Multi-Dialect Vietnamese TTS Final Report (CSCI 5541 NLP)

Cheston Opsasnick, Vy Bui Nguyen, Annalise Xiao, Quenton Ni
Team Lower Expectations

Abstract

This project proposes a multi-dialect Text-to-Speech (TTS) system for Vietnamese. This was achieved by utilizing a continuous flow-matching approach on a dialect-annotated dataset in order to achieve a higher degree of naturalness and intelligibility compared to traditional pipelines. To evaluate the framework's performance, the project measured overall speech intelligibility using Word Error Rate (WER) and Mean Opinion Score (MOS) to verify if the synthesized speech matches the target prompt. The expected outcome is a speech synthesis system capable of generating distinct Vietnamese dialects using minimal reference audio.

1 Introduction

In recent years, Text-to-Speech (TTS) systems have rapidly improved in naturalness and intelligibility with paradigms like continuous flow matching and audio language models. However, most of these models target popular, high-resource languages such as English and Mandarin. Despite substantial advancements in TTS technology, low resource languages such as Vietnamese remain underrepresented. This disparity is particularly evident when attempting to synthesize diverse regional dialects. The strict lexical tones and significant phonetic shifts between Northern, Central, and Southern Vietnamese present complex acoustic challenges that traditional synthesis pipelines and English-centric models struggle to resolve without dedicated, dialect-aware fine-tuning. This project addresses that gap by fine-tuning the F5-TTS architecture as shown in Figure 1 on multi-dialect Vietnamese data.

2 Background

2.1 Current Practices and Limitations

Previous work has focused on dialect identification and speech recognition for Vietnamese, but less attention has been given to how dialect variation affects TTS models. Existing TTS systems are trained on standard Vietnamese, which is heavily based on the Northern dialect. Central and Southern Vietnamese speakers are systematically underrepresented. The three regional Vietnamese dialects differ in consonants, vowels, and tonal contours, all of which carry lexical meaning. No prior work has examined whether a TTS model trained on speech data from multiple Vietnamese dialects can preserve dialect-specific characteristics or whether it collapses toward the dominant dialect during generation.

2.2 Motivation and Novelty

This project investigates whether a TTS model trained on Northern, Central, and Southern Vietnamese dialects can accurately generate dialect-specific acoustic characteristics. Vietnamese is underexplored as a low resource language in speech technology due to the limited availability linguistic data (Bui et al., 2025). Regional dialects with differences in pronunciation and vocabulary may also challenge TTS models. By investigating how a TTS model behaves when trained on multiple Vietnamese dialects, this project contributes to understanding how neural speech synthesis systems handle linguistic variation in low-resource languages. The results may also provide insights into improving dialect-aware TTS models for other low-resource languages. If successful, this work could serve as a blueprint for extending similar approaches to other dialect-rich, low-resource languages. More immediate benefits include improving accessibility in applications like screen readers, navigation tools, and voice assistants for Vietnamese speakers from Central and Southern regions who are underrepresented in existing speech tools.

2.3 F5-TTS Architecture

To address the challenges of dialect rich synthesis, this project utilizes the F5-TTS (Fairytaler that Fakes Fluent and Faithful speech with Flow matching) architecture (Chen et al., 2024). Traditional TTS pipelines typically require complex components, including explicit phoneme aligners and dedicated duration models. F5-TTS circumvents these components by employing a fully non-autoregressive framework based on Condition Flow Matching and a Diffusion Transformer (DiT). Instead of requiring prealigned phonemes, the system converts text directly into a character sequence, which is then run through convolutional layers (ConvNeXt V2) to extract localized linguistic features.

3 Approach

3.1 Hypothesis

The hypothesis is that training on segmented, multi-dialect Vietnamese speech data will produce intelligible speech with distinguishable dialect characteristics, as measured by Word Error Rate (WER) and Mean Opinion Score (MOS).

This design is expected to be effective because high-quality, duration-controlled audio samples with aligned transcripts reduce noise in the training process and improve model convergence. Additionally, explicitly incorporating dialect diversity during training should enable the model to learn distinct phonetic and tonal variations.

3.2 Data Extraction

Speech samples labeled as Northern, Central, and Southern dialects were extracted from the Vietnamese Multi-Dialect (ViMD) dataset (Dinh et al., 2024) by streaming the nguyendv02/ViMD_Dataset repository from HuggingFace. The dataset was streamed rather than downloaded in full to accommodate compute constraints.

3.2.1 Audio Processing and Clip Segmentation

Each speech sample was filtered for metadata validity (non-empty transcript, recognized region label, and non-null audio) and constrained to a duration range of 3 to 17 seconds. All audio was converted to mono and resampled to 24 kHz. Samples within this duration range were retained with their original transcripts. Longer clips exceeding 17 seconds, but under a 30-second threshold, were processed using a two-stage segmentation and transcription pipeline consisting of Silero Voice Activity Detection (VAD) and neural ASR models.

Silero VAD (ax Inc., 2023) was used to identify speech regions and segment audio into candidate chunks, targeting approximately 15 seconds per segment. These segments were subsequently transcribed using vinai/PhoWhisper-medium (Research, 2023), a Vietnamese ASR model built on OpenAI's Whisper architecture and fine-tuned for Vietnamese speech.

Before transcription, audio segments were downsampled from 24 kHz to 16 kHz to meet model input requirements and normalized to mitigate clipping. Transcription was performed with language='vi' and task='transcribe' explicitly specified. Inference was executed under torch.inference_mode() using float16 precision on GPU. The resulting transcripts were post-processed by normalizing whitespace and correcting punctuation spacing before being written to the dataset manifest. Segments that failed to produce valid transcripts were discarded. VAD-derived segments without corresponding ASR outputs were flagged for manual alignment and excluded from training. Audio samples exceeding 30 seconds were omitted entirely.

In total, 17,893 out of 22,828 samples (78.38%) were generated through the VAD-based segmentation and transcription pipeline, while the remaining 4,935 samples (21.62%) retained their original ViMD transcripts. This distribution reflects the predominance of longer, conversational recordings in the source dataset, necessitating segmentation for effective use in TTS training.

3.2.2 Dataset Statistics

The final extracted dataset contained 22,828 samples, totaling to about 71 hours of speech across the three dialects. The average clip duration was roughly 11.1 to 11.2 seconds. The per-dialect breakdown is shown in Table 1.

Dialect	Samples	Total Hours	Avg Duration(s)
Northern	8,007	25.0	11.24
Central	7,615	23.7	11.21
Southern	7,206	22.2	11.10
Total	**22,828**	**70.9**	**11.18**

Table 1: Extracted dataset statistics by dialect region.

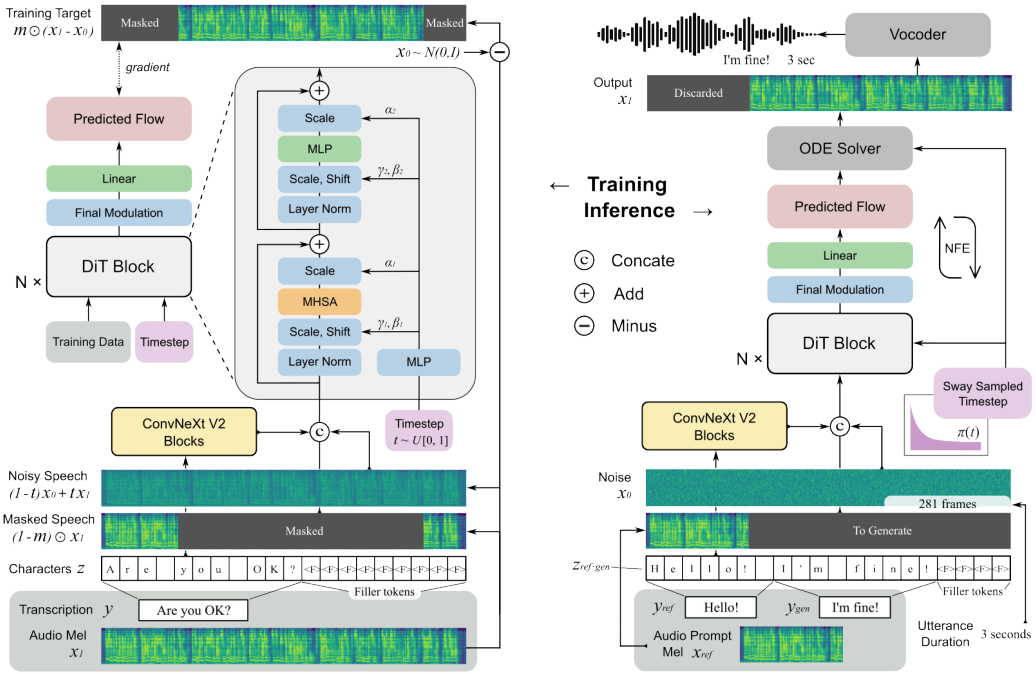

Figure 1: Overview of F5-TTS Training (left) and Inference (right).

3.3 Fine-tuning

3.3.1 Vocabulary Adaptation and Base Model Initialization

The fine-tuning process was initialized using the baseline F5-TTS architecture. The F5-TTS architecture was pre-trained on approximately 100,000 hours of Chinese and English audio data. The default vocabulary mapping was explicitly expanded to incorporate all Vietnamese specific characters (e.g., ă, â, đ, ê, ô, ơ, ư) and their respective tonal variations. Since the vocabulary had to be expanded to accommodate Vietnamese, the base model's embedding layer had to be overwritten and initialized with new weights. This expansion was critical to ensure that the text to spectrogram mechanisms could accurately map acoustic features to their phonetic representations.

3.3.2 Hardware and Hyperparameters

Training was conducted in a Google Colab environment, utilizing an NVIDIA RTX PRO 6000 Blackwell Server Edition, allocated with approximately 96 GB of VRAM. The fine-tuning process was initialized from a mature pre-trained F5-TTS checkpoint (pretrained_model_1250000.safetensors).
The batch size was scaled to 30,000 frames to maximize the utilization of the 96 GB VRAM on the GPU. To maintain stable convergence and prevent gradient explosions under this batch capacity, the learning rate was set to 1×10^{-5}. The model was configured to train for 50 epochs over the extracted 71 hour dataset. Training information, hardware metrics, and loss curves were continuously monitored using Weights & Biases. To track model progression and prevent data loss, the pipeline captured persistent checkpoints every 5,000 updates, with a rolling save state generated every 500 updates. Audio samples synthesized during persistent checkpoints were also logged to help evaluate the model while it trained.

3.4 Inference

Inference was conducted by loading the optimized weights from the fine-tuned F5-TTS checkpoint. F5-TTS uses an in-context approach, requiring a reference audio clip and its corresponding transcript to condition the generated output. To synthesize the distinct regional accents, a set of reference samples were established, representing the Northern, Central, and Southern Vietnamese dialects. During inference, the target dialect can be controlled by swapping these contexts. This methodology allowed the abstraction of a single, unified model to generate multi dialect outputs without needing discrete checkpoints for each region.

One constraint that needed to be managed during inference was the architecture's 30 second maximum context window. Because this limit encompasses both reference clips and generated audio,

reference clips were limited to a maximum duration of 10 seconds. This limit provided a nice balance for the model to capture reference characteristics while still having enough of a buffer for generation.

4 Results

4.1 Training Loss

Figure 2 shows the per-step training loss over 41,000 steps without smoothing, revealing the full variance of the flow-matching objective across individual batches. The median loss (dark line) remains relatively stable throughout training, hovering around 0.58 to 0.62 after an initial settling period in the first 5,000 steps. However, the shaded variance band spans approximately 0.45 to 0.90 at nearly every stage of training, with isolated spikes reaching 1.0. This high step-level variance is characteristic of batch-level instability, rather than model divergence. The running median does not trend upward, indicating that training remained globally stable. The persistent width of the variance band from step 15,000 onward suggests that the model reached a plateau in its ability to reduce per-batch uncertainty.

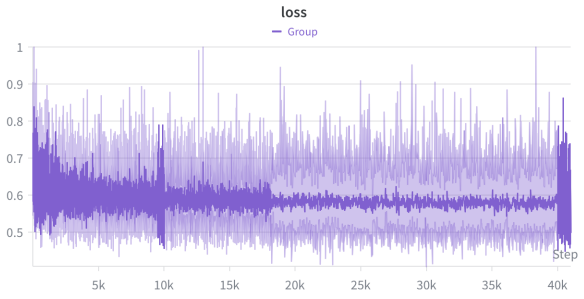

Figure 2: Training Loss

4.2 Time Weighted EMA Loss

Figure 3 shows the training loss curve over 41,000 update steps. The loss decreased rapidly during steps 0 to 5,000, dropping from approximately 0.65 to 0.60. This indicates that the model quickly adapted the pretrained F5-TTS weights to the Vietnamese data distribution. The loss continued to decline at a slower rate, stabilizing at around 0.575 to 0.595 from step 25,000 onward. No significant divergence or upward trend was observed, suggesting that the training process was stable throughout and that the learning rate of 1e-5 was appropriate for fine-tuning at this data scale.

However, the relatively narrow overall loss range (0.58 to 0.65) and the noisy oscillations throughout training indicate that the model did not converge to a well-defined minimum. This behavior is consistent with insufficient training data, where the model lacks enough examples to learn a smooth, stable mapping from text to acoustic features at approximately 71 hours of total speech data. The plateau behavior after step 25,000 further suggests that additional training steps alone would be unlikely to yield meaningful quality improvements without an increase in data volume.

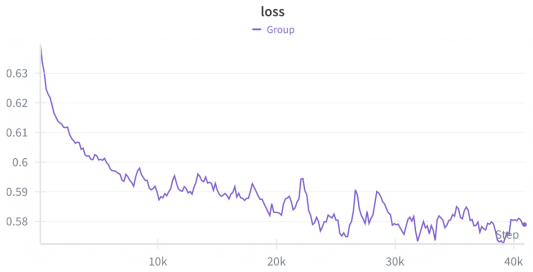

Figure 3: Time Weighted EMA Loss (0.95)

4.3 Audio Samples

Generated and reference audio samples across training checkpoints (2,000 through 40,000) are available on the project webpage.

4.4 Gradio Demonstration

To showcase the capabilities of the fine-tuned model in a practical setting, an interactive web based demonstration using Gradio was developed. The demonstration was designed to run using a Google Colab GPU for inference, while providing an accessible interface that requires no local installation or expertise from the end user.

Upon loading the application, the user is presented with three controls. A dialect selector offering Northern, Central, and Southern Vietnamese options, a text input field where arbitrary Vietnamese text can be entered or loaded from a file, and a synthesize button that triggers generation. After synthesis completes, the generated audio is playable directly in the browser alongside a mel spectrogram visualization of the output.

The demonstration is packaged as two files. A standalone Python script (`mdv_tts_app.py`) containing all application logic, and a Colab notebook (`run_mdv_tts_demo.ipynb`) that handles environment setup such as mounting Google Drive, installing the F5-TTS package, and launching the Gradio server. Upon launch, Gradio generates a

publicly shareable URL, allowing external users to interact with the model without requiring their own GPU or environment configuration. Additionally, a video demonstration is linked in the public project webpage.

4.5 Word Error Rate (WER)

Word Error Rate (WER) was the primary metric used for quantitative evaluation. WER measures the transcription accuracy of an ASR model by computing the edit distance between a reference transcript and a hypothesis transcript, expressed as a fraction of the total words in the reference (ScienceDirect). WER is computed as:

$$WER = \frac{S + D + I}{N} \quad (1)$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the reference transcript. A lower WER indicates higher intelligibility.

In testing, 50 transcriptions were pulled from the ViMD dataset's test split. With this, the model generated audio samples for each transcription, for each dialect, across checkpoints. Each of these generated samples were then passed through PhoWhisper-medium. This created transcripts that were used to compare with the ground truth reference audios. Table 2 displays the WER across the different dialects and checkpoints, where WER exceeding 100% indicates that the ASR system inserted more words than were present in the reference transcript. Although the WER values are high, a downward trend can be observed from step 20,000 to step 40,000 where the overall WER decreased from 118.28% to 106.87%.

Step	Central	North	South	Overall
20,000	118.22%	127.70%	112.80%	118.28%
30,000	110.83%	116.74%	108.33%	111.38%
35,000	117.97%	121.75%	106.61%	114.22%
40,000	104.06%	110.22%	106.65%	106.87%

Table 2: Word Error Rate (%) by dialect across training checkpoints.

4.6 Mean Opinion Score (MOS)

Mean Opinion Score (MOS) was the subjective metric used to evaluate the perceived naturalness and quality of synthesized speech (Milvus). Human listeners rate audio samples on a scale from 1 to 5, where 1 indicates poor quality and 5 indicates excellent, natural-sounding speech. Ratings are averaged across all listeners to produce a single score per condition. MOS is widely used in TTS evaluation because it captures perceptual qualities, such as prosody, fluency, and speaker naturalness, that quantitative metrics like WER cannot fully reflect.

To evaluate the perceptual quality of the fine-tuned model across training, 8 native Vietnamese speakers were asked to rate synthesized audio samples generated at six checkpoint steps. Each listener was presented with audio clips of randomized dialects and rated each sample on the standard 1 to 5 MOS scale. The average scores across all listeners are reported in Table 3.

As training checkpoints progressed, listeners agreed that the intelligibility of generated speech matched closer with the reference audio.

Checkpoint Step	Average MOS
2,000	1.2
10,000	1.9
20,000	3.1
30,000	3.7
35,000	4.3
40,000	4.7

Table 3: Average Mean Opinion Scores (MOS) from 8 native Vietnamese speakers across training checkpoints.

5 Challenges

A number of significant challenges were encountered when attempting to create this model. The most significant challenge were the data constraints. Initially, in the first iteration, instead of processing data through the pipeline described in Section 3.2.1, data was only extracted organically within our training window size. This resulted in about 20 hours of data while creating in a model that generated gibberish. As seen in the previous sections, this issue was successfully mitigated by extracting more data.

6 Discussion

6.1 Replicability

This project is fully replicable given sufficient computational resources. The code is open source and runs in a standard Python environment. However, training is compute-intensive. This project utilized an NVIDIA RTX PRO 6000 Blackwell Server Edition GPU and took approximately 17 hours to train.

6.2 Datasets

The ViMD dataset (Dinh et al., 2024) consists of labeled Vietnamese audio files with corresponding transcripts, speaker metadata, and regional dialect annotations spanning Northern, Central, and Southern varieties. Vietnamese is an underrepresented language in Natural Language Processing (NLP) and speech research, unlike high-resource languages such as English and Mandarin, there is far less publicly available speech data for Vietnamese. Dialect-annotated corpora are particularly scarce. The availability of ViMD enables a range of speech-related tasks beyond TTS, including automatic speech recognition, speaker identification and modeling, dialect classification, and multilingual speech understanding. The dataset's labeled structure makes it especially valuable for supervised learning tasks that require fine-grained linguistic distinctions. This could influence others to build more inclusive speech technologies or expand research beyond English. However, any biases in this project's annotations, like dialect coverage, could also shape what kinds of models others choose to develop or avoid.

6.3 Ethics

The ViMD dataset (Dinh et al., 2024) was collected from consenting participants, ensuring ethical sourcing of the underlying speech data. However, the system developed in this project carries risks associated with voice synthesis technology. A sufficiently capable TTS model can be used to generate speech that mimics real individuals without their consent, enabling voice-based deepfakes that could be used for fraud, impersonation, or spread of misinformation. As a result, the system was not deployed with open-ended voice cloning capabilities. Instead, synthesized output is restricted to three predetermined dialect voices, preventing users from supplying reference audio to clone a specific individual's voice.

6.4 Limitations

One limitation was the constrained training data, where only a subset of the ViMD dataset (total 71 hours) was used, which may limit model generalization. Standard TTS models typically require 50 to 100+ hours minimum to achieve intelligible output, and multi-dialect models need more hours to account for dialect-specific differences.

Another limitation is the transcript quality. Since 78.38% of the training samples were transcribed by the Silero VAD and PhoWhisper pipeline, a portion of the training data likely contains transcription errors. Transcription errors introduce noise into the text-audio alignment that the F5-TTS model learns from, which potentially degrades synthesis intelligibility and consistency.

From qualitative analysis, a significant portion of the data was found to be quite noisy. There was often noticeable background noise or poor microphone quality, which considerably affected overall data quality. This limitation may have also been a contributing reason as to why the WER values were high.

The dataset also demonstrates a gender imbalance, with a disproportionate number of samples originating from male speakers. This imbalance can bias the F5-TTS model toward dominant vocal characteristics, such as pitch range and speaking style, limiting the model's ability to generalize across diverse speaker profiles. Consequently, the synthesized speech may lack variability and perform less effectively when generating voices outside the majority distribution.

6.5 Future Work

One critical next step for this project is to improve data quality. The current training pipeline relied on PhoWhisper-medium for automated transcription of 78.38% of samples, introducing transcription noise that likely degraded text-audio alignment. Future work should incorporate manual transcript verification or a higher-accuracy ASR model to clean these transcripts. Additionally, applying speech enhancement pre-processing to remove background noise from raw ViMD speech recordings would further improve the quality of text-audio pairs available for training.

Another direction is dataset expansion. Future work could incorporate additional publicly available Vietnamese speech sources to address the underrepresentation of Central Vietnamese in the current dataset. A more balanced dialect distribution across a larger total data volume would give the model a stronger foundation for learning dialect-specific acoustic features without converging toward the dominant Northern dialect.

Further evaluation work includes extending the WER pipeline with a dedicated Vietnamese dialect classifier to verify whether synthesized speech is recognized as the correct target dialect, not just whether it is intelligible. Cross-validating WER

results against a secondary ASR model would also help isolate how much of the elevated WER reflects ASR bias toward Northern Vietnamese phonology versus genuine synthesis degradation. A formal MOS study with a larger and more diverse listener pool would additionally strengthen the perceptual evaluation.

References

ax Inc. 2023. SileroVAD: Machine learning model to detect speech segments. https://medium.com/axinc-ai/silerovad-machine-learning-model-to-detect-speech-segments-e99722c0dd41. Accessed: 2026-05-04.

Nhat Bui, Giang Nguyen, Nguyen Nguyen, Bao Vo, Luan Vo, Tom Huynh, Arthur Tang, Van Nhiem Tran, Tuyen Huynh, Huy Quang Nguyen, and Minh Dinh. 2025. Fine-tuning large language models for improved health communication in low-resource languages. *Computer Methods and Programs in Biomedicine*, 263:108655.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.

Nguyen Van Dinh, Thanh Chi Dang, Luan Thanh Nguyen, and Kiet Van Nguyen. 2024. Multi-dialect Vietnamese: Task, dataset, baseline models and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7498, Miami, Florida, USA. Association for Computational Linguistics.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, Yanqing Liu, Sheng Zhao, and Naoyuki Kanda. 2024. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. *arXiv preprint arXiv:2406.18009*.

Milvus. How is mean opinion score (MOS) used in TTS evaluation. https://milvus.io/ai-quick-reference/how-is-mean-opinion-score-mos-used-in-tts-evaluation. Accessed: 2026-05-05.

Ben Pham and Sharynne McLeod. 2016. Consonants, vowels and tones across vietnamese dialects. *International Journal of Speech-Language Pathology*, 18(2):122–134.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

VinAI Research. 2023. PhoWhisper-medium. https://huggingface.co/vinai/PhoWhisper-medium. Accessed: 2026-05-04.

ScienceDirect. Word error rate. https://www.sciencedirect.com/topics/computer-science/word-error-rate. Accessed: 2026-05-05.

Vinpearl. 2023. Central vietnamese accent. https://vinpearl.com/en/central-vietnamese-accent. Accessed: 2026-03-02.