Leveraging Parameter-Efficient Transfer Learning for Multi-Lingual Text-to-Speech Adaptation

Anonymous ACL submission

Abstract

Different languages have distinct phonetic systems and vary in their prosodic features making it challenging to develop a Text-to-Speech (TTS) model that can effectively synthesise speech in multilingual settings. Furthermore, TTS architecture needs to be both efficient enough to capture nuances in multiple languages and efficient enough to be practical for deployment. The standard approach is to build transformer based model such as SpeechT5 and train it on large multilingual dataset. As the size of these models grow the conventional fine-tuning for adapting these model becomes impractical due to heavy computational cost. In this paper, we proposes to integrate parameter-efficient transfer learning (PETL) methods such as adapters and hypernetwork with TTS architecture for multilingual speech synthesis. Notably, in our experiments PETL methods able to achieve comparable or even better performance compared to full fine-tuning with only ∼2.5% tunable parameters[1].

1 Introduction

Multilingual speech synthesis, generating speech in multiple languages from text input, represents a major advancement in speech processing with wide-reaching implications for global communication (Tan et al., 2021; Mehrish et al., 2023b). Unlike single-language systems, multilingual architectures break linguistic barriers, transforming education, entertainment, healthcare, and customer service by facilitating seamless communication across languages (Marais et al., 2020; Seong et al., 2021; Le et al., 2024; Panda et al., 2020).

Current multilingual TTS architectures face challenges (Nuthakki et al., 2023; Kaur and Singh, 2023), including the complexity of modeling diverse linguistic structures, phonetic variations, and prosodic features across languages. Resource constraints, such as the availability of multilingual corpora and linguistic expertise, can impede model development, particularly for low-resource languages or underrepresented dialects (Tan et al., 2021; Mehrish et al., 2023b). Addressing these challenges requires concerted efforts in data collection, model development, and evaluation.

The advancement of architectural designs, coupled with pre-training models such as SpeechT5 (Ao et al., 2021), reflects challenges similar to those encountered in NLP. Achieving optimal performance through fine-tuning these models for diverse downstream tasks or domain adaptations requires substantial task-specific datasets. Moreover, fine-tuning all model parameters necessitates significant memory resources allocated to each task. With limited data available for various underrepresented languages, full fine-tuning can further leads to poor generalization. Researchers have sought solutions to these challenges through the exploration of PETL methods (Li et al., 2023; Oh et al., 2023; Chen et al., 2023; Sathyendra et al., 2022; Vander-reydt et al., 2023; Le et al., 2021). However, their investigation remains limited for TTS adaptation.

In this paper, we extends PETL approaches to the multilingual TTS, focusing on adapter (Houlsby et al., 2019) and Hyper-Network (Üstün et al., 2022). We pioneer the hyper-networks for multilingual TTS adaptation and introduces the *Multi-Conditioned HyperGenerator* for multilingual TTS. Our major contributions includes: (1) Regular & Dynamic Adapters: We embed language-specific parameters into SpeechT5 using regular adapters and explore a hyper-network to generate these parameters, referred to as HyperGenerator. (2) Parameter Efficiency: We achieve comparable or superior performance to full fine-tuning using only about 2.44% of the parameters. (3) Improved Zero-shot Performance: HyperGenerator outperforms full finetuning and regular adapters on an unseen

[1] The code and samples are available at: https://anonymous.4open.science/r/multilingualTTS-BA4C

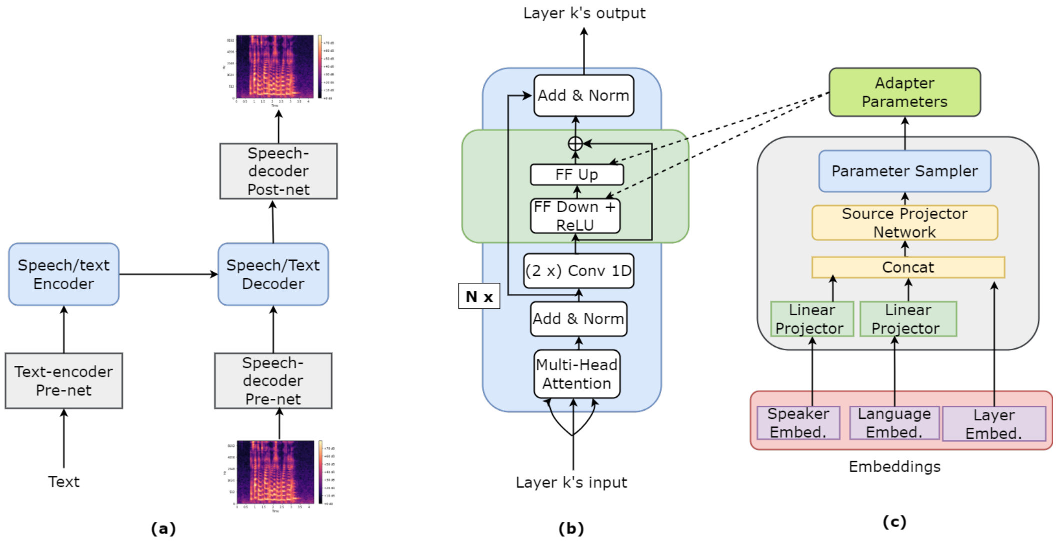


Figure 1: (a) SpeechT5 TTS architecture. (b) Encoder/decoder architecture with adapter. (c) HyperGenrator architecture.

language with the same parameter count.

2 Related work

Research on PETL methods for large multilingual pretrained models like XLSR (Vanderreydt et al., 2023) has gained significant attention, notably in Automatic Speech Recognition (ASR) (Fu et al., 2023; Zhang et al., 2023; Yu et al., 2023; Yang et al., 2023; Shi and Kawahara, 2024). Li et al. (Li et al., 2023) propose a benchmark utilizing XLSR-53 (Conneau et al., 2020), employing PETL such as regular adapters (Pfeiffer et al., 2020), prefix tuning (Li and Liang, 2021), and LoRA (Hu et al., 2021). Le et al. (Le et al., 2021) and Zhao et al. (Zhao et al., 2022) explore multilingual neural machine translation, focusing on lightweight adapter tuning. Morioka et al. (Morioka et al., 2022) advocate for integrating regular adapters with TTS models for few-shot speaker adaptation, while Mehrish et al. (Mehrish et al., 2023a) introduce a mixture of experts for low-resource speaker adaptation.

3 Methodology

3.1 Base Model Architecture: SpeechT5

SpeechT5 (Ao et al., 2021) merges NLP and speech synthesis techniques, extending the transformer-based T5 architecture (Raffel et al., 2020). It integrates self-attention mechanisms and CNNs to capture both temporal dependencies and spectral features in speech. By pre-training on large-scale speech corpora and fine-tuning on specific datasets, SpeechT5 excels in tasks such as speech recognition, TTS, and speech translation.

3.2 Adapter

In this work, we integrate language-specific parameters using adapter modules, commonly employed in the NLP for multilingual or multi-task scenarios (Ansell et al., 2021). Following the formulation of (Houlsby et al., 2019), we insert one adapter block after each convolutional block of every transformer module in the SpeechT5 model as shown in Figure 1. Each adapter module, with fewer parameters compared to the main network (SpeechT5), down-projects the input to a lower-dimensional space, applies a non-linearity, and then up-projects back to the original dimensions. A residual connection is added to produce the final output. During language adaptation, only the adapter parameters are updated while keeping the main network frozen.

3.3 HyperGenerator

HyperGenerator consists of a hyper network (Üstün et al., 2022) that generates the weights of all adapter modules. As depicted in Figure 1, a single hyper-network is employed to create adapters for multiple languages and layers, with conditioning on (s, l, p), where s denotes speaker embeddings, l represents the target language, and p indicates the encoder or decoder layer ID. This method, unlike traditional adapters, promotes cross-language and cross-layer information sharing, enabling the hyper-network to efficiently distribute its capacity among them. By adapting parameters based on speaker characteristics and language specifics, the hyper-network augments the effectiveness of adapters. Further-

Model	de		fr		fi		hu		nl		avg		**Params**
	MCD	CER	MCD	CER	MCD	CER	MCD	CER	MCD	CER	MCD	CER	
Finetune Multilingual w/o Pretrain	4.88	7.19	4.91	14.40	4.76	15.16	4.93	16.29	5.46	10.56	4.99	12.72	144M(100%)
Finetune Multilingual w/ Pretrain	4.87	6.66	4.95	11.82	4.82	12.64	5.00	15.52	5.51	10.28	5.03	11.38	144M(100%)
Adapter Multilingual w/ Pretrain	4.75	**6.30**	4.93	11.81	4.75	9.58	4.92	16.11	5.40	10.37	4.95	10.83	3.56M(2.47%)
HyperAdapter Multilingual w/ Pretrain	4.78	6.52	5.04	15.50	4.67	**7.05**	**4.87**	**13.13**	**5.36**	10.96	4.94	**10.63**	3.52M(2.44%)
Finetune Monolingual w/ Pretrain	4.86	6.44	4.85	**10.80**	4.67	15.09	4.90	15.54	5.53	**8.47**	4.96	11.27	144M(100%)
Adapter Monolingual w/ Pretrain	4.77	6.82	4.82	14.32	**4.63**	9.41	4.94	14.53	5.36	14.22	**4.90**	11.86	3.56M(2.47%)
HyperAdapter Monolingual w/ Pretrain	**4.71**	7.94	**4.82**	14.66	4.77	13.47	4.93	14.00	5.40	13.21	4.93	12.66	3.52M(2.44%)

Table 1: Evaluation results for *seen* languages along with the percentage of parameters updated during training.

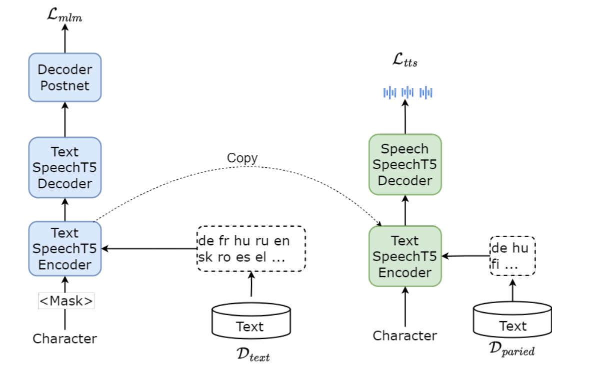


Figure 2: Multilingual Masked Text Pretraining. Where $\mathcal{L}_{mlm}$ and $\mathcal{L}_{tts}$ is mask language modeling and reconstruction loss respectively.

more, to ensure network efficiency, we utilize a shared hyper-network to generate adapter parameters across all layers within the TTS backbone, further conditioning it with the layer ID p.

4 Experimental Setup

4.1 Baseline and Dataset

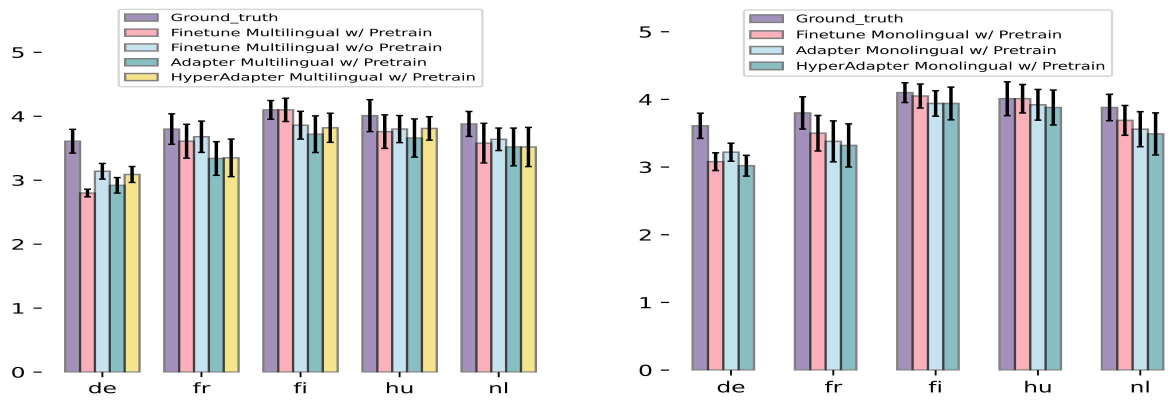
We developed a baselines with the following 3 configurations for comparing the performance of the adapter and HyperGenerator with full fine-tuning.
Monolingual: We finetune the SpeechT5 individually for each language, uniquely optimizing its parameters to enhance speech synthesis performance.
Multilingual: We finetune the SpeechT5 with diverse speech data from multiple languages. This improves the model's ability to understand and generate speech across various linguistic contexts, capturing cross-lingual patterns, phonetic variations, and language-specific features.
Multilingual Masked Text Pretraining: Multilingual models like multilingual BERT (Devlin et al., 2018) have demonstrated strong cross-lingual transfer capabilities in NLP tasks. Leveraging multilingual pre-training improves generalization to other languages without specific target data. In this settings, we extend MLM pre-training to SpeechT5 to enhance pronunciation and prosody transfer. The left side of Figure 2 illustrates the unsupervised pre-training of SpeechT5's text encoder and decoder using text-only data $\mathcal{D}_{text}$ with MLM. The pre-trained text encoder is then integrated into the TTS pipeline, as shown on the right side of Figure 2, and trained on paired speech-text data $\mathcal{D}_{paired}$.

For fine-tuning using monolingual and multilingual configuration, as discussed in Section 4.1, we leverage German (de), French (fr), Finnish (fi), Hungarian (hu), and Dutch (nl)—as the five *seen* European languages from the CSS10 dataset (Park and Mulc, 2019). To evaluate zero-shot performance, we use Spanish (es) as an *unseen* language. For Multilingual Masked Text Pretraining, we utilize transcripts from VoxPopuli (Wang et al., 2021), M-AILABS (Bakhturina et al., 2021), and CSS10 (Park and Mulc, 2019) to pre-train the SpeechT5 text encoder-decoder for a character-based masked language modeling task.

4.2 Training and Evaluation

We follow the data partition outlined in (Saeki et al., 2023). We use pretrained chekcpoint[2] of SpeechT5 for all experiments. Speaker embeddings[3] are set at a dimension of 256, while language embeddings are initialized using pretrained weights from lang2vec (Littell et al., 2017). The layer embedding dimension is set at 64. The bottleneck dimension for adapters is 128, whereas for HyperGenerator, is 32 for ensuring the same number of parameters across both architectures. We employed MCD (Kominek et al., 2008) and assess intelligibility using Character Error Rates (CERs) computed with the multilingual ASR (Radford et al., 2023)[4] as objective metrics. Furthermore, to evaluate naturalness, we conducted listening tests to calculate the MOS of synthesized speech. We recruited five native speaker via Amazon Mechanical Turk (AMT)

[2]https://huggingface.co/microsoft/speecht5_tts
[3]Pretrained speaker verification model (Wan et al., 2018).
[4]https://github.com/openai/whisper



(a) Multilingual performance

(b) Monolingual performance

Figure 3: Subjective evaluation on naturalness: MOS score

for each of the languages.

Model	es	
	MCD	CER
Finetune Multilingual w/o Pretrain	5.75	39.32
Finetune Multilingual w/ Pretrain	5.87	34.80
Adapter Multilingual w/ Pretrain	5.39	45.94
HyperAdapter Multilingual w/ Pretrain	**5.28**	**18.79**

Table 2: Evaluation results for *unseen* language (es).

5 Results

5.1 Objective and Subjective Evaluation

Table 1 shows that Finetune *Multilingual with Pretraining* outperforms Finetune *Multilingual without Pretraining* due to multilingual masked text pretraining. Both Adapter and HyperGenerator achieve similar or better performance than full fine-tuning with significantly fewer parameters. Full fine-tuning updates 144M parameters, while Adapter and HyperGenerator use only 3.56M and 3.52M parameters, respectively, with HyperGenerator showing superior performance in multilingual settings with text pretraining.

While the performance gain for HyperGenerator with *seen* languages is modest, it shows promise for zero-shot multilingual speech synthesis. Table 2 shows that HyperGenerator achieved a CER of 18.79% for Spanish, significantly lower than the over 30% CER for Multilingual Fine-tuning and adapter-based approaches. This highlights HyperGenerator's dynamic adaptability and potential for efficient and accurate zero-shot synthesis. Figure 4 further demonstrates this, as speech from the same language clusters together, indicating HyperGenerator's ability to adjust parameters based on language, unlike static adapters.

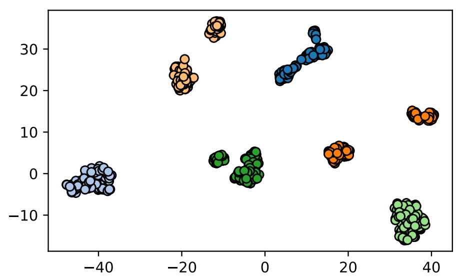


Figure 4: t-SNE plot of HyperGenerator parameters for 6 languages from the CSS10 test set, with same colors denoting speech samples from the same language.

MOS results (Figure 3a and Figure 3b) indicate that Adapters and HyperGenerator perform as well as or better than full fine-tuning in multilingual contexts. HyperGenerator consistently achieved the highest scores, with values of 3.09 for *de* and 3.81 for *hu*, demonstrating superior naturalness. Similar trends in monolingual scenarios highlight HyperGenerator's effectiveness in generating high-quality speech across different languages.

6 Conclusion

In this paper, we advances multilingual speech synthesis using PETL methods like adapter fine-tuning, achieving SOTA performance with fewer parameters. Introducing regular and dynamic adapters with a hyper-network enhances efficiency and zero-shot performance. Future work could optimize adapters for specific languages, improve cross-lingual transfer learning, and reduce model complexity while maintaining high performance

Limitations

While our proposed approach shows promise in advancing multilingual TTS synthesis, there are several limitations that must be acknowledged. Addressing these challenges will be crucial for enhancing the robustness and applicability of our methods across a wider range of languages and use cases. The key limitations are as follows:

- The performance of hypernetworks and adapters can vary greatly depending on the hyperparameters used. Adjusting these settings for each language and task is often a complex and time-consuming process that requires significant computational resources.
- Languages such as Russian and Greek use scripts that differ from the Latin alphabet, like Cyrillic and Greek scripts, respectively. These scripts have unique rules for how letters and sounds are represented. The current PETL methods might not fully address these differences, resulting in lower quality speech synthesis for these languages.
- Symbolic languages, such as Chinese and Japanese, have unique linguistic elements like Chinese logograms and Japanese kana, as well as complex grammatical structures in languages like Russian and Greek. The proposed architecture in its current form can struggle to handle these diverse features effectively, which means modification to these adaptation techniques are needed to improve performance.

Potential Risk

While our research aims to advance multilingual TTS technology, it is crucial to acknowledge the potential risks associated with such systems. We will discuss some associated risks as follows :

- Malicious Use and Disinformation: The ability of TTS systems to generate highly realistic speech could be used to create disinformation. This could lead to the spread of false information, manipulation of opinion, and erosion of trust in digital content.
- Our research utilizes the publicly available CSS10 dataset, however utilization of personalized data to adapt these models can have the risk of privacy violations. Therefore it is important to follow best data management practices that do not inadvertently compromise privacy.
- TTS systems are vulnerable to adversarial attacks where small perturbations to the input can lead to significant changes in the output. Although the proposed framework is robust to noise, the necessary security measures and continuous testing of the system against potential attacks can enhance resilience.

Ethical Considerations

The TTS system could be used to produce misleading or harmful content. For instance, synthesized speech could be exploited to create fake audio recordings that mimic real individuals, potentially leading to misinformation or fraud. Additionally, the accessibility of TTS technology might raise concerns about the unauthorized use of voices, infringing on personal privacy and intellectual property rights.

References

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. 2021. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. *arXiv preprint arXiv:2110.07205*.

Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang. 2021. Hi-fi multi-speaker english tts dataset. *arXiv preprint arXiv:2104.01497*.

Nanxin Chen, Izhak Shafran, Yu Zhang, Chung-Cheng Chiu, Hagen Soltau, James Qin, and Yonghui Wu. 2023. Efficient adapters for giant speech models. *arXiv preprint arXiv:2306.08131*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier. 2023. On the effectiveness of parameter-efficient fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12799–12807.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *ICLR*.

Navdeep Kaur and Parminder Singh. 2023. Conventional and contemporary approaches used in text to speech synthesis: A review. *Artificial Intelligence Review*, 56(7):5837–5880.

John Kominek, Tanja Schultz, and Alan W Black. 2008. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *SLTU*, pages 63–68.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. *arXiv preprint arXiv:2106.01463*.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. 2024. Voicebox: Text-guided multilingual universal speech generation at scale. *NeurIPS*, 36.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Yingting Li, Ambuj Mehrish, Rishabh Bhardwaj, Navonil Majumder, Bo Cheng, Shuai Zhao, Amir Zadeh, Rada Mihalcea, and Soujanya Poria. 2023. Evaluating parameter-efficient transfer learning approaches on sure benchmark for speech understanding. In *ICASSP 2023*, pages 1–5. IEEE.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *EACL*, volume 2, pages 8–14.

Laurette Marais, Johannes A Louw, Jaco Badenhorst, Karen Calteaux, Ilana Wilken, Nina Van Niekerk, and Glenn Stein. 2020. Awezamed: A multilingual, multimodal speech-to-speech translation application for maternal health care. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE.

Ambuj Mehrish, Abhinav Ramesh Kashyap, Li Yingting, Navonil Majumder, and Soujanya Poria. 2023a. Adaptermix: Exploring the efficacy of mixture of adapters for low-resource tts adaptation. *arXiv preprint arXiv:2305.18028*.

Ambuj Mehrish, Navonil Majumder, Rishabh Bharadwaj, Rada Mihalcea, and Soujanya Poria. 2023b. A review of deep learning techniques for speech processing. *Information Fusion*, page 101869.

Nobuyuki Morioka, Heiga Zen, Nanxin Chen, Yu Zhang, and Yifan Ding. 2022. Residual adapters for few-shot text-to-speech speaker adaptation. *arXiv preprint arXiv:2210.15868*.

Praveena Nuthakki, Madhavi Katamaneni, Chandra Sekhar JN, Kumari Gubbala, Bullarao Domathoti, Venkata Rao Maddumala, and Kumar Raja Jetti. 2023. Deep learning based multilingual speech synthesis using multi feature fusion methods. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Changdae Oh, Hyeji Hwang, Hee-young Lee, YongTaek Lim, Geunyoung Jung, Jiyoung Jung, Hosik Choi, and Kyungwoo Song. 2023. Blackvip: Black-box visual prompting for robust transfer learning. In *CVPR*, pages 24224–24235.

Soumya Priyadarsini Panda, Ajit Kumar Nayak, and Satyananda Champati Rai. 2020. A survey on speech synthesis techniques in indian languages. *Multimedia Systems*, 26:453–478.

Kyubyong Park and Thomas Mulc. 2019. Css10: A collection of single speaker speech datasets for 10 languages. *arXiv preprint arXiv:1903.11269*.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. *arXiv preprint arXiv:2007.07779*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*, pages 28492–28518. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Takaaki Saeki, Soumi Maiti, Xinjian Li, Shinji Watanabe, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2023. Learning to speak from text: Zero-shot multilingual text-to-speech with unsupervised text pretraining. In *International Joint Conference on Artificial Intelligence*.

Kanthashree Mysore Sathyendra, Thejaswi Muniyappa, Feng-Ju Chang, Jing Liu, Jinru Su, Grant P. Strimel, Athanasios Mouchtaris, and Siegfried Kunzmann. 2022. Contextual adapters for personalized speech recognition in neural transducers. In *ICASSP 2022)*, pages 8537–8541.

Jiwon Seong, WooKey Lee, and Suan Lee. 2021. Multilingual speech synthesis for voice cloning. In *2021 IEEE International Conference on Big Data and Smart Computing*, pages 313–316. IEEE.

Hao Shi and Tatsuya Kawahara. 2024. Exploration of adapter for noise robust automatic speech recognition. *arXiv preprint arXiv:2402.18275*.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, Gertjan van Noord, and Sebastian Ruder. 2022. Hyper-x: A unified hypernetwork for multi-task multilingual transfer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7934–7949.

Geoffroy Vanderreydt, Amrutha Prasad, Driss Khalil, Srikanth Madikeri, Kris Demuynck, and Petr Motlicek. 2023. Parameter-efficient tuning with adaptive bottlenecks for automatic speech recognition. In *ASRU-2023*, pages 1–7. IEEE.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized end-to-end loss for speaker verification. In *ICASSP-2018*, pages 4879–4883. IEEE.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

Chao-Han Huck Yang, Bo Li, Yu Zhang, Nanxin Chen, Rohit Prabhavalkar, Tara N Sainath, and Trevor Strohman. 2023. From english to more languages: Parameter-efficient model reprogramming for cross-lingual speech recognition. In *ICASSP*, pages 1–5. IEEE.

Yu Yu, Chao-Han Huck Yang, Jari Kolehmainen, Prashanth G Shivakumar, Yile Gu, Sungho Ryu Roger Ren, Qi Luo, Aditya Gourav, I-Fan Chen, Yi-Chieh Liu, et al. 2023. Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition. In *ASRU-2023*, pages 1–8. IEEE.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.

Jinming Zhao, Hao Yang, Ehsan Shareghi, and Gholamreza Haffari. 2022. M-adapter: Modality adaptation for end-to-end speech-to-text translation. *arXiv preprint arXiv:2207.00952*.