer (DiT) [21]. Unlike autoregressive models that predict speech sequentially, F5-TTS enables parallelized speech synthesis, reducing inference latency while maintaining high synthesis quality. Furthermore, F5-TTS is trained using a text-guided speech-infilling task, where text is aligned with speech using a simple padding strategy proposed in E2-TTS [22], rather

than explicit phoneme-level alignment.

F5-TTS consists of three key components: Text Embedding, Input Embedding, and DiT Blocks.

- **Text Embedding:** Converts input into dense representations using ConvNeXt V2 [23] blocks, which enhance expressiveness by capturing both local and hierarchical linguistic patterns.
- **Input Embedding:** Concatenates the processed text embeddings with masked Mel-spectrogram and flow-matching latent variables, ensuring proper conditioning before being fed into the diffusion transformer.
- **DiT Blocks:** A stack of transformer-based diffusion layers that progressively refine the latent space, modeling long-range dependencies and natural prosody in speech generation.

Pre-trained on a multilingual corpus, F5-TTS captures cross-linguistic phonemic and prosodic patterns, making it ideal for cross-lingual adaptation. However, direct fine-tuning on a limited single-speaker dataset risks catastrophic forgetting of multi-speaker diversity, necessitating parameter-efficient strategies tailored to its architecture.

3.2. Parameter-Efficient Fine-Tuning with Adapter

As illustrated in Figure 1, we introduce three adapter modules to efficiently fine-tune F5-TTS for low-resource language adaptation while preserving pre-trained language capabilities.

3.2.1. Conditioning Adapter

We attach a Conv-Adapter [24] to the depth-wise convolution layers of the ConvNeXt V2-based text embedding module. The Conv-Adapter is designed with a depth-wise convolution followed by a point-wise convolution, and it incorporates squeeze-and-excitation (SE) parameter that modulates the feature responses. The compression factor γ significantly affects performance by balancing between parameter efficiency and adaptation capability.

3.2.2. Prompt Adapter

We apply LoRA to the linear projection layer which follows the concatenation of text and audio features. Notably, our training data consists of a single-speaker, our experiments indicate that this layer serves a more critical role than simple dimensional consistency, as its trainability directly influences the trade-off between pronunciation accuracy and speaker similarity. To further fine-tune this balance, we incorporate a DropPath mechanism, allowing more controlled adaptation to new languages. DropPath, a regularization technique that randomly drops residual paths during training, helps the network learn robust representations and avoid overfitting to new language or speaker features. This mechanism also plays a crucial role in maintaining audio quality.

3.2.3. DiT LoRA Adapter

We apply LoRA to DiT blocks, as they are most closely associated with speaker characteristics. This allows the model to preserve its pre-trained capabilities while effectively adapting to the training data. Notably, since the training data consists of a single-speaker, the adaptation process must be carefully managed to prevent the model from overfitting to speaker-specific traits while maintaining generalizability. This effect becomes even more pronounced in multi-speaker scenarios, where the model must balance speaker variation while retaining language adaptation quality. In our approach, we set the LoRA rank to 16, a relatively small value that helps maintain the model’s flexibility while limiting speaker-specific bias.

4. Experiments and Results

4.1. Experimental Setup

4.1.1. Dataset

For fine-tuning, we use the KSS (Korean Single Speaker) dataset [25], a 12.65-hour Korean speech corpus consisting of 12,853 samples recorded by a professional female voice actress. Compared to the English and Chinese subsets of the Emilia dataset [26], a 95K-hour multilingual corpus used for pre-training, the KSS dataset is significantly smaller and consists of a single-speaker, making the adaptation process more challenging. The dataset includes 44.1 kHz high-quality audio aligned with transcriptions. We downsampled the KSS dataset to 24 kHz to fine-tune the model, and split the dataset into 12,653 and 200 utterances for train and test subsets.

To evaluate the model’s ability to adapt to a single-speaker setting, we randomly create 100 test pairs with prompt and target speech pairs for inference. Additionally, to assess the model’s robustness in a multi-speaker scenario, we utilize 164 utterances (87 pairs) from AI Hub’s Korean Multi-Speaker Speech Synthesis Dataset [27], which includes a balanced male-to-female ratio, providing diverse phonetic and prosodic variations. Following F5-TTS, we simply estimate the total duration using the ratio of the text sequences in prompt and target.

4.1.2. Training

We fine-tune the F5-TTS model on the KSS dataset to adapt it to the Korean language while preserving its pre-trained multilingual capabilities. The base model consists of 22 layers with 16 attention heads and 1024/2048 embedding/feed-forward network (FFN) dimensions for the DiT blocks. The text embedding module is based on ConvNeXt V2 with 4 layers and 512/1024 embedding/FFN dimensions, totaling 335.8M parameters.

For fine-tuning, we trained the model for 355,000 steps with a batch size of 3,200 Mel-spectrogram frames, using the AdamW optimizer. The learning rate is set to 1e-5, following a linear warm-up, and then decays over the course of training. While training from scratch, we trained the model for 710,000 steps with 7.5e-5 learning rate. We apply gradient clipping with a max norm of 1.0 to stabilize optimization. Training is conducted on a single NVIDIA TITAN RTX GPU, leveraging the PEFT method to reduce computational costs.

The proposed method only updates 1.72% of the total parameters, significantly reducing memory consumption and computation time. This approach achieves approximately twice the training speed of full fine-tuning, as we do not compute gradients or maintain optimizer states for most parameters.

Unlike traditional Conv-Adapter, we set the compression factor to 0.25 and use a kernel size of 3, optimizing the trade-off between parameter efficiency and adaptation performance. DropPath with a rate of 0.3 is applied to enhance stability and prevent overfitting. For LoRA fine-tuning in the DiT blocks, we apply LoRA only to the query and value projection layers with a dropout rate of 0.05. The LoRA matrices are initialized using Kaiming uniform initialization with $A \sim U(-\sqrt{5}, \sqrt{5})$ and $B = 0$, ensuring stability during the early stages of training.

4.1.3. Evaluation

We evaluate the fine-tuned model using a cross-sentence synthesis task, where the model generates speech for a given reference text while mimicking the characteristics of a provided speech prompt. The evaluation includes both objective and subjective metrics to assess pronunciation accuracy and speaker similarity.

For objective evaluation, we measure character error rate (CER) and word error rate (WER) for linguistic accuracy utilizing Whisper-large-v3 [28] for Korean transcription. Speaker similarity is evaluated using SIM-O, which quantifies the acoustic resemblance between the synthesized speech and the reference speech. SIM-O is computed using WavLM [29] speaker representations to provide an accurate similarity score. Additionally, we report UTMOS [30], an automated MOS predictor, to provide further insights into synthesis quality.

For subjective evaluation, we adopt naturalness mean opinion score (NMOS) and similarity mean opinion score (SMOS). For NMOS, eight listeners are presented with randomly ordered synthesized speech. For SMOS, eight listeners are to score the similarity between the synthesized and prompt speech; note that SMOS is only conducted on the multi-speaker test set.

4.2. Single-speaker TTS Results

We compared PEFT-TTS models with F5-TTS using the same Korean dataset. Furthermore, we utilized an open-source zero-shot TTS model, CosyVoice 2 [31] trained with large-scale multilingual datasets including a Korean dataset. First, we found that training the F5-TTS model from scratch in low-resource environments using only a single GPU and a small dataset is challenging, often resulting in misalignment between text and speech. The results show that using a pre-trained F5-TTS with English and Chinese datasets enables learning a new language even with a single GPU and a small dataset. Table 1 shows the effectiveness of the proposed PEFT modules. Although we only utilized 1.72% of the trainable parameters compared to the baseline model, PEFT-TTS shows better pronunciation in terms of CER and WER while achieving a comparable performance in terms of similarity. Additionally, our model achieves better NMOS performance than the fully fine-tuned F5-TTS models.

Table 1: *Single-Speaker Korean TTS Results. * denotes large-scale multilingual text-to-speech models including Korean dataset. Full denotes fine-tuning all parameters, C-Adpt, P-Adpt, DP denote conditioning adapter, prompt adapter, DropPath, respectively.*

Methods	Pre-train	Text Enc. (γ)	Prompt Emb.	DiT (rank)	#Trainable Params	CER (\downarrow)	WER (\downarrow)	SIM-O (\uparrow)	UTMOS (\uparrow)	NMOS (\uparrow)
GT	-	-	-	-	-	6.519	12.679	0.625	3.808	4.74 \pm 0.16
CosyVoice 2*	-	-	-	-	0.5B	6.812	17.956	0.583	3.447	3.49 \pm 0.37
F5-TTS	\times	Full	Full	Full	336M (100%)	17.345	45.883	0.641	3.152	2.64 \pm 0.41
F5-TTS	\checkmark	Full	Full	Full	336M (100%)	5.671	12.032	0.659	3.503	3.58 \pm 0.32
PEFT-TTS	\checkmark	Full	P-Adpt w. DP (0.3)	LoRA (16)	5.81M (1.72%)	4.906	12.625	0.640	3.338	3.88 \pm 0.16
Ablation (Separated Enc.)	\checkmark	Full	P-Adpt w. DP (0.3)	LoRA (32)	7.25M (2.15%)	4.570	11.338	0.645	3.264	-
	\checkmark	Full	P-Adpt w. DP (0.3)	LoRA (64)	10.1M (3.01%)	4.916	12.454	0.646	3.168	-
	\checkmark	Full	P-Adpt	LoRA (16)	5.81M (1.72%)	4.282	10.804	0.638	3.241	-
	\checkmark	Full	Freeze	LoRA (16)	5.70M (1.69%)	5.088	12.888	0.637	3.235	-
Ablation (C-Adpt)	\checkmark	C-Adpt (0.25)	P-Adpt w. DP (0.3)	LoRA (16)	5.81M (1.72%)	5.029	12.742	0.643	3.225	-
	\checkmark	C-Adpt (0.25)	P-Adpt w. DP (0.5)	LoRA (16)	5.81M (1.72%)	5.513	13.923	0.642	3.185	-
	\checkmark	C-Adpt (0.25)	Freeze	LoRA (16)	5.70M (1.69%)	5.288	13.143	0.649	3.236	-
	\checkmark	C-Adpt (0.25)	P-Adpt	LoRA (16)	5.81M (1.72%)	4.940	12.383	0.633	3.274	-
	\checkmark	C-Adpt (1)	P-Adpt	LoRA (16)	2.57M (0.76%)	6.099	15.293	0.636	3.213	-
	\checkmark	C-Adpt (4)	P-Adpt	LoRA (16)	1.78M (0.52%)	7.242	18.192	0.639	3.289	-

Table 2: *Multi-Speaker Korean TTS Results. Note that F5-TTS and PEFT-TTS are fine-tuned with a single-speaker Korean dataset.*

Methods	Pre-train	Text Enc. (γ)	Prompt Emb.	DiT (rank)	#Trainable Params	CER (\downarrow)	WER (\downarrow)	SIM-O (\uparrow)	UTMOS (\uparrow)	NMOS (\uparrow)	SMOS (\uparrow)
GT	-	-	-	-	-	3.488	10.295	0.834	3.604	4.45 \pm 0.24	4.88 \pm 0.06
CosyVoice 2*	-	-	-	-	0.5B	3.662	10.887	0.806	3.756	4.26 \pm 0.28	4.36 \pm 0.09
F5-TTS	\times	Full	Full	Full	336M (100%)	36.482	107.541	0.250	2.545	1.13 \pm 0.16	1.12 \pm 0.08
F5-TTS	\checkmark	Full	Full	Full	336M (100%)	13.454	39.413	0.627	3.446	2.24 \pm 0.31	2.35 \pm 0.11
PEFT-TTS	\checkmark	Full	P-Adpt w. DP (0.3)	LoRA (16)	5.81M (1.72%)	7.220	21.591	0.641	3.435	3.06 \pm 0.31	2.81 \pm 0.21
Ablation (Separated Enc.)	\checkmark	Full	P-Adpt w. DP (0.3)	LoRA (32)	7.25M (2.15%)	5.358	15.547	0.614	3.310	-	-
	\checkmark	Full	P-Adpt w. DP (0.3)	LoRA (64)	10.1M (3.01%)	6.140	17.927	0.605	3.322	-	-
	\checkmark	Full	P-Adpt	LoRA (16)	5.81M (1.72%)	6.662	20.021	0.639	3.251	-	-
	\checkmark	Full	Freeze	LoRA (16)	5.70M (1.69%)	6.705	19.500	0.618	3.367	-	-
Ablation (C-Adpt)	\checkmark	C-Adpt (0.25)	P-Adpt w. DP(0.3)	LoRA (16)	5.81M (1.72%)	8.540	25.000	0.655	3.325	-	-
	\checkmark	C-Adpt (0.25)	P-Adpt w. DP(0.5)	LoRA (16)	5.81M (1.72%)	8.904	26.144	0.631	3.316	-	-
	\checkmark	C-Adpt (0.25)	Freeze	LoRA (16)	5.70M (1.69%)	10.630	31.562	0.636	3.277	-	-
	\checkmark	C-Adpt (0.25)	P-Adpt	LoRA (16)	5.81M (1.72%)	11.134	35.493	0.657	3.350	-	-
	\checkmark	C-Adpt (1)	P-Adpt	LoRA (16)	2.57M (0.76%)	15.376	44.319	0.642	3.401	-	-
	\checkmark	C-Adpt (4)	P-Adpt	LoRA (16)	1.78M (0.52%)	18.192	52.139	0.656	3.406	-	-

4.3. Multi-speaker TTS Results

We also evaluated multi-speaker TTS performance with the same baselines. Note that although we use only a single-speaker Korean dataset, our proposed PEFT methods enable the generation of novel speakers while preserving the zero-shot TTS capabilities of the pre-trained models. Table 2 shows that fine-tuning all model parameters results in a loss of zero-shot TTS ability. Furthermore, the results also show higher CER and WER due to hallucinations in the generated speech, such as word repeating and skipping. The results show that our proposed methods could synthesize the speech of unseen speakers, achieving higher similarity compared to the baselines. However, we found that it is crucial to fine-tune all parameters of the text encoder for robust text-to-speech performance in terms of CER and WER. Still, it has much lower performance than the model trained with large-scale multilingual dataset including Korean.

4.4. Ablation Study

4.4.1. Conditioning Adapter

We adopt a Conv-Adapter to generate speech from the text sequences of new language. In our preliminary study, we could fine-tune the model with a large γ of 4 for English dataset. However, it is difficult to learn a new language with a large γ . The results show that learning new language requires more trainable parameters. In this regard, we fine-tuned all parameters of the pre-trained text encoder. Note that the original text encoder could be used for the pre-trained languages.

4.4.2. Prompt Adapter

We found that freezing the prompt embedding could increase the audio quality. However, it shows lower speaker similarity such as intonation of new language, so we fine-tune it with

LoRA adapter. Additionally, we found that training it with DropPath could improve the similarity and preserve the audio quality of the pre-trained model.

4.4.3. DiT LoRA Adapter

We compared the rank of LoRA from 16 to 64. Although increasing the rank size could improve the pronunciation of new language, it loses its speaker adaptation ability and decreases the audio quality. It indicates that it is important to use low rank when using a small-scale fine-tuning dataset for preserving the generative capability of the pre-trained model. We believe that scaling up the training dataset simply improves performance with a small rank size. Additionally, using a rank of 32 could improve the pronunciation. However, we use a rank of 16 to preserve the zero-shot capability in this work.

5. Conclusion

We propose PEFT-TTS, PEFT methods for cross-lingual continual learning in low-resource environments. We investigated three adapters for each module: a Conditioning Adapter for text conditioning, a Prompt Adapter for zero-shot TTS, and a DiT LoRA adapter for acoustic generation. We successfully fine-tuned F5-TTS for a new language using only a single GPU and a single-speaker dataset. Furthermore, our proposed method preserves the zero-shot TTS capability despite training with a single-speaker. However, we observed that the generated samples of zero-shot TTS might contain hallucinated results. While fine-tuning the entire text encoder mitigates this phenomenon, it also requires more trainable parameters. In the future, we will explore more robust and efficient fine-tuning methods for text conditioning layers and develop improved alignment methods between text and speech to reduce hallucinations.

6. Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2025-02283048, Developing the Next-Generation General AI with Reliability, Ethics, and Adaptability, IITP-2025-RS-2023-00255968, the Artificial Intelligence Convergence Innovation Human Resources Development, No.RS-2021-II212068, Artificial Intelligence Innovation Hub, 2022-0-01077, the National Program for Excellence in SW), and Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea)&Gwangju Metropolitan City.

7. References

- [1] 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,” *arXiv preprint arXiv:1703.10135*, 2017.
- [2] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” *arXiv preprint arXiv:2006.04558*, 2020.
- [3] 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.
- [4] 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.
- [5] S.-H. Lee, S.-B. Kim, J.-H. Lee, E. Song, M.-J. Hwang, and S.-W. Lee, “Hierspeech: Bridging the gap between text and speech by hierarchical variational inference using self-supervised representations for speech synthesis,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 624–16 636, 2022.
- [6] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar *et al.*, “Voicebox: Text-guided multilingual universal speech generation at scale,” *Advances in neural information processing systems*, vol. 36, 2024.
- [7] S. Kim, K. Shih, J. F. Santos, E. Bakhturina, M. Desta, R. Valle, S. Yoon, B. Catanzaro *et al.*, “P-flow: a fast and data-efficient zero-shot tts through speech prompting,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [8] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, “F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching,” *arXiv preprint arXiv:2410.06885*, 2024.
- [9] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, “Deep networks with stochastic depth,” in *European Conference on Computer Vision*. Springer, 2016, pp. 646–661.
- [10] 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. [Online]. Available: <https://openreview.net/forum?id=nZeVKeeFYf9>
- [11] 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*, 2024, pp. 4413–4417.
- [12] N. Hounsby, A. Giurghi, S. Jastrzebski, B. Morrone, Q. De Larousilhe, A. Gesmundo, M. Attariyan, and S. Gelly, “Parameter-efficient transfer learning for nlp,” in *International conference on machine learning*. PMLR, 2019, pp. 2790–2799.
- [13] B. Thomas, S. Kessler, and S. Karout, “Efficient adapter transfer of self-supervised speech models for automatic speech recognition,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7102–7106.
- [14] X. Yue, X. Gao, X. Qian, and H. Li, “Adapting pre-trained self-supervised learning model for speech recognition with light-weight adapters,” *Electronics*, vol. 13, no. 1, p. 190, 2024.
- [15] S. Udupa, J. Bandekar, S. Kumar, S. Murthy, P. Pai, S. Raghavan, R. Nanavati, P. K. Ghosh *et al.*, “Adapter pre-training for improved speech recognition in unseen domains using low resource adapter tuning of self-supervised models,” in *Proc. Interspeech 2024*, 2024, pp. 2529–2533.
- [16] N. Inoue, S. Otake, T. Hirose, M. Ohi, and R. Kawakami, “Elp-adapters: Parameter efficient adapter tuning for various speech processing tasks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [17] N. Morioka, H. Zen, N. Chen, Y. Zhang, and Y. Ding, “Residual adapters for few-shot text-to-speech speaker adaptation,” *arXiv preprint arXiv:2210.15868*, 2022.
- [18] A. Mehrish, A. R. Kashyap, L. Yingting, N. Majumder, and S. Poria, “Adaptmix: Exploring the efficacy of mixture of adapters for low-resource tts adaptation,” *arXiv preprint arXiv:2305.18028*, 2023.
- [19] Y. Li, A. Mehrish, B. Chew, B. Cheng, and S. Poria, “Leveraging parameter-efficient transfer learning for multi-lingual text-to-speech adaptation,” *arXiv preprint arXiv:2406.17257*, 2024.
- [20] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, “Flow matching for generative modeling,” *arXiv preprint arXiv:2210.02747*, 2022.
- [21] W. Peebles and S. Xie, “Scalable diffusion models with transformers,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.
- [22] S. E. Eskimez, X. Wang, M. Thakker, C. Li, C.-H. Tsai, Z. Xiao, H. Yang, Z. Zhu, M. Tang, X. Tan *et al.*, “E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts,” in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 682–689.
- [23] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, “Convnext v2: Co-designing and scaling convnets with masked autoencoders,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 133–16 142.
- [24] H. Chen, R. Tao, H. Zhang, Y. Wang, X. Li, W. Ye, J. Wang, G. Hu, and M. Savvides, “Conv-adapter: Exploring parameter efficient transfer learning for convnets,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024, pp. 1551–1561.
- [25] K. Park, “Korean single speaker speech dataset,” 2018, accessed: 2024-02-17. [Online]. Available: <https://github.com/Kyubyong/ks>
- [26] H. He, Z. Shang, C. Wang, X. Li, Y. Gu, H. Hua, L. Liu, C. Yang, J. Li, P. Shi, Y. Wang, K. Chen, P. Zhang, and Z. Wu, “Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation,” in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 885–890.
- [27] AI Hub, “Korean Multi-Speaker Speech Synthesis Dataset,” 2024, accessed: 2024-12-12. [Online]. Available: <https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=542>
- [28] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.
- [29] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [30] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, “Utmos: Utokyo-sarulab system for voicemos challenge 2022,” *Interspeech 2022*, 2022.
- [31] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang *et al.*, “Cosyvoice 2: Scalable streaming speech synthesis with large language models,” *arXiv preprint arXiv:2412.10117*, 2024.