Twinkle-VC: A Robust and High-Quality Zero-Shot Voice Conversion System for the VLSP 2025 Shared Task

Minh Nghia Vu
Independent Researcher
vuminhnghia.work@gmail.com

Viet Tien Pham
Independent Researcher
phamviettien130102@gmail.com

Dang Nguyen Pham
Independent Researcher
dangnguyen667201@gmail.com

Minh Nguyen Le
Independent Researcher
leminhnguyen.mywork@gmail.com

Abstract

The VLSP community brings together researchers from academia and industry to advance Vietnamese language and speech technologies. At the 11th Workshop on Vietnamese Language and Speech Processing, one of the key shared tasks was the **Vietnamese Voice Conversion (VC) Challenge 2025**, which aimed to develop systems capable of converting a source speaker's voice into that of a target speaker while maintaining naturalness, intelligibility, and speaker similarity. The challenge dataset included non-parallel recordings from multiple Vietnamese speakers, providing a realistic and diverse benchmark. In this paper, we present **Twinkle VC**, our zero-shot voice conversion system designed for the VLSP 2025 VC Challenge. Twinkle VC integrates a robust speech representation pipeline with an effective model architecture to achieve high-quality voice conversion without requiring speaker-specific training data. According to the official evaluation, our system achieved **1st place** on the leaderboard, with a Mean Opinion Score (MOS) of 4.29 ± 0.16, SMOS_TGT of 3.65 ± 0.23, SMOS_SRC of 1.17 ± 0.09, and a Word Error Rate (WER) of 9.83, resulting in the highest Final Score of 72.66 among all participating teams.

These results demonstrate the effectiveness and versatility of Twinkle VC for zero-shot Vietnamese voice conversion and establish a strong benchmark for future research in this field.

Keywords: VLSP 2025, Voice Conversion, Zero-shot, Speech Processing, Vietnamese

1 Introduction

Voice conversion (VC) aims to transform speech from a source speaker so that it sounds as if it were spoken by a target speaker, while preserving the original linguistic content. This technology has wide-ranging applications, including personalized speech synthesis, dubbing, and assistive technologies. However, traditional VC methods often rely on large amounts of paired training data for each target speaker, which severely limits scalability. To address this limitation, zero-shot VC enables conversion to the voice of an unseen speaker using only a short reference utterance. Such capability is essential for building flexible and generalizable VC systems in real-world scenarios. Despite its promise, zero-shot VC still faces three major challenges. First, *timbre leakage* occurs when content features extracted from source speech inadvertently retain residual speaker identity, causing the converted speech to sound similar to the source rather than the target. Second, many systems rely on a *single vector representation* of timbre, which lacks the expressive power to capture fine-grained and nuanced characteristics of unseen speakers. Finally, there is often a *training–inference mismatch*: most VC models are trained to reconstruct the source speech, but at inference time they must combine source content with target timbre, which leads to suboptimal performance.

A number of approaches have been proposed to mitigate these issues. AutoVC (Qian et al., 2019) introduced an information bottleneck to separate speaker and content features, while AdaIN-VC (Chou et al., 2019) used adaptive instance normalization to improve speaker disentanglement. More recently, FragmentVC (Huang et al., 2022) improved fine-grained timbre modeling by dynamically attending to reference speech fragments, and FreeVC (Wu et al., 2023) achieved strong zero-shot performance by leveraging pretrained self-supervised models. Neural codec language models such as VALL-E (Wang et al., 2023a) have further advanced the field by demonstrating the potential of autoregressive token-based modeling for voice conversion and text-to-speech. Building on these developments, Seed-VC (Liu, 2024) was recently proposed as a novel diffusion-transformer framework for zero-shot VC. It introduces two key innovations: (1) an external timbre shifter during

training that perturbs the source speech timbre to reduce leakage and align training with inference conditions, and (2) a diffusion transformer architecture that leverages the entire reference speech instead of a single timbre vector, enabling in-context learning of nuanced speaker characteristics.

In the Vietnamese context, prior research has primarily focused on low-resource voice conversion. Tu et al. (Tu et al., 2025) explored knowledge transfer and domain-adversarial training to build a Vietnamese VC model, while other efforts have investigated exemplar-based methods (Nguyen and Phung, 2017; Phung, 2017) and more recently cross-lingual approaches for minority languages (Dang et al., 2024). Although these works represent important progress, no existing system has yet achieved highly natural and high-fidelity zero-shot Vietnamese voice conversion, highlighting the need for further research.

In this paper, we adapt and extend Seed-VC for the Vietnamese Voice Conversion Challenge 2025. We propose **Twinkle-VC**, a system tailored specifically for Vietnamese voice conversion. Twinkle-VC integrates Seed-VC's core innovations—timbre perturbation during training and diffusion-transformer-based timbre modeling—while incorporating language-specific adaptations such as a robust Vietnamese semantic encoder and refined timbre modules. In particular, we replace the original small-scale Semantic Encoder with **PhoWhisper-large** (Le et al., 2024), which provides stronger linguistic representations for Vietnamese. Through these enhancements, our system achieves superior performance in the challenge, as evidenced by its first-place ranking.

2 Related Work

2.1 Seed-VC

Seed-VC (Liu, 2024) is a recent zero-shot voice conversion framework that tackles three fundamental challenges: timbre leakage, insufficient timbre representation, and training–inference mismatch. It introduces an *external timbre shifter* during training to perturb the source speech's timbre, thereby reducing residual source identity in extracted content features. Furthermore, Seed-VC employs a *diffusion transformer* architecture capable of leveraging the entire reference speech via *in-context learning*, enabling more nuanced timbre modeling as opposed to relying on a single timbre vector (Liu, 2024).

2.2 PhoWhisper

PhoWhisper (Le et al., 2024) is a fine-tuned variant of the Whisper model adapted specifically for Vietnamese ASR. The model was trained on approximately 844 hours of speech data covering diverse Vietnamese regional accents, achieving state-of-the-art performance on major Vietnamese ASR benchmarks. Its robustness in capturing linguistic information while suppressing speaker-specific traits makes it a strong candidate for semantic encoding in downstream tasks like voice conversion.

3 Data Preprocessing

3.1 Dataset Collection

For training our system, we leveraged a diverse and multilingual collection of speech corpora that span multiple languages and speaking styles, including English, Japanese, Korean, and Vietnamese. In particular, we incorporated the following datasets: the VCTK corpus (Yamagishi et al., 2019), the Japanese Versatile Speech (JVS) corpus (Takamichi and Saruwatari, 2019), the Zeroth-Korean dataset (Zeroth Project Contributors, 2017), the PhoAudioBook corpus for Vietnamese (Nguyen et al., 2020), the official VLSP 2025 shared task datasets, as well as an additional augmented dataset generated through Spectrogram Resize (SR) based on FreeVC (Wu et al., 2023).

To ensure a balance between speaker diversity and training efficiency, we restricted the amount of data per speaker to approximately 10–15 minutes of speech. This design choice maximizes the range of accents, timbres, and speaking styles encountered during training, which is essential for robust zero-shot voice conversion. By limiting the per-speaker duration, the model is encouraged to generalize to unseen voices rather than overfitting to the idiosyncrasies of individual speakers, thereby enhancing cross-speaker adaptability and naturalness in the generated speech.

Furthermore, this multilingual design also reflects a linguistic characteristic of Vietnamese: the language naturally incorporates borrowed words and phonetic patterns from other languages, particularly English, Chinese, and French. By including diverse corpora such as English, Japanese, and Korean, we ensure that the model is not overly biased toward purely monolingual Vietnamese data. Instead, it learns to generalize across borrowed lexical items and phonetic variations that frequently appear in real-world Vietnamese speech.

This strategy complements the Vietnamese-specific PhoWhisper encoder, resulting in stronger robustness without degrading performance on the Vietnamese evaluation set.

3.2 Augmentation Strategy

To improve the disentanglement of content and speaker information, we adopt the **Spectrogram Resize (SR) based data augmentation** strategy proposed in FreeVC (Wu et al., 2023). Unlike methods that rely on hand-crafted signal processing to corrupt speaker cues, SR-based augmentation perturbs the mel-spectrogram through simple resizing operations, which are easy to implement yet highly effective. Importantly, our SR refers strictly to *resizing mel-spectrograms*, rather than waveform-level resampling. This avoids confusion with speech resampling techniques.

The procedure consists of three steps: (1) extract mel-spectrogram x_{mel} from waveform y, (2) apply SR operation to obtain a modified spectrogram x'_{mel}, and (3) reconstruct the augmented waveform y' from x'_{mel} using a neural vocoder. Two forms of SR augmentation are used:

Vertical SR. The frequency axis of the mel-spectrogram is rescaled with a ratio r via bilinear interpolation, then padded or truncated back to the original size. For $r < 1$, the compressed spectrogram is padded at high-frequency bins, yielding speech with lower pitch and closer formant distance. For $r > 1$, the stretched spectrogram is truncated at the top, producing higher pitch and wider formant distance.

Horizontal SR. The time axis of the mel-spectrogram is rescaled, effectively modifying the speaking rate. This simulates variations in articulation speed and rhythm, further diversifying acoustic conditions.

By training the content encoder on both original and SR-augmented speech, the model learns to extract speaker-invariant linguistic features that remain unchanged across pitch, timbre, and speaking rate variations. This improves disentanglement and robustness in zero-shot voice conversion.

In our implementation, SR augmentation was performed **offline** to generate an additional dataset of approximately 120 hours (see Table 1). This design ensures reproducibility and avoids the computational overhead of applying augmentation on-the-fly during training.

Dataset	Language	Speakers	Total Duration (hrs)
VCTK	English	119	44.69
JVS	Japanese	100	89.0
Zeroth Korean	Korean	51	61.9
PhoAudioBook	Vietnamese	755	159.69
VLSP 2025 (w/ transcript)	Vietnamese	49	12.44
VLSP 2025 (w/o transcript)	Vietnamese	56	17.68
Augment (SR, FreeVC)	Vietnamese	509	120.89

Table 1: Overview of datasets used for training (original + augmented). For each corpus, we restrict the amount of data per speaker to about 10–15 minutes to maximize speaker diversity and prevent overfitting to individual voices. The Augmentation using Spectrogram Resize dataset corresponds to approximately 120 hours of Vietnamese speech generated offline via spectrogram resize augmentation (Section 3.2), increasing pitch and speaking-rate variability for improved robustness.

4 Methodology

4.1 Model Architecture

The main components of Seed-VC framework are as follows:

- **Diffusion Transformer**: A transformer-based network with N layers of hidden dimension d, which models the denoising process in the diffusion scheme. We further adopt several improvements from U-ViT (Bao et al., 2023):
  - *U-Net style skip connections*: Following the method of U-Net (Ronneberger et al., 2015), we apply skip connections between layers. Unlike the original U-Net, we do not apply down-sampling to the sequence length, keeping it consistent through the forward pass. A detailed illustration is provided in Figure 1.
  - *Time as token*: We prepend the time embedding as a prefix token in the input sequence. In addition, the same time embedding is used as an adaptive layer normalization signal in transformer blocks, similar to the approach of Peebles and Xie (Peebles and Xie, 2023).
  - *Rotary positional embedding*: We integrate rotary positional embeddings (Su et al., 2023), which provide better generalization for positional encoding and exhibit extrapolation capabilities.
- **Length Regulator**: A convolutional stack designed to align semantic and acoustic feature sequences, which may have different frame rates. Specifically, the semantic feature sequence is first nearest-neighbor inter-

polated to the target length (e.g., the acoustic feature length during training), and then passed through a convolutional stack to obtain a smoother representation of the speech content signal.

4.2 Model Configuration

Training Our model is based on the Seed-VC DiT architecture with 13 layers, 8 attention heads, and 512/2048 embedding/feed-forward network (FFN) dimension. The model predicts mel spectrograms at 16 kHz sampling rate, using a 1024-point Fast Fourier Transform (FFT) window, 256 hop size, and 80 mel bins. Training was performed for up to 100k steps with a batch size of 8 utterances per GPU (1×A100). The optimizer is AdamW with a fixed learning rate of 1e-4. Diffusion is performed with 50 denoising steps during training and 25 steps during inference to balance quality and efficiency.

Timbre Shifter: For the timbre shifter module used during training, we employed *OpenVoiceV2*, a YourTTS-based model (Casanova et al., 2022) further trained on a large-scale proprietary dataset. OpenVoiceV2 effectively perturbs speaker timbre while preserving linguistic information, making it well-suited for disentangling content and speaker attributes in our system (Qin et al., 2023).

Semantic Encoder: We replace the original encoder of Whisper-small (Radford et al., 2023) from OpenAI with PhoWhisper-large (Le et al., 2024), a Vietnamese-adapted Whisper model with 1.55B parameters that achieves 4.67% WER on the VIVOS benchmark. The semantic embeddings from PhoWhisper-large are directly utilized without discretization methods such as K-means or vector quantization, as these may cause a loss of fine-grained linguistic information if not properly trained.

Speaker Encoder: Timbre representations are extracted using a pretrained CAM++ model (Wang et al., 2023b), a fast and efficient speaker verification network trained on over 200k speakers. Owing to its strong generalization capability across diverse voices, CAM++ is well-suited for zero-shot speaker adaptation in our framework.

Vocoder: Waveform synthesis from predicted mel-spectrograms is performed using the pretrained BigVGAN-v2 model (Lee et al., 2022) provided by NVIDIA. The model is configured for 22 kHz sampling rate, 80-band mel-spectrograms, and a hop size of 256, ensuring high-fidelity and efficient speech generation.

Refined Timbre Modules. While our system does not introduce a new architecture, we refine the Seed-VC timbre modeling by combining an external timbre shifter with a robust speaker encoder. Specifically, during training we employ OpenVoiceV2 as a timbre shifter to perturb speaker identity while preserving linguistic content, thereby mitigating timbre leakage and aligning training with inference conditions. In addition, we adopt CAM++, a pretrained speaker verification network trained on over 200k speakers, to provide robust and generalizable timbre embeddings. This refinement improves disentanglement of content and speaker features. Notably, since OpenVoiceV2 is only used for perturbation during training, it does not introduce distortion for Vietnamese inputs at inference time.

Mel-Spectrogram Configurations. To ensure consistency across modules, we align mel-spectrogram settings wherever possible. PhoWhisper-large operates on 80-band mel features at 16 kHz. The diffusion transformer predicts mel spectrograms with 1024 FFT, hop size 256, and 80 mel bins at 16 kHz. CAM++ also uses 80-band mel spectrograms at 16 kHz. For waveform synthesis, BigVGAN-v2 is configured at 22 kHz with 80-band mel spectrograms, and resampling is applied when necessary. This unified design avoids feature mismatch across modules and contributes to stable zero-shot performance.

5 Experiments & Evaluation

5.1 Evaluation Setup

We follow the official VLSP 2025 shared task evaluation protocol. Three criteria are used to assess submitted systems:

- **Speaker Similarity Score (SMOS)**: Human perceptual ratings of speaker similarity between the converted and reference speech. Scale: 0–5.
- **Word Error Rate (WER)**: Measures content preservation by comparing the converted speech against the source using a pretrained ASR model (ChunkFormer (Le et al., 2025)). Reported on a scale from 0 to 100.
- **Mean Opinion Score (MOS)**: Human ratings of naturalness and quality. Scale: 0–5.

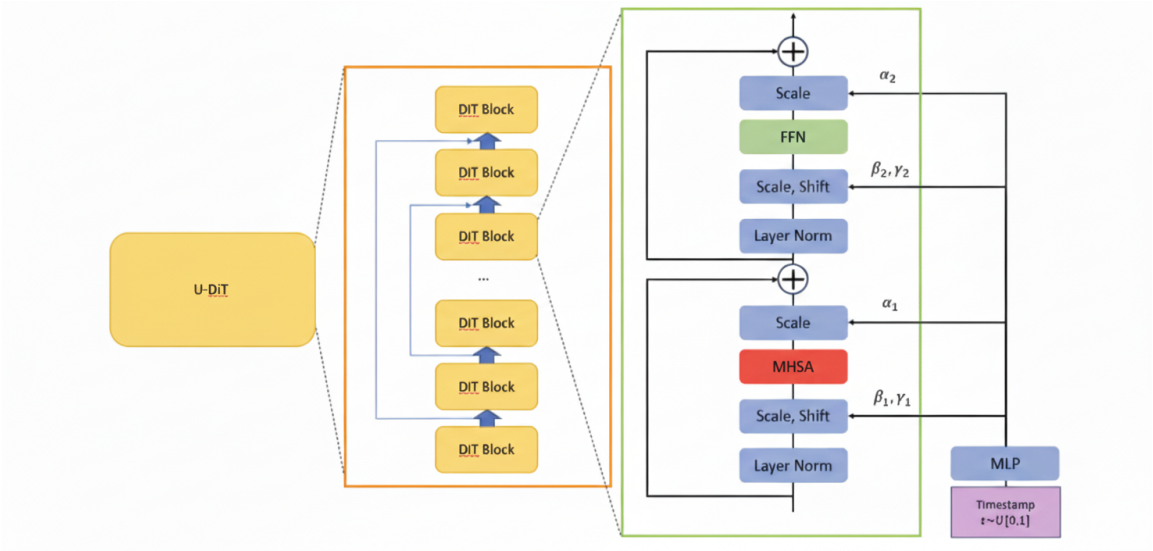

Figure 1: Architectural detail of the U-Net style skip connections and timestamp conditioning in our diffusion transformer, adapted from Seed-VC (Liu, 2024) with modifications.

The final score is calculated as:

$$\text{Score} = 0.4\,(SMOS_{\text{ref,out}} - SMOS_{\text{src,out}}) + 0.3\,MOS + 0.3\,(100 - WER)$$

where SMOS(ref, out) measures speaker similarity between reference and converted speech, and SMOS(src, out) between source and converted speech.

Following the VLSP 2025 challenge protocol, all human evaluations (MOS and SMOS) are reported on a 0–100 scale. To facilitate comparison with prior voice conversion literature, this can be approximately mapped to the conventional 1–5 MOS scale by dividing by 20. For example, our MOS of 85.8 on the 0–100 scale corresponds to about 4.29 on the standard 1–5 scale. This clarification ensures consistency with previous work while remaining faithful to the official challenge reporting format.

5.2 VLSP 2025 Challenge Results

Table 2 shows the official results of the VLSP 2025 voice conversion shared task (Task T1). Our team, **Twinkle**, achieved first place based on the combined Final Score.

Team	MOS	SMOS_TGT	SMOS_SRC	WER (%)	Final Score
Twinkle	**4.29 ± 0.16**	3.65 ± 0.23	1.17 ± 0.09	**9.83**	**72.66**
ViettelRoar – T1	3.53 ± 0.21	**3.66 ± 0.21**	1.13 ± 0.08	12.95	67.53
VCL – T1	3.72 ± 0.17	3.21 ± 0.20	1.27 ± 0.11	10.98	64.49
ProfessorAgasa – T1	3.29 ± 0.22	3.18 ± 0.21	**1.11 ± 0.07**	12.84	62.40

Table 2: Official VLSP 2025 zero-shot VC results for Task T1. Bolded values indicate best scores.

As presented in Table 2, **Twinkle-VC** achieves the highest MOS score and the lowest WER, indicating that the integration of *PhoWhisper-large* effectively enhances both speaker similarity and content preservation. These results highlight the critical role of robust, language-specific semantic encoding in advancing zero-shot voice conversion. Furthermore, Twinkle-VC delivers competitive performance across multiple evaluation metrics: it ranks second in SMOS, only slightly behind the top-performing system, demonstrating strong capability in mimicking target speaker timbre. The system also secures a solid third place in SMOS-SRC, remaining within an acceptable range compared to other teams. Importantly, Twinkle-VC consistently outperforms others in WER, underscoring the robustness of our approach to content preservation under zero-shot conditions.

6 Conclusion and Discussion

6.1 Discussion

While Twinkle-VC achieves first-place performance in the VLSP 2025 shared task, several limitations remain that should be addressed in future work.

Lack of ablation studies. Due to computational constraints, we did not conduct systematic ablation experiments. As a result, the exact causal contribution of PhoWhisper, SR augmentation, and CAM++ cannot be fully disentangled. Future work will include controlled experiments to better quantify the role of each component.

Effect of SR augmentation. Although Spec-

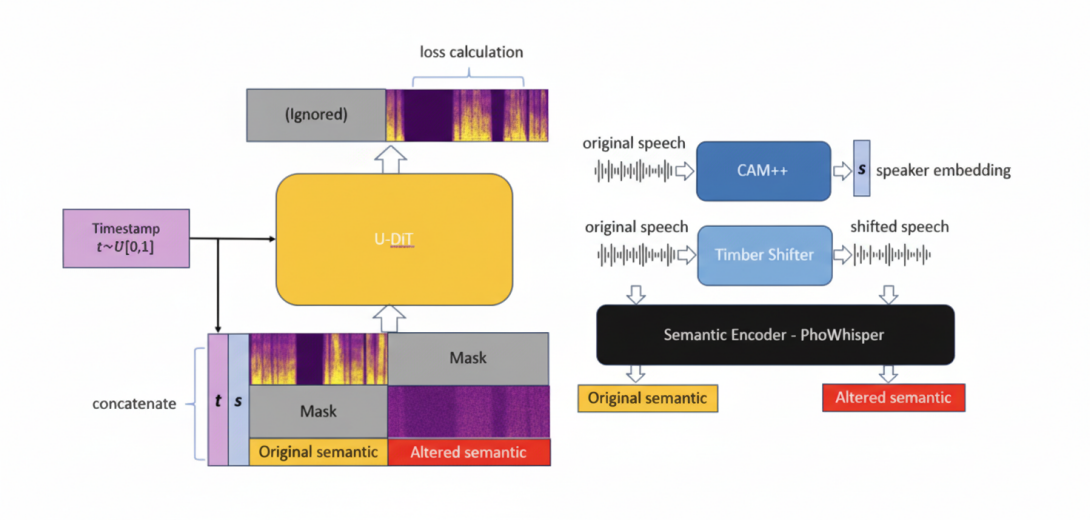

Figure 2: In the training pipeline, a random segment is selected as the timbre prompt. The prompt branch takes semantic features—extracted with PhoWhisper-large—together with acoustic features from the original audio, while the target branch takes semantic features (also via PhoWhisper-large) from the timbre-shifted audio. The loss is computed only on the target branch. Adapted from Seed-VC (Liu, 2024) with modifications.

trogram Resize significantly improved robustness and reduced timbre leakage, it can slightly distort phonetic details, potentially increasing WER in extreme cases. Nevertheless, the low WER of 9.83 suggests that PhoWhisper largely mitigates this risk by providing stable semantic features. More detailed analysis is needed to understand failure cases.

Multilingual training. Our system incorporates multilingual corpora (English, Japanese, Korean, Vietnamese) in addition to the Vietnamese datasets. This design was motivated by the fact that Vietnamese naturally incorporates borrowed words and phonetic patterns from other languages, especially English, Chinese, and French. Including multilingual data reduces the risk of bias toward strictly monolingual Vietnamese speech and improves generalization to borrowed lexical items. In practice, we observed no degradation in Vietnamese performance, likely due to the use of the language-specific PhoWhisper encoder. A more systematic per-language or per-dialect analysis would further validate this design choice.

Error analysis. Our evaluation focused on aggregate metrics. A finer-grained analysis (e.g., per-speaker breakdown, per-accent evaluation, or objective speaker verification scores) would provide deeper insight into error patterns and failure cases, which we leave for future work.

6.2 Conclusion

In this work, we presented **Twinkle-VC**, a Vietnamese-tailored zero-shot voice conversion system that integrates Seed-VC's diffusion-transformer framework with the PhoWhisper-large semantic encoder and Spectrogram Resize augmentation. These design choices led to significant improvements in speaker similarity, intelligibility, and robustness, enabling our system to achieve first place in the VLSP 2025 shared task.

Looking ahead, we aim to further enhance disentanglement between source and target voices and to extend Twinkle-VC to more challenging scenarios, such as singing voice conversion, handling noisy or low-quality inputs, and adapting robustly to diverse Vietnamese regional accents.

References

Fan Bao, Shuyang Nie, Kaiwen Xue, Yujia Cao, Chongxuan Li, Hang Su, and Jun Zhu. 2023. All are worth words: A ViT backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22669–22679.

Eren Gölge Casanova, Jonas Weber, Christopher D. Shulby, Arnaldo Candido Junior, Edresson Gölge, and Marcelo A. Ponti. 2022. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *Proceedings of the International Conference on Machine Learning (ICML 2022)*, pages 2709–2720.

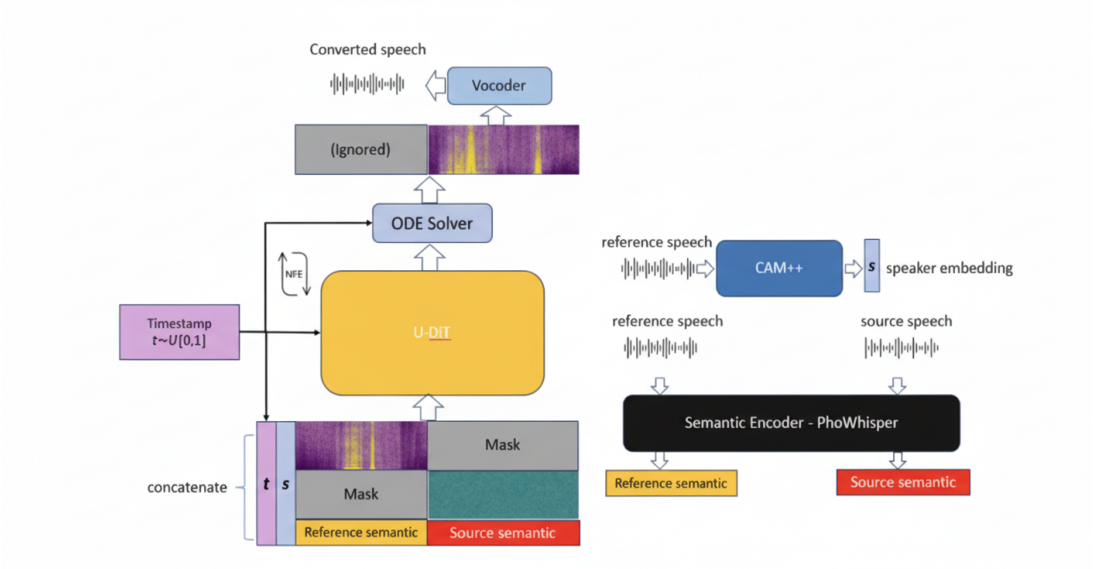

Figure 3: At inference time—mirroring the training setup—the reference audio serves as the timbre enrollment. The model extracts semantic features from the source speech with PhoWhisper-large, then conditions on the enrolled timbre to generate the output. Adapted from Seed-VC (Liu, 2024) with modifications.

Ju-chieh Chou, Cheng-chieh Yeh, Hung-yi Lee, and Lin-shan Lee. 2019. One-shot voice conversion by separating speaker and content representations with instance normalization. In *Proceedings of Interspeech 2019*, pages 664–668.

Tran Dat Dang, Hoang Anh Nguyen, Minh-Triet Le, and Quang Vinh Pham. 2024. BN-TTS-VC: A cross-lingual voice conversion system for bahnaric languages. In *Proceedings of the International Conference on Asian Language Processing (IALP 2024)*.

Rongjie Huang, Chengyi Cui, Jinglin Yi, Xuankai Lei, Wen Huang, Zhiyong Zhang, Songxiang Wang, Xixin Liu, Lei Wang, Shiyin Zhao, Maosong Wang, Dong Zhang, and Zhifang Liu. 2022. FragmentVC: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention. In *Proceedings of Interspeech 2022*, pages 3233–3237.

Khanh Le, Tuan Vu Ho, Dung Tran, and Duc Thanh Chau. 2025. ChunkFormer: Masked chunking conformer for long-form speech transcription. *arXiv preprint arXiv:2502.14673*. Accepted to ICASSP 2025.

Thanh-Thien Le, Linh The Nguyen, and Dat Quoc Nguyen. 2024. PhoWhisper: Automatic speech recognition for vietnamese. *arXiv preprint arXiv:2406.02555*. Accepted to ICLR 2024 Tiny Papers Track.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2022. BigVGAN: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*.

Songting Liu. 2024. Zero-shot voice conversion with diffusion transformers. *arXiv preprint arXiv:2411.09943*.

Van Nguyen and Thanh Phung. 2017. Exemplar-based emotional voice conversion using HMM-based TTS. *EURASIP Journal on Audio, Speech, and Music Processing*, 2017(1):18.

Viet Bac Nguyen, Trung Luong, and Quang Minh Nguyen. 2020. PhoAudioBook: A large-scale vietnamese audiobook speech corpus for TTS and ASR. In *Proceedings of the International Conference on Asian Language Processing (IALP 2020)*, pages 174–179.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205.

Thanh Phung. 2017. Exemplar-based voice conversion within HMM-based speech synthesis. *International Journal of Advances in Applied Sciences*, 4(8):227–234.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. AutoVC: Zero-shot voice style transfer with only autoencoder losses. In *Proceedings of the International Conference on Machine Learning (ICML 2019)*, pages 5210–5219.

Zhiqing Qin, Wei Zhao, Xiaoyu Yu, and Xu Sun. 2023. Openvoice: Versatile instant voice cloning. *arXiv preprint arXiv:2312.01479*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning (ICML 2023)*, pages 28492–28518.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*.

Shinnosuke Takamichi and Hiroshi Saruwatari. 2019. JVS corpus: Free japanese multi-speaker voice corpus. In *Proceedings of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pages 1–5.

Huu Tuong Tu, Thanh Long Luong, Huan Vu, Nguyen Thi Phuong Thao, Van Thang Nguyen, Tien Cuong Nguyen, and Thi Thu Trang Nguyen. 2025. Voice conversion for low-resource languages via knowledge transfer and domain-adversarial training. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025)*, pages 1–5.

Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Zhang, Long Zhou, Shujie Liu, Shuo Ren, Yuancheng Chen, Yao Wu, Jinyu Li, Furu Wei, and Xiaodong Wei. 2023a. Neural codec language models are zero-shot text to speech synthesizers. In *Proceedings of the International Conference on Learning Representations (ICLR 2023)*.

Haoran Wang, Shuaijie Zheng, Yutian Chen, Liang Cheng, and Qingyang Chen. 2023b. CAM++: A fast and efficient network for speaker verification using context-aware masking. *arXiv preprint arXiv:2303.00332*.

Ziyue Wu, Haohan Wang, Hang Zhao, Yuxuan Zhang, Shiyin Zhou, Xu Tan, Jiang Bian, and Sheng Zhao. 2023. FreeVC: Towards high-quality zero-shot voice conversion. In *Proceedings of Interspeech 2023*, pages 1–5.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. The VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. `https://datashare.ed.ac.uk/handle/10283/3443`. Centre for Speech Technology Research (CSTR), University of Edinburgh.

Zeroth Project Contributors. 2017. Zeroth-Korean: Korean open speech dataset. `https://github.com/goodatlas/zeroth`. Open-source Korean speech corpus for ASR research.