Interspeech 2024
1-5 September 2024, Kos, Greece

Learning Fine-Grained Controllability on Speech Generation via Efficient Fine-Tuning

Chung-Ming Chien[†,1], *Andros Tjandra*[2], *Apoorv Vyas*[2], *Matt Le*[2], *Bowen Shi*[2], *Wei-Ning Hsu*[2]
[1]Toyota Technological Institute at Chicago, USA [2]AI at Meta, USA
cmchien@ttic.edu, androstj@meta.com

Abstract

As the scale of generative models continues to grow, efficient reuse and adaptation of pre-trained models have become crucial considerations. In this work, we propose Voicebox Adapter, a novel approach that integrates fine-grained conditions into a pre-trained Voicebox speech generation model using a cross-attention module. To ensure a smooth integration of newly added modules with pre-trained ones, we explore various efficient fine-tuning approaches. Our experiment shows that the LoRA with bias-tuning configuration yields the best performance, enhancing controllability without compromising speech quality. Across three fine-grained conditional generation tasks, we demonstrate the effectiveness and resource efficiency of Voicebox Adapter. Follow-up experiments further highlight the robustness of Voicebox Adapter across diverse data setups.

Index Terms: Speech generation, fine-grained conditioning, efficient fine-tuning

1. Introduction

Large-scale speech pre-training has demonstrated remarkable competence across various applications [1, 2]. In contrast to discriminative pre-training [3, 4], which aims to acquire speech representations beneficial for downstream tasks, generative pre-training [5] directly learns the data distribution from speech corpora [6, 7]. Through pre-training on diverse data, speech generation models learn the distribution of speech of different styles, speakers, and various transient events, including pauses, stress, and non-verbal vocalizations such as laughter [8].

Recent speech generation models have demonstrated one-shot capabilities, allowing control over the speaker and style of generated speech with a short speech prompt [9, 10]. We term this form of controlled generation as *global control*, where the speaker and style remain consistent throughout the entire generated utterance. In contrast, *fine-grained control* involves adding conditioning only to specific parts of the generated utterance, such as emphasizing certain words or introducing pauses at specific points. Despite the intrinsic global controllability observed in many generative pre-training methods [9, 10], the exploration of post-hoc integration of fine-grained controllability into pre-trained speech generation models remains limited.

In this work, we propose Voicebox Adapter, a method to integrate fine-grained controllability into Voicebox [10], a text-conditioned speech generation model. We explore three fine-grained conditions: punctuation, emphasis, and laughter. We hypothesize that these fine-grained vocalizations are inherently learned during pre-training, but the absence of fine-grained conditioning mechanisms in Voicebox restricts its ability to generate speech with precise fine-grained vocalizations. To address this limitation, we introduce cross-attention modules to the Transformer layers of the pre-trained Voicebox to extract and integrate fine-grained condition information. Additionally, we employ parameter-efficient adaptation methods to seamlessly connect the pre-trained parameters with the new modules. Experimental results demonstrate that Voicebox Adapter achieves performance comparable to fine-tuning the entire model, with adapter parameters comprising a small percentage of the model.

Our contributions are as follows: (1) we propose Voicebox Adapter, which augments Voicebox, a pre-trained speech generation model, with fine-grained controllability; (2) we explore different efficient fine-tuning methods to bridge the gap between pre-trained parameters and new fine-grained conditioning modules; (3) we show that Voicebox Adapter can generalize across various fine-grained conditions, attaining performance comparable to that achieved by fine-tuning the entire model with significantly fewer fine-tuned parameters; (4) we conduct experiments using varying amounts of fine-tuning data and different hidden dimension sizes, analyzing the performance of Voicebox Adapter under different setups.

2. Related works

2.1. Adaptive fine-grained controllable speech generation

Adaptation has become an important research topic for speech generation. A widely adopted approach involves the initial pre-training of a speech generation model on a large and diverse dataset, followed by fine-tuning a specific subset of parameters with a smaller target dataset [11]. Adaptive methods have shown great success in global speaker [11, 12] and style control [13], and can also provide fine-grained controllability [14]. In this work, our goal is to develop a comprehensive adaptive framework that transcends task-specific design considerations [13] and can be applied to various fine-grained conditions.

2.2. Efficient fine-tuning for Transformers

The growing size of language models has made efficient fine-tuning of Transformer models an increasingly important research topic [15]. Extensive studies have been made to explore the use of adapters [15] and Low-Rank Adaptation (LoRA) [16] across diverse tasks. Instead of limiting fine-tuning only to newly added modules, recent research also advocates for unlocking specific pre-trained parameters, such as the normalization, bias, and scale of linear layers [17]. In this paper, we systematically investigate various efficient fine-tuning strategies applied to the Transformer layers of the Voicebox model.

3. Background

Voicebox [10] is a speech generation framework consisting of a duration model and an acoustic model, both with a Transformer architecture. Given a phoneme sequence as input, the duration model is trained to predict the duration of each phoneme using an L_1 regression loss. The flow-matching-based [18] acoustic model defines a vector field to transform a Gaussian prior $p(x)$ into the real distribution of Mel spectrograms $q(x)$.

Let z_p be a time-aligned phoneme embedding sequence,[1]

[†]Work done during an internship at Meta.

[1]During training, a forced-alignment tool is used to obtain the alignment between the phonemes and the ground-truth Mel spectrogram. During inference, the duration predicted by the duration model is used.

10.21437/Interspeech.2024-562

and x_1 be the associated Mel spectrogram sampled from the real data distribution $q(x)$. The acoustic model defines a time-dependent vector field v_t, which is used to construct a flow ϕ_t as described by the differential equation:

$$\frac{d\phi_t(x)}{dt} = v_t(\phi_t(x), m(x_1), z_p; \theta) \quad (1)$$

Here, θ represents the model parameters, $t \in [0, 1]$ is the time parameter, and $m(\cdot)$ is a mask function applied to the Mel spectrogram x_1. Given x_0 sampled from the Gaussian prior $p(x)$, an ODE solver can be used to evaluate $\phi_1(x_0)$ with the initial condition $\phi_0(x_0) = x_0$.

The training objective is to align the flow-transformed distribution $p_1(x) = p(\phi_1^{-1}(x)) \det[\frac{\partial \phi_1^{-1}(x)}{\partial x}(x)]$ with the real data distribution $q(x)$. To achieve this, the acoustic model is trained to minimize the loss:

$$\mathbb{E}_{t,(x_1,z_p)\sim q,x_0\sim p}||v_t(\psi_t(x_0, x_1), m(x_1), z_p; \theta) - (x_1 - (1 - \sigma_{min})x_0)||^2 \quad (2)$$

with $\psi_t(x_0, x_1) = (1 - (1 - \sigma_{min})t)x_0 + tx_1, \sigma_{min} = 10^{-5}$. During training, t is uniformly sampled from $[0, 1]$ and is encoded as a sinusoidal positional embedding, which is concatenated with the phoneme embedding z_p, the masked spectrogram $m(x_1)$,[2] and the sampled $\psi_t(x_0, x_1)$ as the model input.

Sampling from the learned audio distribution $p_1(x|m(x_1), z_p)$ starts with a noise x_0 sampled from the Gaussian prior $p(x)$, followed by solving the ODE in equation (1) to obtain $\phi_1(x_0)$. During inference, we can choose to mask out all of x_1 for zero-shot text-to-speech or mask parts of x_1 to provide additional information to the Voicebox model.

4. Proposed method

4.1. Voicebox Adapter

Voicebox Adapter extends a pre-trained Voicebox model by incorporating additional modules to handle fine-grained conditions. Let z_f be the fine-grained condition, and θ' denote the parameters of the new fine-grained conditioning modules. The vector field modeled by the acoustic model is redefined as:

$$\frac{d\phi_t(x)}{dt} = v_t(\phi_t(x), m(x_1), z_p, z_f; \theta, \theta') \quad (3)$$

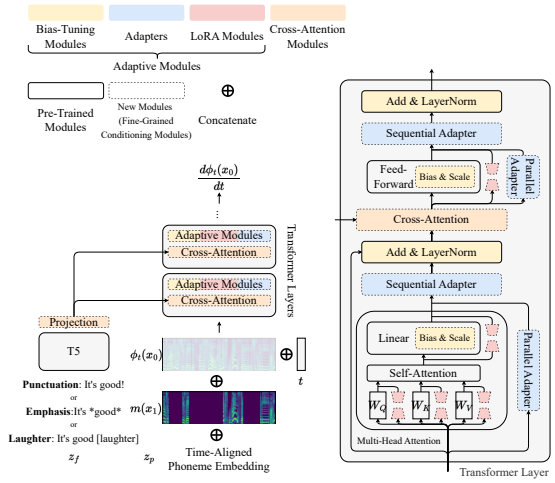
During fine-tuning, we keep the pre-trained parameters θ frozen and solely optimize the new parameters θ' with the loss:[3]

$$\mathbb{E}_{t,(x_1,z_p,z_f)\sim q,x_0\sim p}||v_t(\psi_t(x_0, x_1), m(x_1), z_p, z_f; \theta, \theta') - (x_1 - (1 - \sigma_{min})x_0)||^2 \quad (4)$$

An illustration of the acoustic model architecture is provided in Fig.1(a). The newly introduced modules comprise a frozen T5 encoder [19], a trainable linear projection layer, as well as cross-attention modules and adaptive modules within the Transformer stack. The T5 model takes fine-grained conditions as inputs and outputs a vector sequence, which is subsequently projected into a 768-dimensional space. Cross-attention modules attend to the projected vector sequence to integrate fine-grained conditions into the pre-trained model. Within each Transformer layer, parameter-efficient adaptive modules are used to ensure smooth integration of extracted information with the hidden features in the pre-trained Voicebox acoustic model. As Fig. 1(a) shows, the fine-grained conditions are formulated as transcripts with special annotations. This framework can potentially be applied to various fine-grained conditioning tasks, provided that the condition can be processed by the T5 encoder.

[2]During training, the mask function $m(\cdot)$ randomly masks out all or parts of the frames in the ground-truth Mel spectrogram x_1.

[3]Bias-tuning (see section 4.2 for details) is an exception, where the pre-trained LayerNorm layers remain trainable during fine-tuning.



(a) Voicebox Adapter *(b) Zoom-in view of a Transformer layer*

Figure 1: *The model architecture of Voicebox Adapter, and a zoom-in view of a Transformer layer.*

The implementation of the duration model resembles that of the acoustic model, where the frozen T5 encoder, projection layer, cross-attention modules, and adaptive modules are incorporated to provide fine-grained conditions for duration prediction. During fine-tuning, we freeze the pre-trained modules and only optimize the new modules as done in the acoustic model.

4.2. Efficient fine-tuning methods

To facilitate the integration of fine-grained conditions, we explore several efficient fine-tuning strategies, shown in Fig. 1(b).

- Adapters: We follow the configurations of parallel and sequential adapters in prior work [20] and add them to the self-attention and feed-forward layers of each Transformer layer.
- LoRA: We explore two different configurations. By default, we apply LoRA to the input projection matrices of the self-attention module. We also experiment with adding LoRA to every linear layer in the model, as prior work suggests [21].
- Bias-tuning: Following the approach of Llama Adapter V2 [17], we apply bias-tuning as an optional add-on to LoRA fine-tuning. In our implementation, bias vector $\mathbf{b}$ and scale vector $\mathbf{s}$ are used to modify the output of every linear layer ($\odot$ means element-wise product):

$$Linear_{bias-tuning}(\mathbf{x}) = (Linear(\mathbf{x}) + \mathbf{b}) \odot \mathbf{s} \quad (5)$$

  During fine-tuning, we optimize the bias and scale vectors jointly with the LayerNorm parameters while keeping the linear layer and other components frozen.

For all the adaptive modules, we employ zero-initialization [5] to ensure a smooth start to the fine-tuning process.

5. Experimental setup

5.1. Pre-training details

Our pre-trained Voicebox acoustic model comprises 12 Transformer layers with a model dimension of 768, and the hidden size of the feed-forward layers is 3,072. The model has 12 attention heads. We train the model for 750k updates on 32 GPUs, with a batch size of 8k tokens per GPU. The duration model has a similar architecture but has only 8 Transformer layers, a model dimension of 512, and a feed-forward hidden size of 2,048, and is trained for 400k updates on 32 GPUs. Other details of the models follow from the original setup of Voicebox [10].

For pre-training, we utilize a combination of three English datasets: a 60k-hour audiobook reading-style dataset, a 100-

hour podcast recording conversation-style dataset, and a 1.7k-hour telephone conversation-style dataset. These datasets collectively have over 20k speakers, providing wide coverage of speaking styles. To address the unbalanced size of the datasets, we implement data re-sampling to ensure equal exposure to the three datasets. We consider audio x to be an 80-dimensional log-scaled Mel spectrogram extracted with a 40ms window at a 100Hz frame rate, which can be converted to raw waveform with a HiFi-GAN vocoder [22].

5.2. Fine-grained conditional fine-tuning setup

The cross-attention module for integrating fine-grained conditions has 12 attention heads, each having a hidden dimension of 64. The configurations of the adaptive modules are as follows:

- Adapters: Sequential and parallel adapters are both 2-layer ReLU-activated feed-forward layers with a hidden size of 64.
- LoRA: We set the rank (also known as the hidden dimension) r and the scaling parameter α of LoRA both to 64 and use a dropout rate of 0.05.
- Bias-tuning: The weight and bias in the pre-trained LayerNorm layers and the newly introduced scale and bias vectors in linear layers are trained.

During fine-tuning, only the linear layer on top of the T5 encoder, the cross-attention, and the adaptive modules are trained. Other optimization setups remain identical to our pre-training setup, except for fewer training updates (see below).

We use different datasets for each fine-grained conditional fine-tuning task. For the punctuation task, we use a 550-hour audiobook dataset with punctuated cased transcription.[4] For the emphasis task, we use a 20-hour expressive speech dataset with word emphasis annotations in the transcription. For the laughter task, we use a 250-hour conversation dataset with special annotations for laughter. The acoustic model is fine-tuned for 50k updates and the duration model is fine-tuned for 100k updates in the punctuation and laughter tasks. In the emphasis task, due to limited fine-tuning data, we reduce the fine-tuning updates for the acoustic and duration models to 30k/50k, respectively.

5.3. Inference setup

The primary focus of Voicebox Adapter is on the zero-shot text-to-speech setup, where we mask out all content in $m(x_1)$, so the generation is solely conditioned on the text z_p and the fine-grained conditions z_f. To explore different aspects of our approach, we also conduct experiments using the prompted generation setup in selected cases. In this variant, we use the initial 3 seconds of the ground-truth recording as the speech prompt $m(x_1)$ to provide global information such as speaker, style, and environment noises, as opposed to the fine-grained information modeled by the proposed method. The prompted generation setup enables us to study the disentanglement of global and fine-grained conditions in Voicebox Adapter. Implementation details for both setups follow Voicebox [10].

5.4. Objective evaluation methods

5.4.1. Fine-grained controllability

We use automatic annotation tools to evaluate the adherence of the generated utterances to the fine-grained conditions. For emphasis and laughter tasks, we follow the approach of the word emphasis detector in the EmphAssess benchmark [23] to train the annotation models with our fine-grained conditional fine-tuning data. For punctuation, we use the English-only Whisper-small model [24] to add punctuation marks to the text transcription of generated speech by constraining the output of the Whisper decoder to either the next ground-truth text token or a punctuation mark in each decoding step. The pseudo labels predicted by the annotators are compared with the original fine-grained conditions, and the results are reported in F_1 scores[5] as in prior work [23]. As a sanity check, our annotation models achieve F_1 scores of 86.7%, 92.4%, and 62.1% on the validation set of the emphasis, laughter, and punctuation tasks, respectively.

5.4.2. Intelligibility and speaker similarity

To assess whether fine-grained conditional fine-tuning affects the speech generation capabilities of our models, we use word error rate (WER) to evaluate the content correctness and intelligibility of utterances generated in the zero-shot scenario, and report speaker similarity (SIM-o) between the speech prompts and the utterances generated in the prompted scenario. The setup of these evaluation methods follows Voicebox [10].

5.5. Subjective evaluation methods

For the subjective evaluation, we recruit 34 trained subjects with audio relevant experience to rate the generated speech samples and report our results using the 5-scale Mean Opinion Score (MOS). In the fine-grained controllability MOS test (FC-MOS), listeners are instructed to focus on the alignment between the generated utterances and the fine-grained conditions, while disregarding speech quality, style, speaker characteristics, and other irrelevant aspects. Conversely, in the quality MOS test (Q-MOS), listeners focus solely on the quality of the generated speech and disregard other factors. In each subjective evaluation, we randomly select 200 samples from the evaluation set (with the exception of 100 samples for the emphasis task due to limited data availability) and collect 5 ratings for each sample. We use the recommended recipes from the CrowdMOS [25] package to filter outliers and address inaccurate ratings, and report the averaged ratings with a 95% confidence interval.

6. Results

6.1. Effectiveness of different efficient fine-tuning methods

We use three datasets to evaluate Voicebox Adapter on the fine-grained conditional generation tasks. For the punctuation task, we use the *test* set of the LibriTTS dataset [26]. For the emphasis task, we randomly select 100 utterances with emphasis annotations from the Expresso dataset [27]. For the laughter task, we sample 27 speakers from the Switchboard dataset [28] and use their utterances with laughter annotations for evaluation.

First, we would like to identify the optimal strategy for fine-tuning the pre-trained Voicebox model with fine-grained conditions. Table 1 shows the results of the efficient fine-tuning

Table 1: *Performance of Voicebox Adapter with different adaptive modules on the fine-grained conditional generation tasks.*

Adaptive Modules	Punctuation F_1(%)	Emphasis F_1(%)	Laughter F_1(%)	params.*
Sequential adapter	63.6	66.9	39.9	2.4M
Parallel adapter	63.4	70.5	44.4	2.4M
LoRA (self-attention only)	63.5	74.7	44.5	3.5M
+ bias-tuning	63.8	75.6	50.4	3.6M
LoRA (all linear layers)	63.2	72.9	47.5	4.4M

*The number of parameters in the adaptive modules of the acoustic model. The pre-trained acoustic model has 93M parameters in total.

[4] Practically, we select the 8 most common punctuation marks (commas, periods, question marks, exclamation marks, colons, semicolons, hyphens, and quotes) for this task and remove other punctuation marks.

[5] For the punctuation task, we first compute the F_1 scores for each type of punctuation mark, and then report the micro-average across the 8 punctuation marks considered.

Table 2: *Main results for the evaluations of fine-grained controllability, quality, intelligibility, and speaker similarity.*

Models	Configuration			Punctuation (LibriTTS)					Emphasis (Expresso)			Laughter (Switchboard)		
	PT	X-attn.	Adapt.	F_1(%)	FC-MOS	WER(%)	SIM-o	Q-MOS	F_1(%)	FC-MOS	Q-MOS	F_1(%)	FC-MOS	Q-MOS
(a) Ground-truth				62.1	4.04±0.09	5.3	0.718	3.84±0.09	86.7	4.00±0.13	3.98±0.13	92.4	4.27±0.10	3.78±0.10
(b) Voicebox Adapter	✓	✓	LoRA+BT	63.8	4.18±0.07	3.2	0.593	3.88±0.09	75.6	3.57±0.18	3.82±0.13	50.4	2.15±0.17	3.67±0.09
(c) Fine-tune all*	✓	✓	✗	63.5	4.20±0.07	3.3	0.568	3.89±0.08	71.9	3.62±0.16	3.82±0.13	57.5	2.55±0.19	3.64±0.09
(d) No pre-training*†	✗	✓	✗	63.5	4.20±0.07	3.6	0.512	3.90±0.09	68.4	3.30±0.16	3.61±0.15	57.1	2.61±0.19	3.45±0.10
(e) Voicebox*†	✗	✗	✗	57.7	3.85±0.08	3.5	0.565	3.82±0.09	39.3	2.99±0.13	3.78±0.14	32.7	1.46±0.12	3.64±0.09

PT: pre-training; X-attn.: cross-attention; Adapt.: adaptive modules; BT: bias-tuning.
*For models without adaptive modules, we unfreeze all model parameters (except for T5) when training the models on the fine-grained conditioning datasets.
†For models without pre-training, we train the acoustic and duration models for 150k/400k updates on the fine-grained conditioning datasets to ensure full convergence.

strategies we explore. We find that parallel adapters outperform sequential adapters, which is consistent with prior research in both text and speech models [20, 29]. Notably, the LoRA + bias-tuning configuration demonstrates superior results across all tasks. Considering that the differences in the parameter counts of the adaptive modules for each fine-tuning method are marginal — accounting for less than 5% of the pre-trained parameters — we adopt LoRA + bias-tuning as our best setup and use it as our default configuration for the remainder of the paper.

6.2. Main results

To verify the effectiveness of Voicebox Adapter, we conduct a comprehensive comparison against the baseline Voicebox model and other alternative training configurations. The results of our objective and subjective evaluations are shown in Table 2.

In the fine-grained controllability metrics F_1 and FC-MOS, we observe that models incorporating fine-grained conditioning inputs (rows (b), (c), and (d)) consistently outperform the baseline Voicebox model (row (e)). Despite fine-tuning only a small portion of parameters, Voicebox Adapter (row (b)) shows performance comparable to fine-tuning the entire model (row (c)). The comparison between rows (b) and (d) highlights the importance of pre-training as Voicebox Adapter achieves comparable performance with significantly fewer fine-tuning iterations.

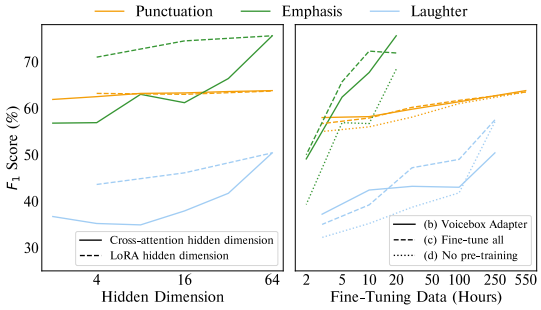
For models without applying efficient fine-tuning (row (c) and (d)), we find that pre-training does not make large differences to the punctuation and laughter tasks, but results in an improvement in the emphasis task, where the amount of data is limited. This further shows the potential of the pre-training & fine-tuning strategy in resource-constrained scenarios.

We notice that all models have worse performance on the Laughter task. This challenge may be attributed to limited presence of laughter in the pre-training data. Our analysis shows that only 1.4% of frames in the pre-training data trigger the laughter detector developed in section 5.4, whereas 6.3% of the frames trigger the emphasis detector. This discrepancy highlights the importance of adequate representation of various speech characteristics during pre-training.

The Q-MOS evaluation further shows that the efficient fine-tuning of Voicebox Adapter does not compromise the quality of the generated speech (row (b) vs. others). Further, the comparison between rows (b) (c) against row (d) shows that pre-training enhances the quality of generated speech, particularly on the Expresso and Switchboard datasets. The WER and SIM-o evaluations on the LibriTTS dataset validate that Voicebox Adapter retains the intelligibility and one-shot speech generation ability of Voicebox [10] and even exhibits improved performance, which may be attributed to the pre-training on large corpora.

6.3. Hyper-parameter and data ablation studies

In Fig. 2(a), we present the results of models with different cross-attention dimensions and LoRA hidden dimensions r. In general, we observe a positive correlation between the hidden dimensions and the performance of the models. These findings contrast with the results of the original LoRA experiments on Large Language Models (LLMs) [16], which may be attributed to the size difference between Voicebox Adapter and state-of-the-art LLMs. As reported, small intrinsic dimensions in large models [30] potentially lead to the good performance of LLMs with small LoRA hidden dimensions. Given that Voicebox Adapter has less than 0.1% of the parameters of GPT-3 [31], it is conceivable that a smaller r could be more effective as the model size scales up. However, this hypothesis requires further investigation, which we plan to explore more in future work.



(a) Voicebox Adapter with different hidden dimension sizes *(b) Voicebox Adapter and compared models trained with different amounts of data*

Figure 2: *Performance of Voicebox Adapter and compared models with different hidden dimensions and data configurations.*

In Figure 2(b), we fine-tune Voicebox Adapter on subsets of the fine-tuning data and compare their performance. We find that the efficient fine-tuning method consistently achieves performance comparable to fine-tuning the entire model, regardless of the amount of data used in the fine-tuning process. This observation holds true across different tasks, affirming the robustness of Voicebox Adapter across various data setups.

7. Conclusion

In this paper, we introduce Voicebox Adapter, which augments a pre-trained Voicebox speech generation model with fine-grained conditions. We inject fine-grained conditions into the Voicebox model through cross-attention modules and investigate various efficient fine-tuning methods to facilitate the integration of pre-trained parameters with newly introduced modules. On the three fine-grained conditional generation tasks — punctuation, emphasis, and laughter — Voicebox Adapter successfully adheres to the fine-grained conditions while maintaining the naturalness and intelligibility of the generated speech. In our subjective and objective evaluations, Voicebox Adapter shows superior performance over the baseline Voicebox model and achieves comparable performance to fine-tuning the entire model.Future work involves extending the proposed method to more diverse fine-grained conditions, as well as addressing challenges in specific tasks such as laughter generation.

8. References

[1] S. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. Lee, "SUPERB: Speech processing universal performance benchmark," in *Interspeech*, 2021.

[2] A. Mohamed, H. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

[4] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[5] A. H. Liu, M. Le, A. Vyas, B. Shi, A. Tjandra, and W.-N. Hsu, "Generative pre-training for speech with flow matching," in *ICLR*, 2024.

[6] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed, and E. Dupoux, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[7] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour, "AudioLM: A language modeling approach to audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2523–2533, 2023.

[8] T. A. Nguyen, E. Kharitonov, J. Copet, Y. Adi, W.-N. Hsu, A. Elkahky, P. Tomasello, R. Algayres, B. Sagot, A. Mohamed, and E. Dupoux, "Generative spoken dialogue language modeling," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 250–266, 2023.

[9] C. Wang, S. Chen, Y. Wu, Z. 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," *preprint arXiv:2301.02111*, 2023.

[10] 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 *NeurIPS*, 2023.

[11] Y. Chen, Y. M. Assael, B. Shillingford, D. Budden, S. E. Reed, H. Zen, Q. Wang, L. C. Cobo, A. Trask, B. Laurie, Ç. Gülçehre, A. van den Oord, O. Vinyals, and N. de Freitas, "Sample efficient adaptive text-to-speech," in *ICLR*, 2019.

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

[13] Y. Yan, X. Tan, B. Li, G. Zhang, T. Qin, S. Zhao, Y. Shen, W.-Q. Zhang, and T.-Y. Liu, "Adaptive text to speech for spontaneous style," in *Interspeech*, 2021.

[14] R. Huang, Y. Ren, J. Liu, C. Cui, and Z. Zhao, "Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech," in *NeurIPS*, 2022.

[15] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *ICML*, 2019.

[16] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *ICLR*, 2022.

[17] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue, H. Li, and Y. Qiao, "LLaMA-Adapter V2: Parameter-efficient visual instruction model," *preprint arXiv:2304.15010*, 2023.

[18] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *ICLR*, 2023.

[19] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[20] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig, "Towards a unified view of parameter-efficient transfer learning," in *ICLR*, 2022.

[21] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in *NeurIPS*, 2023.

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

[23] M. de Seyssel, A. D'Avirro, A. Williams, and E. Dupoux, "EmphAssess: A prosodic benchmark on assessing emphasis transfer in speech-to-speech models," *preprint arXiv:2312.14069*, 2023.

[24] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023.

[25] F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer, "CROWDMOS: An approach for crowdsourcing mean opinion score studies," in *ICASSP*, 2011.

[26] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Interspeech*, 2019.

[27] S. Communication, L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenthaler, P.-A. Duquenne, B. Ellis, H. Elsahar, J. Haaheim, J. Hoffman, M.-J. Hwang, H. Inaguma, C. Klaiber, I. Kulikov, P. Li, D. Licht, J. Maillard, R. Mavlyutov, A. Rakotoarison, K. R. Sadagopan, A. Ramakrishnan, T. Tran, G. Wenzek, Y. Yang, E. Ye, I. Evtimov, P. Fernandez, C. Gao, P. Hansanti, E. Kalbassi, A. Kallet, A. Kozhevnikov, G. M. Gonzalez, R. S. Roman, C. Touret, C. Wong, C. Wood, B. Yu, P. Andrews, C. Balioglu, P.-J. Chen, M. R. Costa-jussà, M. Elbayad, H. Gong, F. Guzmán, K. Heffernan, S. Jain, J. Kao, A. Lee, X. Ma, A. Mourachko, B. Peloquin, J. Pino, S. Popuri, C. Ropers, S. Saleem, H. Schwenk, A. Sun, P. Tomasello, C. Wang, J. Wang, S. Wang, and M. Williamson, "Seamless: Multilingual expressive and streaming speech translation," *preprint arXiv:2312.05187*, 2023.

[28] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *ICASSP*, 1992.

[29] Q. Li, B. Li, D. Hwang, T. Sainath, and P. M. Mengibar, "Modular domain adaptation for conformer-based streaming ASR," in *InterSpeech*, 2023.

[30] A. Aghajanyan, S. Gupta, and L. Zettlemoyer, "Intrinsic dimensionality explains the effectiveness of language model fine-tuning," in *ACL-IJCNLP*, 2021.

[31] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *NeurIPS*, 2020.