ViettelRoar: Voice conversion approach for VLSP 2025

Hong Nhat Tran, Tung Luong Nguyen , Thanh Hai Tran
ViettelAI, Viettel Group
Hanoi, Vietnam
{nhatth2, luongnt29, haitt43}@viettel.com.vn

Abstract

In this paper, we propose a two-stage cross-lingual voice conversion system for Vietnamese combining an Automatic Speech Recognition (ASR) module with a zero-shot Text-to-Speech (TTS) module. The ASR leverages ChunkFormer for efficient long-form transcription, while the F5-TTS synthesizes natural, intelligible speech and preserves the target speaker's timbre and prosody using phoneme-level representations. Trained on the combination of ViVoice and VCTK corpora, our system achieves high-quality Vietnamese synthesis and strong cross-lingual generalization. Experimental results on the VLSP 2025 Voice Conversion evaluation dataset show a MOS of 3.53 and the highest SMOS between reference and output audio of 3.66.

1 Introduction

Voice conversion (VC) (Sisman et al., 2020) aims to transform the voice characteristics of a source speaker into those of a target speaker while preserving the original linguistic content (Zhang et al., 2020). In recent years, VC has attracted significant attention due to its wide range of applications, such as personalized virtual assistants, dubbing in multimedia, language learning, and speech rehabilitation (Casanova et al., 2023). Among the various VC techniques, cross-lingual zero-shot voice conversion is particularly challenging, where the system must generate speech in a language different from the input while adapting to an unseen target speaker using only a short reference audio sample.

Traditional VC approaches often rely on paired training data or speaker-dependent models, which limits their scalability and applicability in real-world scenarios. To overcome these challenges, modern systems typically adopt a two-stage pipeline combining Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) modules (Sun and Nagamatsu, 2020). By explicitly separating the extraction of linguistic content from the synthesis of the target voice, this design improves flexibility, naturalness, and cross-lingual robustness (Morioka et al., 2022). For example, in multilingual conversational agents, a zero-shot VC system can accurately transcribe a Vietnamese utterance using ASR and generate high-quality English speech in the target speaker's voice using TTS, even if that speaker was unseen during training.

However, developing such systems poses several challenges. First, long-form speech transcription requires ASR models that are both memory-efficient and highly accurate. Second, synthesizing natural-sounding speech for unseen speakers across multiple languages demands TTS models capable of zero-shot voice conversion. Finally, achieving cross-lingual consistency between input content and generated speech requires robust phoneme-level representations to handle language-specific pronunciation rules.

In this paper, we propose a two-stage cross-lingual zero-shot voice conversion framework. The ASR module leverages ChunkFormer (Le et al., 2025), a transformer-based model optimized for efficient long-form transcription with masked batching and chunk-wise processing. For speech synthesis, we adopt a F5-TTS (Chen et al., 2025) model capable of generating high-fidelity, multilingual, and speaker-consistent speech from short reference samples. To enhance cross-lingual performance, we employ a training strategy on multilingual corpora, including the ViVoice (Phuoc et al., 2024) dataset (1000 hours of Vietnamese speech) and the VCTK (Veaux et al., 2017) English dataset, combined with phoneme-based inputs to improve pronunciation accuracy and speaker similarity.

Experimental results on the VLSP 2025 Voice Conversion evaluation dataset demonstrate the effectiveness of our approach. Our system achieved a MOS of 3.53 and the highest speaker similarity between reference and converted audio (3.66), rank-

ing second overall among all participating teams. These results highlight the potential of our method for real-world cross-lingual zero-shot voice conversion applications.

2 Methodology

Our proposed system adopts a two-stage pipeline that combines an Automatic Speech Recognition (ASR) module with a Zero-shot Text-to-Speech (TTS) module. This design allows us to explicitly separate the extraction of linguistic content from the generation of target speech, which improves flexibility, cross-lingual robustness, and naturalness in the converted audio.

2.1 ASR-TTS Pipeline

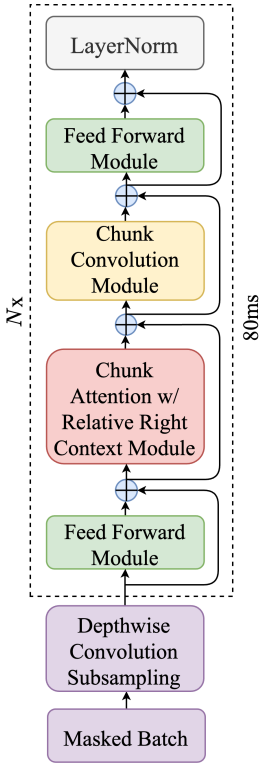

Figure 1: Chunkformer (Le et al., 2025) block repeated N times. Each block operates on an 80 ms chunk with a relative *right* context and residual connections (circled $\oplus$).

ASR Stage We adopt ChunkFormer (Le et al., 2025), a transformer-based ASR model optimized for efficient long-form speech transcription on low-memory GPUs. ChunkFormer processes audio using a chunk-wise mechanism with relative right context, allowing it to handle recordings up to 16 hours while maintaining accuracy comparable to existing models. To improve efficiency, it employs a Masked Batching technique that eliminates unnecessary padding, significantly reducing memory usage when decoding batches with variable-length audio. These characteristics make ChunkFormer scalable, robust, and suitable for both streaming and non-streaming ASR scenarios.

In our pipeline, the ASR model is responsible for converting source speech into accurate transcriptions.

TTS Stage For the speech synthesis stage, we employ a F5-TTS model. F5-TTS is a high-fidelity, multi-speaker, and multilingual text-to-speech system designed for zero-shot voice conversion.

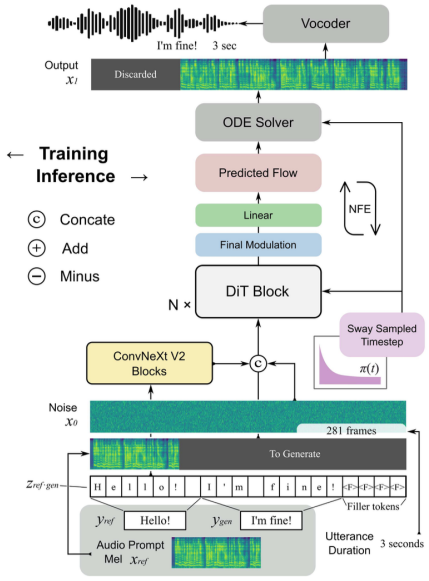

Figure 2: F5-TTS (Chen et al., 2025): flow-matching DiT conditioned on text and a reference mel. At inference, an ODE solver transforms noise x_0 into a mel x_1 using a timestep policy $\pi(t)$ with a chosen number of function evaluations (NFE); a vocoder produces the waveform.

Given a short reference audio sample from the target speaker, the model is capable of synthesizing speech that matches the target speaker's timbre, prosody, and speaking style, even if the speaker was not seen during training.

Moreover, by operating at the phoneme level, F5-TTS can synthesize natural and intelligible speech even when the linguistic input and target speaker are from different languages.

At test time, the inputs are a source utterance $\mathbf{s}_{\text{src}}$ and an optional target-speaker reference $\mathbf{s}_{\text{ref}}$. We extract mels $X_{\text{src}} = \text{Mel}(\mathbf{s}_{\text{src}})$ and $x_{\text{ref}} = \text{Mel}(\mathbf{s}_{\text{ref}})$. Chunkformer (Fig. 1) transcribes the

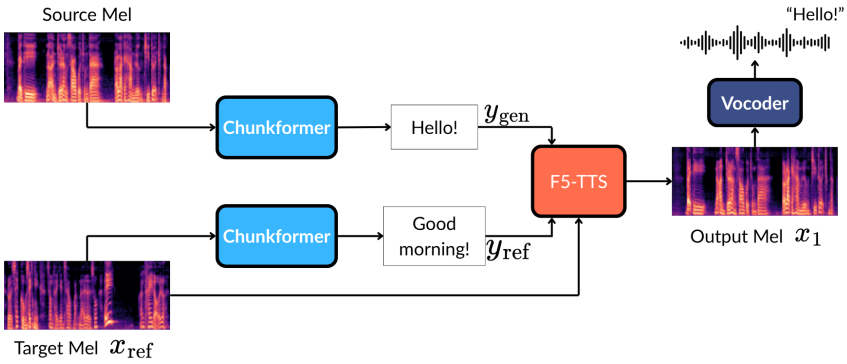

Figure 3: ASR–TTS pipeline. Chunkformer (Le et al., 2025) transcribes both the source mel (to get y_{gen}) and the reference mel (to get y_{ref}). F5-TTS (Chen et al., 2025) conditions on $\{y_{\text{gen}}, y_{\text{ref}}, x_{\text{ref}}\}$ and generates a mel x_1, which the vocoder converts to the final audio.

source by processing X_{src} in chunks of C frames with a *relative right* context of R frames: for chunk i,

$$X^{(i)} = X_{\text{src}}[:, iC : (i+1)C + R],$$

which is passed through N stacked blocks; only the first C frames of the output are kept and concatenated across i before decoding with beam search to produce the content transcript y_{gen}. Running the same ASR on x_{ref} yields a reference transcript y_{ref} that captures speaker prosody and phrasing. F5-TTS (Fig. 2) conditions on $c = \{y_{\text{gen}}, y_{\text{ref}}, x_{\text{ref}}\}$ and samples $x_0 \sim \mathcal{N}(0, I)$. The model predicts a velocity field $v_\theta(x_t, c, t)$ and an ODE solver with the sway timestep policy $\pi(t)$ and a chosen number of function evaluations (NFE) integrates

$$\frac{dx_t}{dt} = v_\theta(x_t, c, t), \quad t \in [0, 1], \quad x_{t=0} = x_0,$$

to obtain a mel x_1. A short warm-up prefix is discarded (the *Discarded* part in Fig. 2), and ⟨F⟩ filler tokens align phoneme length with target frames. Finally, a neural vocoder $g(\cdot)$ converts the mel to waveform $\hat{\mathrm{s}} = g(x_1)$. The end-to-end flow and notations $(y_{\text{gen}}, y_{\text{ref}}, x_{\text{ref}}, x_0, x_1, \pi(t), \text{NFE})$ match Fig. 3.

2.2 Cross-lingual training strategy

To develop a robust zero-shot text-to-speech (TTS) system for cross-lingual voice conversion, we adopt a carefully designed training strategy that combines multilingual data with phoneme-based representations in the synthesis stage.

Phoneme-level representations A major difficulty in diffusion-based TTS models such as F5-TTS lies in duration estimation when using character-level inputs. Since a single character can map to multiple phonetic realizations (e.g., the English letter "x" may be pronounced as /ks/ or /gz/), the alignment between characters and acoustic frames becomes ambiguous. As a workaround, F5-TTS estimates durations by taking the ratio of character lengths between y_{gen} and y_{ref}, and pads with filler tokens when the total length falls short of the mel sequence. While simple, this heuristic highlights the instability of character-level duration modeling.

To address this issue, we use phoneme-level inputs. Phonemes provide a one-to-one mapping between symbolic units and acoustic segments, removing ambiguity in duration prediction and yielding more consistent alignments. This also improves prosodic control, which is especially important for tonal languages such as Vietnamese, where vowel length and tone are directly tied to duration and must be modeled explicitly.

In practice, we use vPhon (Kirby, 2008) to convert Vietnamese text into IPA-based phoneme sequences with explicit tone marking, preserving both segmental and suprasegmental distinctions. English text, by contrast, is left at the grapheme level. To prevent symbol collisions, the two vocabularies are assigned disjoint index ranges but share a unified embedding space to support cross-lingual training. For example, a Vietnamese greeting is tokenized into a phoneme sequence with tone-marked vowels, while the English word *hello* is represented as [h, e, l, l, o]. At the encoder, a routing mechanism sends Vietnamese phoneme tokens through the extended phoneme encoder and English grapheme tokens through the base encoder. This design ensures tonal precision for Vietnamese while maintaining compatibility with English in a

single unified framework.

Training The model is trained in a single stage using a combined dataset to achieve both high-quality Vietnamese synthesis and strong cross-lingual generalization. We merge the ViVoice (Phuoc et al., 2024) corpus, which contains approximately 1,000 hours of diverse, multi-speaker Vietnamese speech, with the VCTK (Veaux et al., 2017) corpus, a widely adopted English multi-speaker dataset. This joint training strategy allows the model to simultaneously learn rich prosodic patterns and speaker variability in Vietnamese while also gaining the ability to handle English phonetic structures.

3 Experiment

3.1 Settings

Datasets We used the official evaluation dataset from the VLSP 2025 Voice Conversion shared task to assess system performance in both intra-lingual and cross-lingual settings. For training, we employed two multilingual speech corpora: ViVoice (Phuoc et al., 2024), a 1000-hour Vietnamese dataset covering diverse speakers and accents with rich phonetic coverage; and VCTK (Veaux et al., 2017), an English multi-speaker dataset with 109 speakers and a range of accents, widely used for cross-lingual voice conversion research.

Model and Parameters We adopt the original F5-TTS architecture without structural modifications. Training and inference use: classifier-free guidance (CFG) strength is set to 2, the number of function evaluations (nfe) is 32, the Euler ODE solver is used, and sway sampling is fixed at –1. Our ASR model employs the large Chunkformer configuration, consisting of 17 encoder layers with 8 attention heads and 512 hidden dimensions.

Hardware All experiments were conducted on a single NVIDIA A100 GPU (40GB) with mixed-precision (fp16) training enabled to optimize memory and runtime.

Evaluation metrics The VLSP 2025 Voice Conversion task specifies three criteria: Mean Opinion Score (MOS) for naturalness on a 1–5 scale (higher is better), Speaker Similarity (SMOS) with respect to target and source ($\mathrm{SMOS}_{\mathrm{TGT}} \equiv \mathrm{SMOS}_{\mathrm{ref,out}}$, $\mathrm{SMOS}_{\mathrm{SRC}} \equiv \mathrm{SMOS}_{\mathrm{src,out}}$), and Word Error Rate (WER) for content accuracy on a 0–100 scale (lower is better), computed with a pretrained ASR model (ChunkFormer). The official scoring is defined as

$$\begin{aligned}\mathrm{Score} = {} & 0.4\,(\mathrm{SMOS}_{\mathrm{TGT}} - \mathrm{SMOS}_{\mathrm{SRC}}) \\ & + 0.3\,\mathrm{MOS} + 0.3\,(100 - \mathrm{WER}).\end{aligned} \tag{1}$$

3.2 Main Results

Table 1 summarizes the official results for the VLSP 2025 Voice Conversion Task 1. Our system, ViettelRoar, achieved a strong overall performance with a Final Score of 67.53, ranking second among all participating teams. It obtained the best $\mathrm{SMOS}_{\mathrm{TGT}}$ (3.66 ± 0.21), showing strong target speaker preservation.

Overall, these results indicate that our method provides a good balance of naturalness, intelligibility, and, most importantly, strong target speaker similarity. The approach is particularly effective for zero-shot voice conversion, where preserving the identity of the target speaker is critical.

3.3 Discussion

Our experiments show that phoneme-level inputs improve synthesis quality for Vietnamese, particularly in capturing tonal contrasts, validating that the one-to-one mapping between phonemes and acoustic units yields more stable alignments and prosodic control. Joint training with English data enables cross-lingual generalization, while merging ViVoice and VCTK brings diverse speaker styles and broader phonetic coverage. The integration of phoneme and grapheme vocabularies into a shared embedding space allows the model to leverage commonalities across languages without symbol collisions. Compared to character-based baselines such as VietVoice, our system produces speech that listeners rate as more natural and intelligible.

Despite these gains, several limitations remain. Vietnamese tonal and dialectal variation are not explicitly modeled, and the system instead relies on reference audio to provide prosodic and regional cues. This dependence can lead to mismatches when tones are under-realized or dialects diverge. In addition, the asymmetry between Vietnamese phonemes and English graphemes may restrict cross-lingual alignment. Finally, the known weaknesses of diffusion-based TTS models, such as repetition and hallucination on long text inputs, persist in our system.

Team	MOS ↑	SMOS_TGT ↑	SMOS_SRC ↓	WER ↓	Final Score ↑
3N	$\mathbf{4.29 \pm 0.16}$	3.65 ± 0.23	1.17 ± 0.09	**9.83**	**72.66**
VCL	3.72 ± 0.17	3.21 ± 0.20	1.27 ± 0.11	10.98	64.49
ProfessorAgasa	3.29 ± 0.22	3.18 ± 0.21	$\mathbf{1.11 \pm 0.07}$	12.84	62.40
ViettelRoar (ours)	3.53 ± 0.21	$\mathbf{3.66 \pm 0.21}$	1.13 ± 0.08	12.95	67.53

soic

Table 1: Evaluation results on VLSP 2025 Voice Conversion Task 1.

4 Conclusion

We introduced ASR-TTS, a two-stage cross-lingual voice conversion framework designed for Vietnamese. The system separates linguistic content extraction and speech synthesis by combining Chunkformer-based ASR for accurate long-form transcription with a F5-TTS for high-fidelity, zero-shot speech generation using phoneme-level representations. Through training on the ViVoice and VCTK corpora, the approach achieves natural Vietnamese synthesis while generalizing effectively to cross-lingual scenarios. Evaluation on the VLSP 2025 Voice Conversion benchmark demonstrates competitive overall performance, highlighted by strong target speaker similarity and balanced naturalness and intelligibility, showing the effectiveness of the proposed design.

References

Edresson Casanova, Christopher Shulby, Alexander Korolev, Arnaldo Candido Junior, Anderson da Silva Soares, Sandra Aluísio, and Moacir Antonelli Ponti. 2023. Asr data augmentation in low-resource settings using cross-lingual multi-speaker tts and cross-lingual voice conversion. In *Interspeech 2023*, pages 1244–1248.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271, Vienna, Austria. Association for Computational Linguistics.

James Kirby. 2008. vphon: a vietnamese phonetizer. Version 2.1.1. Retrieved on October 3, 2025 from `http://github.com/kirbyj/vPhon/`.

Khanh Le, Tuan Vu Ho, Dung Tran, and Duc Thanh Chau. 2025. Chunkformer: Masked chunking conformer for long-form speech transcription. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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*.

Gia Thinh Le Phuoc, Minh Tuan Pham, Quoc Hung Nguyen, Quoc Trung Nguyen, and Hoang Vinh Truong. 2024. vivoice: Enabling vietnamese multi-speaker speech synthesis.

Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. 2020. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:132–157.

Qinghua Sun and Kenji Nagamatsu. 2020. Building Multilingual TTS using Cross-Lingual Voice Conversion. *arXiv preprint arXiv:2012.14039*.

Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, and 1 others. 2017. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 6:15.

Jing-Xuan Zhang, Li-Juan Liu, Yan-Nian Chen, Ya-Jun Hu, Yuan Jiang, Zhen-Hua Ling, and Li-Rong Dai. 2020. Voice conversion by cascading automatic speech recognition and text-to-speech synthesis with prosody transfer. *arXiv preprint arXiv:2009.01475*.