LoRA-TTS: Efficient Dialect Adaptation in Text-to-Speech using LoRA

Andrew Zhou
University of Washington
Seattle, United States

Huy Huynh
University of Washington
Seattle, United States

***Abstract*—Large-scale Text-to-Speech (TTS) models often struggle to adapt to specific dialects or accents that fall outside their training distribution, resulting in robotic or distorted prosody. This paper addresses the challenge of adapting a Standard Mandarin-trained model (XTTS-V2) to Taiwanese Mandarin using a low-resource dataset of only 500 examples. We propose using Low-Rank Adaptation (LoRA) to bridge this accent gap efficiently. We compare three approaches: a zero-shot baseline, full fine-tuning, and LoRA adaptation. Our results show that LoRA achieves comparable prosodic naturalness and speaker similarity to full fine-tuning while reducing the number of trainable parameters by approximately 98%. This demonstrates that high-fidelity accent adaptation is achievable with minimal compute and data, offering a viable solution for democratizing diverse speech synthesis.**

I. Introduction

Recent advancements in generative Text-to-Speech (TTS) have enabled remarkable capabilities in zero-shot voice cloning. State-of-the-art multi-speaker models, such as XTTS-V2 [1] and YourTTS [2], can now synthesize speech in the voice of unseen speakers given only a few seconds of reference audio. These models leverage large-scale multilingual datasets to learn a latent space of speaker embeddings, allowing for the simulation of diverse timbres and styles without the need for retraining.

However, a significant challenge remains in adapting these models to specific dialects or accents that fall outside their primary training distribution. While zero-shot models are effective at cloning vocal timbre (the ”sound” of the voice), they often fail to capture fine-grained local prosody, such as specific tonal patterns or accent-dependent articulation. A prominent example of this failure case is the adaptation of Standard Mandarin models to Taiwanese Mandarin. Despite sharing a character set, Taiwanese Mandarin exhibits distinct tonal flatness and lacks the retroflex sounds (e.g., zh, ch, sh) characteristic of the Beijing-centric Standard Mandarin usually found in training data. Consequently, when zero-shot models are prompted with Taiwanese reference audio, the resulting synthesis frequently suffers from ”robotic” artifacts and prosodic distortion, as the model attempts to force an out-of-distribution accent into its learned latent space.

Traditionally, addressing this domain shift requires Full Fine-Tuning (FFT) of the model on the target dialect. While effective at repairing prosodic mismatches, FFT is computationally prohibitive for many applications. It requires updating hundreds of millions of parameters, demanding significant GPU resources and storage. Furthermore, full fine-tuning on small datasets carries the risk of overfitting or ”catastrophic forgetting,” where the model degrades in its general linguistic capabilities.

To bridge this gap, we propose the application of Low-Rank Adaptation (LoRA) [3] to the TTS domain for efficient dialect adaptation. Originally developed for Large Language Models (LLMs), LoRA freezes the pre-

trained model weights and injects trainable rank-decomposition matrices into the architecture's layers. This allows the model to adapt to new domains by optimizing only a minute fraction of the total parameters.

In this work, we investigate the efficacy of LoRA for adapting a Standard Mandarin XTTS-V2 model to Taiwanese Mandarin using a low-resource dataset of only 500 examples. We conduct a comparative analysis between a zero-shot baseline, a fully fine-tuned model, and a LoRA-adapted model. Our code can be found at https://github.com/andouu/cse493s-final. Our primary contributions are as follows:

1) Low-Resource Accent Adaptation: We demonstrate that high-fidelity accent adaptation is achievable with as few as 500 training samples, significantly lowering the data barrier for custom voice synthesis.
2) Efficiency vs. Quality Analysis: We show that LoRA achieves prosodic naturalness and speaker similarity comparable to full fine-tuning while reducing the number of trainable parameters by approximately 98%.
3) Mitigation of Domain Shift: We provide empirical evidence that fine-tuning (both FFT and LoRA) successfully eliminates the robotic artifacts present in zero-shot inference, validating the necessity of weight adaptation for out-of-distribution accents.

II. Methodology

In this section, we describe the architecture of the base Text-to-Speech model, the Low-Rank Adaptation (LoRA) formulation applied to the attention mechanism, and the data preparation process for low-resource dialect adaptation.

A. Base Model Architecture

We utilize XTTS-v2 [1] as our foundational model. XTTS is a multi-speaker, autoregressive TTS system built on a GPT-like architecture. Unlike traditional Tacotron-based systems that predict mel-spectrograms directly, XTTS operates over discrete audio tokens. The framework consists of three primary components:

1) VQ-GAN Encoder: Compresses raw audio into a sequence of discrete latent tokens (audio codes).
2) GPT Predictor: An autoregressive transformer that predicts these audio tokens conditioned on input text tokens and a reference speaker embedding.
3) HifiGAN Decoder: Reconstructs the high-fidelity waveform from the predicted audio tokens.

During adaptation, the VQ-GAN and HifiGAN components are frozen. We focus our fine-tuning exclusively on the GPT Predictor, which is responsible for modeling the prosody, rhythm, and linguistic-acoustic alignment of the target speech.

B. Low-Rank Adaptation (LoRA

To enable efficient adaptation to the Taiwanese Mandarin dialect without the computational cost of full fine-tuning, we apply Low-Rank Adaptation (LoRA) [3]. Standard fine-tuning updates the entire pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$ by computing a full update ΔW. LoRA hypothesizes that the change in weights during adaptation has a low intrinsic rank. Therefore, we freeze W_0 and inject two trainable low-rank decomposition matrices A and B:

$$W = W_0 + \Delta W = W_0 + BA$$

where $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$, and the rank $r \ll \min(d, k)$.

For our experiments, we set the rank $r = 8$. We apply LoRA modules specifically to the `c_attn` and `c_proj` layers, which are the inputs and outputs to the self-attention mechanism of the model, respectively. This configuration allows us to update the model's prosodic attention mechanisms while reducing the number of trainable parameters by approximately 98% compared to full fine-tuning.

C. Dataset

Our objective is to adapt the model to a Taiwanese Mandarin accent using a low-resource dataset. We utilize a subset of the Common Voice Chinese (Taiwan) dataset.

- Data Selection: We sampled 500 audio-text pairs from native Taiwanese speakers. The total audio duration is approximately 30 minutes.
- Language Distribution: While the base XTTS model includes Mandarin Chinese data, it is predominantly Standard Mandarin. Our subset represents a domain shift toward Taiwanese Mandarin, characterized by flatter tonal contours and specific phonological variances.

D. Experimental Conditions

We compare three experimental conditions:

1) Zero-Shot (Base): The pre-trained XTTS-v2 model conditioned on a reference Taiwanese audio clip without weight updates.
2) Full Fine-Tuning (FFT): Updating all weights of the GPT predictor
3) LoRA Adaptation: Updating only the low-rank adapters ($r = 8$) injected into the attention layers.

.

All experiments were conducted on a Google Colab instance using a single NVIDIA T4 GPU (16GB VRAM). Models were trained for 30 epochs using the AdamW optimizer with a learning rate of 5×10^{-4} for regular fine-tuning, 2×10^{-4} for LoRA, and a batch size of 3. The LoRA implementation utilized the peft library for efficient parameter injection.

III. Results

In this section, we present the quantitative and qualitative evaluation of our proposed LoRA adaptation compared to the zero-shot baseline and full fine-tuning (FFT). We focus on three key dimensions: speech intelligibility, speaker similarity, and parameter efficiency.

A. Quantitative Evaluation

To assess the quality of the synthesized speech quantitatively, we employ two standard metrics: Character Error Rate (CER) and Speaker Encoder Cosine Similarity (SECS). We utilized a pre-trained Whisper (Large-v3) model to transcribe the generated audio for CER calculation, measuring intelligibility. For SECS, we extracted speaker embeddings using a Resemblyzer encoder to measure the similarity between the generated speech and the ground truth reference audio.

Our results indicate a close, but clear hierarchy in performance. The Zero-Shot Base model exhibited the highest CER, reflecting the ”robotic” artifacts and pronunciation errors caused by the domain shift between Standard and Taiwanese Mandarin. Full Fine-Tuning (FFT) achieved the best SECS, effectively resolving the prosodic mismatch. The LoRA (Rank 8) model demonstrated competitive performance, bridging the gap between the baseline and FFT. While the LoRA model's SECS were slightly lower than the fully fine-tuned model, it significantly outperformed the zero-shot baseline and the FFT model in CER, validating that low-rank adaptation is sufficient for capturing dialect-specific features.

B. Qualitative Evaluation

To strictly quantify the ”robotic” artifacts observed in the baseline and measure accent preservation, we conducted a Mean Opinion Score (MOS) study. A group of native Mandarin speakers, blinded to the model types listened to 5 randomly selected samples from each condition. They rated each sample on a 5-point Likert scale for Naturalness (prosodic realism) and Accent Similarity (faithfulness to the Taiwanese target dialect).

As shown in Table I, the results establish a clear hierarchy in performance:

1) Full Fine-Tuning (FFT): This model ranked closest to the ground truth. By updating all parameters, it successfully

re-aligned the prosody to the Taiwanese dialect.

2) LoRA (Rank 8): Our proposed method ranked third but remained highly competitive. While it occasionally retained minor prosodic stiffness in complex sentence structures, it successfully captured the target accent's timbre.
3) Zero-Shot (Base): The baseline consistently ranked last, resulting in a similarity score of only 3.32. This confirms that zero-shot conditioning alone is insufficient for this degree of domain shift.

TABLE I
COMPARISON OF QUANTITATIVE METRICS (CER, SECS) AND QUALITATIVE MEAN OPINION SCORES (MOS) FOR TAIWANESE MANDARIN ADAPTATION. THE BEST SCORES ARE IN **BOLD**.

Method	CER ↓	SECS ↑	Nat. ↑	Sim. ↑
Base Model	0.4691	0.77	3.44	3.32
FFT	0.4617	**0.81**	3.44	**3.72**
LoRA	**0.3960**	0.77	**3.76**	3.68

C. Efficiency Analysis

While FFT ranks highest in subjective quality, the trade-off in computational cost is substantial. As detailed in the methodology, the FFT model requires updating and storing the complete parameter set (∼500M parameters). In contrast, the LoRA model, which trails the FFT model by only 0.04 points in similarity MOS, and outperforms FFT by 0.32 in naturalness MOS, reduces the trainable parameter count by 98% (∼2M parameters). This result positions LoRA as the optimal solution for scalable deployment: it delivers performance significantly closer to the Ground Truth than the Base Model, with a computational footprint a fraction of the size of FFT.

IV. DISCUSSION

A. Bridging the Prosodic Gap

Our experiments confirm that zero-shot adaptation, while powerful for general voice cloning, struggles significantly with out-of-distribution prosody. The base XTTS-v2 model, despite having a strong grasp of Standard Mandarin phonology, failed to produce the characteristic tonal flatness and soft articulation of the Taiwanese accent. This supports our hypothesis that the "robotic" artifacts observed in the baseline are not due to a lack of speaker identity information, but rather a domain shift in prosodic mapping. By fine-tuning the model, we effectively realigned this mapping. The Full Fine-Tuning (FFT) model achieved the highest naturalness scores because it was allowed to update all parameters to fit this new domain.

B. The Efficiency of Low-Rank Adaptation

Crucially, our LoRA implementation demonstrated that this prosodic realignment does not require a full model update. By modifying only the attention mechanisms (via Rank 8 decomposition), the LoRA model successfully mitigated the robotic artifacts found in the baseline. While the quantitative metrics for LoRA were slightly lower than those of the fully fine-tuned model, the perceptual improvement over the zero-shot baseline was substantial. This suggests that the core "identity" of an accent is largely encoded in the attention layers of the transformer, which regulate how the model attends to linguistic tokens over time.

C. Efficiency Trade-Offs

The most significant finding of this study is the efficiency-quality trade-off. The full fine-tuning approach required storing a complete copy of the model parameters (∼2GB) for a single speaker. In contrast, the LoRA checkpoint requires only ∼30MB of storage, a reduction of approximately 98% in trainable parameters. For deployment scenarios involving thousands of custom voices, this difference is non-trivial. Our results indicate that the slight degradation in objective metrics (compared to FFT) is a justifiable cost for the massive gain in storage efficiency and portability.

D. Low-Resource Feasibility

Finally, we demonstrated that this adaptation is feasible with an extremely small dataset (500 samples, roughly 30 minutes of audio). Previous dialect adaptation methods often require tens of hours of high-quality studio data. Our success with a small, randomly sampled subset suggests that modern pre-trained TTS backbones are robust few-shot learners when paired with parameter-efficient fine-tuning techniques.

V. Conclusion and Future Work

In this paper, we addressed the limitations of zero-shot TTS models in synthesizing out-of-distribution dialects, specifically targeting the adaptation of Standard Mandarin models to Taiwanese Mandarin. We proposed the use of Low-Rank Adaptation (LoRA) as a computationally efficient alternative to full fine-tuning.

Our results show that LoRA adaptation effectively bridges the "accent gap," eliminating the robotic artifacts associated with zero-shot inference. While full fine-tuning remains the upper bound for quality, LoRA achieves comparable prosodic fidelity and speaker similarity while reducing the number of trainable parameters by 98%. Furthermore, we showed that this adaptation is effective even with low-resource datasets as small as 500 samples.

These findings suggest that LoRA is a viable strategy for democratizing high-fidelity speech synthesis, allowing for the rapid creation of diverse, accented voices on consumer-grade hardware without the prohibitive costs of traditional training.

A. Future Work

While this study demonstrated the efficacy of LoRA for single-speaker accent adaptation, broader opportunities lie in the disentanglement of accent from speaker identity. Future research should investigate whether "Accent LoRAs" can be trained independently of speaker embeddings, allowing a single dialect adapter to be applied to multiple different voices (i.e., universal accent transfer).

Additionally, as the field moves toward non-autoregressive architectures like Flow Matching and Diffusion transformers, validating the efficiency of low-rank adaptation on these newer backbones remains a critical open question. Finally, the development of robust objective metrics for accent fidelity—beyond simple speaker similarity—would significantly accelerate progress in dialect adaptation by reducing the reliance on resource-intensive human evaluation.

VI. Limitations

While our method improves significantly over the baseline, it is not without limitations. First, the quality of LoRA adaptation is highly sensitive to the acoustic quality of the training data. Given the low-resource nature of the 500-shot dataset, noisy or reverberant samples can introduce artifacts that the LoRA layers may overfit to, degrading the clarity of the synthesized speech.

Second, while LoRA effectively realigns prosody, it is constrained by the phonological inventory of the base model. If the target dialect contains unique phonemes or articulation patterns that are entirely absent from the base model's pre-training data, LoRA cannot "invent" these sounds from scratch, potentially leading to approximate pronunciations.

Finally, our experimental scope was constrained by a limited computational budget. We restricted our training to a small subset of data and relatively few training steps due to hardware limitations. It is likely that with a larger dataset and extended training time, the performance gap between the LoRA-adapted model and the zero-shot baseline would be even more pronounced, further validating the efficacy of the method.

VII. Contributions

A. Andrew Zhou

- Formatted Mozilla Common Voice Taiwanese Mandarin dataset into LJSpeech format

- Implemented regular fine tuning on xTTS-v2 model
- Implemented LoRA fine tuning on xTTS-v2 model

B. Huy Huynh

- Ran inference on FFT, LoRA, and base models on the test set
- Implemented and ran evaluation code for metrics
- Created and collected data from user study

REFERENCES

[1] E. Casanova *et al.*, "XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model," in *Proc. INTERSPEECH*, 2024. [Online]. Available: https://arxiv.org/abs/2406.04904

[2] E. Casanova, J. Weber, C. 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 *International Conference on Machine Learning (ICML)*, 2022, pp. 2709–2720.

[3] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," in *Proc. International Conference on Learning Representations (ICLR)*, 2022. [Online]. Available: https://arxiv.org/abs/2106.09685