Towards Cost-Effective Voice Cloning System for Vietnamese TTS: A Case Study at HCMUT

Vinh Q. Vo[1], Bao G. Quach[1], Quyen T. Bui[1], Khai Q. Truong[1], Long S. T. Nguyen[1], Fabien Baldacci[2], and Tho T. Quan[1(✉)]

[1] URA Research Group, Ho Chi Minh City University of Technology (HCMUT), Vietnam National University Ho Chi Minh City (VNU-HCM), Ho Chi Minh City, Vietnam
{vinh.vo1205,bao.quach24hcmut,quyen.bui080405,
khai.truongquang711,long.nguyencse2023,qttho}@hcmut.edu.vn

[2] Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, 33400 Talence, France
fabien.baldacci@u-bordeaux.fr

Abstract. The increasing demand for personalized, emotionally expressive synthetic voices in Text-to-Speech (TTS) systems presents the challenge of generating voices that sound genuinely friendly, especially with limited data and computational resources. This study proposes a voice cloning method for data-scarce scenarios, requiring only 10–30 minutes of speech and optimized for low-complexity environments like Google Colab. By leveraging So-VITS-SVC, a robust voice conversion model, and the acoustic precision of Parselmouth (Praat), we extract and refine vocal attributes such as intonation, pitch, and timbre for more accurate and emotionally resonant voice generation. The method was deployed in an automated answering system at HCMUT, where it generates voices that are friendly and relatable for students. Preliminary evaluations show that the system outperforms baseline models with lower F0 RMSE (19.4 Hz), higher MOS-N (4.4), MOS-I (4.7), and reduced WER (3.2%), TER (2.7%). These results highlight the effectiveness of prosody conditioning in improving voice intelligibility and naturalness. Future work will focus on refining emotional expressiveness, improving training data, and addressing ethical concerns in voice cloning.

Keywords: Voice Cloning · Text-to-Speech · Deep Learning · Tonal Languages · Prosody Modeling · Speech Synthesis · Personalization · Praat

1 Introduction

Text-to-Speech (TTS) systems convert text into natural, intelligible speech, supporting applications such as virtual assistants, customer service, and education. Recent neural architectures (e.g., WaveNet [6], Tacotron 2 [8], FastSpeech 2 [7])

T. T. Quan et al. (Eds.): MIWAI 2025, LNAI 16355, pp. 213–225, 2026.
https://doi.org/10.1007/978-981-95-4963-4_18

have improved synthesis quality but require large corpora and heavy computation, limiting their applicability in low-resource, tonal languages such as Vietnamese, where lexical meaning depends on fine-grained pitch contours [3]. Few-shot voice-cloning methods [1] reduce data demands but lack explicit prosodic control, while classical concatenative and HMM-based approaches [10] often yield oversmoothed speech that impairs intelligibility in tone-sensitive contexts.

We propose a lightweight, prosody-aware voice-cloning framework based on So-VITS-SVC [9], which integrates explicit prosodic features–F0, intensity, and duration–extracted with Parselmouth [3] into both the variance adaptor [11] and training objectives. The framework follows a two-stage pipeline: (i) generating neutral speech with a baseline TTS, followed by (ii) prosody-aware conversion that transfers speaker identity while preserving tonal contours. It requires less than 30 min of target-speaker data and can be trained efficiently on a single GPU. Deployment at Ho Chi Minh City University of Technology (HCMUT) demonstrates improved tonal fidelity and speaker similarity in real-world conditions. In summary, our contributions are as follows.

- A lightweight, prosody-aware voice-cloning framework for tonal low-resource settings.
- Integration of explicit prosodic features (F0, intensity, duration) into both architecture and loss functions for robust tone preservation.
- Real-world deployment at HCMUT, achieving improved prosody and speaker similarity with less than 30 min of training data.

2 Related Work

Classical concatenative and formant-based TTS rely on fixed databases or rule-based parameters, offering only limited flexibility in prosody. Statistical parametric models, such as the HMM-based approach by Tokuda et al. [10], improved generalization with context-dependent modeling but often produced oversmoothed, robotic speech–particularly problematic for tonal languages.

Neural approaches have since advanced naturalness. WaveNet [6] demonstrated high-quality waveform generation but required large datasets and heavy computation. Tacotron 2 [8] simplified feature engineering with a seq2seq pipeline, yet remained data-hungry and error-prone. FastSpeech 2 [7] enabled faster, non-autoregressive synthesis with variance prediction, but still depended on large corpora.

More recent work targets low-resource adaptation. YourTTS [1] enables multilingual voice cloning with minimal data but lacks explicit prosody control. So-VITS-SVC [9] achieves high-fidelity conversion from limited data but underperforms in fine-grained prosody for TTS. These gaps motivate our framework, which explicitly integrates prosodic conditioning to enhance intelligibility and naturalness in tonal, low-resource settings.

3 Methodology

3.1 Voice Cloning Framework: So-VITS-SVC

Our voice cloning framework builds upon So-VITS-SVC [9], an extension of the original VITS architecture [5]. VITS is an end-to-end speech synthesis model that integrates variational autoencoders, normalizing flows, and adversarial training to generate high-fidelity waveforms with natural prosody and temporal coherence, without requiring explicit alignment. By disentangling speaker identity from linguistic content, So-VITS-SVC inherits these strengths while enabling non-autoregressive voice conversion. The architecture combines three key components: a SoftVC content encoder, the VITS generative backbone, and an NSF HiFiGAN vocoder [6]. This hybrid design allows high-fidelity, natural-sounding conversions with robust prosodic transfer. Crucially, So-VITS-SVC leverages HuBERT [2], a self-supervised content encoder, to extract phonetic representations directly from raw speech. This removes the need for explicit phoneme labels or alignment and enables the model to generalize across speakers and linguistic content. Unlike traditional approaches that rely on speaker encoders [4] or i-vector embeddings, So-VITS-SVC conditions the generator solely on HuBERT-derived content features and speaker-labeled training data. As illustrated in Fig. 1, this design supports flexible speaker adaptation from limited data while preserving both speaker identity and prosodic consistency in the converted speech.

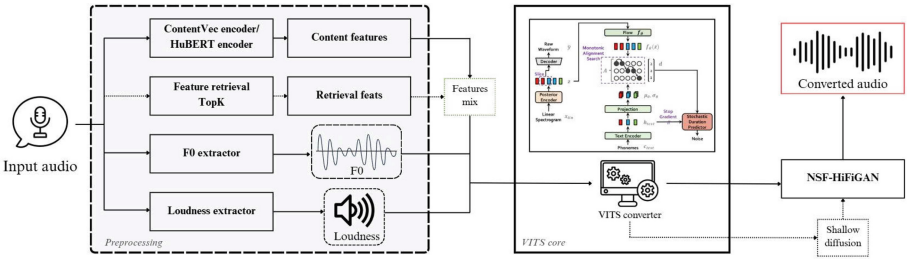

Fig. 1. Illustrated So-VITS-SVC architecture combining a SoftVC content encoder, the VITS backbone, and an NSF HiFiGAN vocoder, conditioned on HuBERT content features for high-fidelity voice conversion.

3.2 Prosody Extraction and Conditioning

To enhance the expressiveness and naturalness of synthesized speech–particularly for tonal languages such as Vietnamese–we introduce explicit prosodic modeling into the So-VITS-SVC framework. While the original architecture captures prosody implicitly through adversarial and reconstruction objectives, it lacks mechanisms for fine-grained prosodic control, which is critical

in tone-sensitive contexts. In our approach, we extract three core prosodic features–pitch (F0), energy (intensity), and temporal alignment–from short reference utterances using Parselmouth [3]. Pitch contours are estimated via autocorrelation-based tracking to capture the tone patterns essential for distinguishing lexical meaning in Vietnamese. Energy is derived through short-time energy analysis, providing information on vocal effort and emphasis, while duration is inferred from energy transitions and formant dynamics, reflecting natural rhythm and segment timing. These features are incorporated into the model's variance adaptor [11], enabling the generator to modulate its output based on explicit prosodic cues. Crucially, this design allows the system to generalize effectively under limited-data conditions by compensating for sparse speaker recordings with rich prosodic supervision. This is especially important in Vietnamese, where tonal inflections are not merely expressive, but semantically determinative.

3.3 Training Strategy

One of the main challenges in building adaptive TTS and voice conversion systems lies in the requirement for high-quality, speaker-diverse datasets, which are essential for generating natural and expressive speech. This challenge becomes even more pronounced in real-world scenarios where target speakers are fixed individuals, but may lack the time or resources to record long, studio-quality speech samples. To address this constraint, we employ HuBERT for phonetic feature extraction and Praat (via Parselmouth) for prosodic analysis, these raw audio corpora will be converted into a format suitable for prosody-aware voice conversion.

The Preprocessing Workflow. The preprocessing pipeline consists of four stages: resampling, data structuring, feature extraction, and configuration setup. First, all audio files are resampled to 44.1 kHz mono WAV format to ensure compatibility with both HuBERT and Parselmouth. Next, we extract phonetic content features using HuBERT and prosodic features–including pitch (F0), intensity, and duration–from each utterance. This end-to-end workflow is illustrated in Fig. 2, which summarizes the complete data preparation process required to construct a reliable database for downstream modeling.

The Training Stage. The training process enables high-fidelity voice conversion by mapping speaker-independent linguistic features to target speech waveforms within a modified VITS architecture. As shown in Fig. 3, the model receives three inputs: (i) HuBERT-extracted content embeddings, (ii) Parselmouth-extracted pitch contours ($F0$), and (iii) raw waveforms for reconstruction and adversarial feedback. To optimize synthesis quality, we employ a composite objective that balances multiple criteria critical to speech fidelity and prosody. The total loss is defined as:

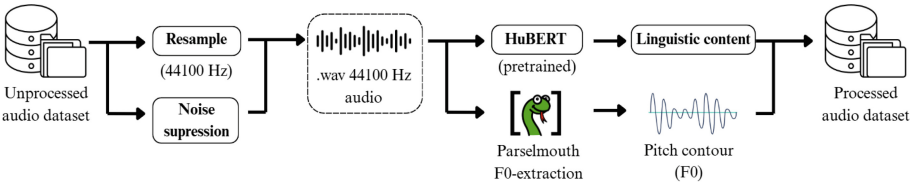

Fig. 2. Illustrated preprocessing workflow showing resampling, feature extraction, and configuration steps for building a reliable database.

$$\mathcal{L}_{\text{total}} = \lambda_{\text{mel}}\,\mathcal{L}_{\text{mel}} + \lambda_{F0}\,\mathcal{L}_{F0} + \lambda_{\text{adv}}\,\mathcal{L}_{\text{adv}} + \lambda_{\text{FM}}\,\mathcal{L}_{\text{FM}} + \lambda_{\text{KL}}\,\mathcal{L}_{\text{KL}}.$$

Here, $\mathcal{L}_{\text{mel}}$ is the $\mathcal{L}_1$ loss on mel-spectrograms, providing strong frame-level spectral reconstruction. $\mathcal{L}_{F0}$ is the $\mathcal{L}_2$ loss on log-scaled fundamental frequency predictions, promoting accurate pitch modeling for natural prosody. $\mathcal{L}_{\text{adv}}$ is a least-squares adversarial loss, encouraging perceptual realism, particularly in high-frequency regions. $\mathcal{L}_{\text{FM}}$ denotes the feature-matching loss using intermediate discriminator activations, improving training stability. Finally, $\mathcal{L}_{\text{KL}}$ is the Kullback–Leibler divergence between the approximate posterior and the prior, regularizing the latent space. The weighting factors λ are tuned empirically to emphasize spectral accuracy and prosodic fidelity while maintaining stable adversarial learning and well-structured latent representations.

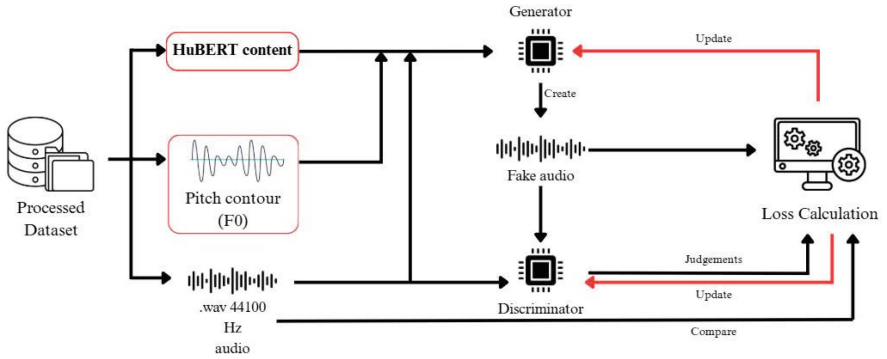

Fig. 3. Illustrated training pipeline of the proposed voice-cloning system. The model integrates HuBERT content embeddings, Parselmouth-derived $F0$, and raw waveforms, optimized with a composite loss that balances spectral accuracy, prosody, and perceptual realism.

The Inference Stage. After training, the system enters the inference phase to perform voice conversion on novel input text. As illustrated in Fig. 4, the

process unfolds in several steps. First, input text is synthesized into Vietnamese speech using an off-the-shelf TTS engine (e.g., Coqui or Edge TTS). The resulting audio is resampled to 44.1 kHz and normalized. A clean reference recording from the target speaker–single-speaker, high SNR, non-clipped, approximately 5–10 s in length, and phonetically diverse–is selected to guide conversion. Next, the synthesized TTS output is adjusted in Praat to match the target speaker's prosodic and spectral profile, aligning key features such as $F0$, formants, duration, and intensity. From this adjusted audio, linguistic features are extracted with a pretrained HuBERT model, while pitch contours ($F0$) are computed with Parselmouth. Finally, these representations are passed to the trained voice conversion model, which generates speech that preserves the content of the source text while enhancing target-speaker similarity.

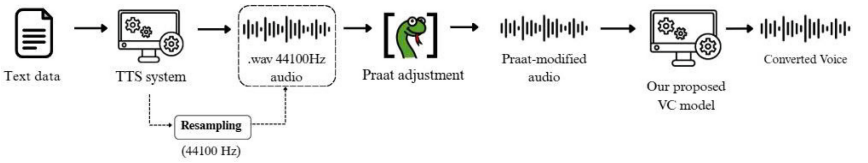

Fig. 4. Illustrated inference pipeline of the proposed TTS system. Input text is first synthesized into neutral speech, aligned with target-speaker prosody via Praat adjustment, then processed through HuBERT and Parselmouth features before voice conversion to produce speech with enhanced similarity.

4 Experiments

4.1 Dataset

For training and evaluation of the voice conversion model, we constructed a small-scale Vietnamese speech dataset tailored for low-resource speaker adaptation tasks. The dataset comprises approximately 10 to 30 min of speech from a single Vietnamese speaker, recorded in a closed room with minimal background noise and no acoustic treatment–using standard consumer-grade recording equipment such as basic smartphone recorders.

The recordings were manually segmented into around 200 utterances, each ranging from 5 to 10 s. Instead of employing a neural audio tokenizer for segmentation (e.g., EnCodec or Whisper), we opted for fixed-length utterance segmentation in the time domain. This decision was motivated by the need to preserve the natural prosody, pitch contour, and speaker-specific tonal characteristics, which are critical for effective voice conversion. All audio files were resampled to 44.1 kHz, converted to mono-channel 16-bit PCM WAV format, and underwent silence trimming and amplitude normalization prior to feature extraction. For model development, we adopted a 90/10 data split strategy–allocating 90%

of the utterances for training and 10% for testing. This split was applied during the dataset preparation phase, ensuring a consistent and speaker-balanced distribution of utterances across both sets.

While our dataset comprises recordings from a single Vietnamese speaker, this design choice was made to emphasize feasibility in low-resource scenarios. However, we acknowledge that a single-speaker dataset does not capture inter-speaker variability, such as gender-based pitch ranges, regional tonal differences, or idiosyncratic prosodic patterns. As a result, the system's generalizability to diverse speakers is limited. Future work will expand the dataset to include multi-speaker and gender-balanced recordings to better evaluate robustness across tonal and demographic variation.

4.2 Training Setup

We provide the key training and implementation details to support reproducibility. The model was trained on the preprocessed dataset using our modified approach to the So-VITS-SVC framework[1] with explicit F0 conditioning extracted via Parselmouth. The content encoder used pretrained HuBERT-base weights[2], while frame-level F0 contours were extracted via Parselmouth's autocorrelation-based analysis. Training was performed using the Adam optimizer (learning rate: 0.0001) for 50,000 steps with a batch size of 8. We trained the model for 2000 epochs using a step-based evaluation schedule that adapts its frequency according to validation behavior. During the bulk of training, we used a coarse evaluation interval (800 training steps) to conserve compute. When the validation loss showed consistent downward trends and both F0-RMSE and TER reached within 10% of their lowest observed values, we considered the model to be entering a "promising region". At this point, we increased evaluation frequency (interval reduced to 200 steps) to capture transient performance peaks more reliably. This adaptive schedule balances compute efficiency with reproducibility by explicitly linking evaluation frequency to validation metrics.

F0 Extraction Parameters (Parselmouth). We used Parselmouth's autocor-relation-based method with the following settings: voiced/unvoiced threshold = 0.3, pitch floor = 50 Hz, pitch ceiling = 500 Hz, and frame shift = 10 ms. These values were chosen to balance robustness across male and female speakers while maintaining accurate pitch tracking for tone-sensitive Vietnamese speech.

Loss Weights. We prioritize two objectives during training: (i) stable spectral reconstruction via mel-spectrogram alignment, and (ii) faithful pitch-contour preservation to minimize tone errors. To prevent instability on small datasets, adversarial training is introduced gradually, with the feature-matching (FM) loss

[1] https://github.com/svc-develop-team/so-vits-svc.
[2] https://huggingface.co/lj1995/VoiceConversionWebUI.

acting as a damping force on GAN dynamics. The KL divergence is kept small and annealed slowly to avoid over-regularizing the latent space. The final weights were set to $\lambda_{\text{mel}} = 45$, $\lambda_{F0} = 1$, $\lambda_{\text{FM}} = 2$, $\lambda_{\text{adv}} = 1$, and $\lambda_{\text{KL}} = 1$. We adopt a short warm-up strategy: during the first 5% of training steps, GAN components are disabled by setting $\lambda_{\text{adv}} = \lambda_{\text{FM}} = 0$; afterwards, they are restored to their target values, and λ_{KL} is gradually annealed from 0 to 1. These values were selected through a ±25% grid search around a base setting on a 20-utterance development set unseen during training. The selection criterion minimized F0-RMSE and Tone Error Rate (TER), with Word Error Rate (WER) used as a tie-breaker. This configuration consistently reduced tone errors while avoiding high-frequency artifacts, and preserved naturalness in low-resource Vietnamese settings. As shown in Fig. 5, the training process converged stably across all objectives, with both mel-spectrogram and F0 losses decreasing smoothly alongside the total loss.

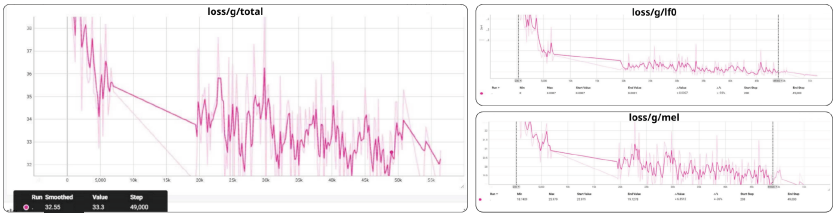

Fig. 5. Training performance metrics: total loss, F0 loss, and mel-spectrogram loss over time, illustrating stable convergence and effective pitch preservation.

Loss Statistics. The total loss (L_{total}) stabilized at around 32.865, aligning with expected trajectories for VITS-based systems (initial values typically 36–40, targeting < 30 for high-fidelity output in low-resource scenarios). The F0 loss (L_{F0}) reached a low of 0.8926 (comparable to g_{kl} in related architectures, where < 1 ensures strong prosody preservation critical for tonal languages like Vietnamese), e.g., reducing errors in words like 'ma' (ghost) vs. 'má' (cheek). The mel-spectrogram loss (L_{mel}) settled at 18.5964, which is reasonable for minimizing timbre discrepancies but improvable below 15 through extended training or refined prosody conditioning via Parselmouth. These trends indicate stable adversarial dynamics (reflected in minimal fluctuations post-50,000 steps) and near-perfect posterior encoding, confirming robust model convergence without artifacts. The results are comparable to benchmarks in voice conversion literature, underscoring the system's efficiency for resource-constrained TTS applications at HCMUT.

4.3 Evaluation Metrics

The system's performance was evaluated using both objective and subjective metrics to assess prosodic accuracy, intelligibility, and naturalness.

For objective evaluation, three automatic metrics were used: (1) F0-RMSE (Hz), measuring frame-wise pitch contour deviations to assess prosody accuracy; (2) model size (MB), indicating the storage efficiency; and (3) inference time (s/utterance), reflecting computational cost during deployment.

Subjective evaluation was conducted with 30 native Vietnamese participants. For both intelligibility and naturalness, listeners rated 15 synthesized utterances–generated from novel text prompts unseen during training–on a 5-point Likert scale, yielding Mean Opinion Scores for naturalness (MOS-N) and intelligibility (MOS-I). Transcription and tonal accuracy were measured using a listening test with the reference transcript. For each of the 15 utterances, raters saw the ground-truth text while listening and marked any mismatches. WER was scored at the word level by marking substitutions, deletions, and insertions against the transcript. TER was scored at the syllable level by flagging tone mistakes–i.e., wrong Vietnamese tone and diacritic/accent mark relative to the reference. When raters disagreed on a token, we used majority vote to choose a single label. Scores were averaged across utterances and raters, yielding separate measures of segmental (WER) and prosodic (TER) accuracy.

4.4 Results and Analysis

We conducted extensive experiments to evaluate the performance of the proposed prosody-aware voice conversion system against several baselines. The first baseline is a standard VITS-based voice conversion model without variance adaptor conditioning on F0. The second baseline is So-VITS-SVC, which does not explicitly model pitch and instead relies on emergent prosody. In contrast, our proposed system integrates Parselmouth-based F0 extraction with variance adaptor prosody conditioning, enabling more precise pitch and prosody control.

As shown in Table 1, the proposed system achieves the lowest F0 RMSE (19.4 Hz), demonstrating superior pitch accuracy compared to both VITS-VC (27.3 Hz) and So-VITS-SVC (21.8 Hz). Importantly, these gains are obtained with negligible increases in model size (237 MB) and inference time (1.18 s), confirming that explicit prosody modeling does not introduce significant computational overhead.

Table 1. Objective evaluation results comparing different voice conversion systems.

Model	F0 RMSE (Hz)↓	Model Size (MB)↓	Inference Time (s)↓
VITS-VC (no F0)	27.3	230	1.10
So-VITS-SVC	21.8	235	1.15
Proposed (F0 + Parselmouth)	**19.4**	237	1.18

Subjective evaluations, summarized in Table 2, further confirm the benefits of explicit F0 conditioning. The proposed system obtains the highest mean opinion scores–MOS-N = 4.4 for naturalness and MOS-I = 4.7 for intelligibility–surpassing both baselines. Error rates are also substantially reduced, with WER

Table 2. Subjective evaluation results showing Mean Opinion Scores and Error Rates.

Model	MOS-N↑	MOS-I↑	WER (%)↓	TER (%)↓
VITS-VC (no F0)	3.2	3.5	8.7	14.7
So-VITS-SVC	3.7	4.0	6.5	8.7
Proposed (F0 + Parselmouth)	**4.4**	**4.7**	**3.2**	**2.7**

dropping to 3.2% and TER to 2.7%, compared to 8.7% and 14.7% for the plain VITS-VC baseline. These results indicate that preserving pitch contours plays a critical role in improving comprehensibility for tonal languages such as Vietnamese.

Taken together, the objective and subjective results validate that explicit prosody modeling via Parselmouth-based F0 extraction significantly enhances voice conversion quality. Compared to approaches that rely solely on implicit prosody learning, the proposed system yields both technically robust and perceptually superior outputs, particularly in tonal contexts where accurate prosodic control is crucial for intelligibility.

4.5 Implementation

In the automated answering system at the Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), we applied our proposed Praat-modified So-VITS-SVC voice conversion framework. A practical challenge was that faculty secretaries–who serve as the system's voices–are not professional voice actors, making large-scale data collection infeasible. To address this, we developed a lightweight model capable of effective voice cloning from less than 30 min of recorded speech per speaker. Figure 6 illustrates the real-world deployment of the optimized model within the faculty's automated answering service, confirming its ability to deliver high-quality responses under practical hardware constraints.

Model Deployment Optimization. For deployment, the key requirements were high synthesis quality and fast inference on mid-range NVIDIA GPUs available in office computers. To meet these constraints, we converted the trained model to half-precision (FP16). This variant (`half_precision.pth`, 89.79 MB) reduces VRAM usage by nearly 50% compared to the original full-precision model (237 MB) while preserving naturalness and intelligibility. The effect on prosodic accuracy was negligible, with only a slight increase in WER and TER (0.4–0.7%). The FP16 model can be enabled directly via `.half()` during initialization, requiring no architectural changes. As shown in Table 3, the half-precision model achieves a 2.64× compression ratio while maintaining performance close to the full-precision baseline. This lightweight adaptation enables real-time inference in the target deployment environment without compromis-

ing audio quality, making it the most suitable choice for HCMUT's answering service.

Table 3. Comparison of deployment model variants in terms of size, compression ratio, quantization type, and evaluation metrics.

Model variant	Size (MB)↓	Compression ratio↑	Quantization type	WER↓	TER↓
Original	237.0	1.00×	None	3.2	2.7
Half-precision	89.8	2.64×	FP16	3.4	3.0

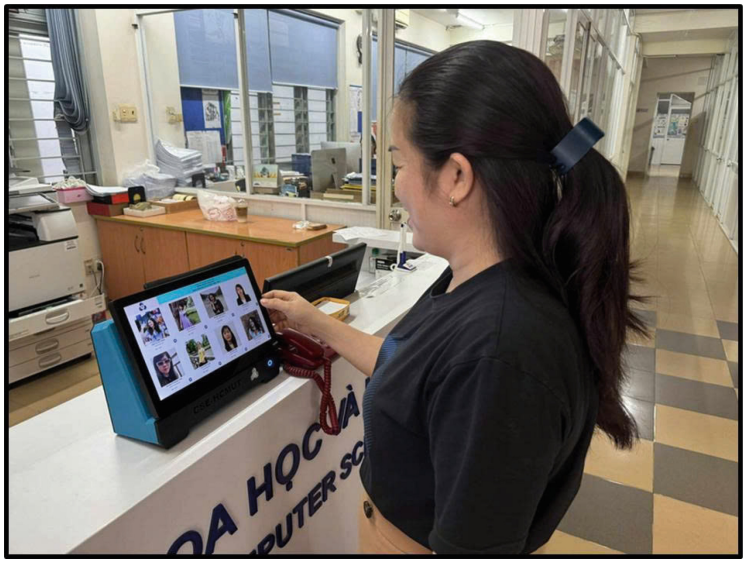

Fig. 6. Real-world deployment of the proposed Praat-modified So-VITS-SVC model in the HCMUT automated answering system.

5 Conclusion

In this work, we presented a prosody-aware voice conversion framework for Vietnamese based on So-VITS-SVC. The system integrates explicit F0 extraction via Parselmouth and self-supervised linguistic features from a pretrained HuBERT encoder, enabling few-shot speaker adaptation with limited data. We trained the model on a 30-minute single-speaker dataset using Google Colab platform. Evaluation results demonstrate that explicit prosody conditioning improves pitch

preservation and intelligibility while maintaining model efficiency, highlighting its potential for low-resource scenarios and tonal languages like Vietnamese. While promising, the current system remains an early-stage prototype. Future work will focus on refining prosody extraction, improving robustness, expanding speaker coverage, and exploring advanced expressiveness control for more scalable and practical deployment.

Acknowledgments. We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

Ethics and Consent. This study complies with institutional and academic ethical standards governing data collection and dissemination. Written informed consent was obtained from all participants, and all personally identifying metadata was removed. Publicly available materials are limited to short audio excerpts, while full-length recordings are stored under restricted access and may be shared only upon justified request. To reduce risks of misuse, all released synthetic audio is watermarked, model checkpoints are distributed solely for approved research purposes, and the project repository includes a responsible-use statement delineating acceptable applications and explicitly prohibiting malicious use.

References

1. Casanova, E., Weber, J., Shulby, C.D., Junior, A.C., Gölge, E., Ponti, M.A.: YourTTS: towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 2709–2720. PMLR (2022). https://proceedings.mlr.press/v162/casanova22a.html
2. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: HuBERT: self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 3451–3460 (2021). https://doi.org/10.1109/TASLP.2021.3122291
3. Jadoul, Y., Thompson, B., de Boer, B.: Introducing parselmouth: a python interface to praat. J. Phon. **71**, 1–15 (2018). https://doi.org/10.1016/j.wocn.2018.07.001
4. Jia, Y., et al.: Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc. (2018). https://proceedings.neurips.cc/paper_files/paper/2018/file/6832a7b24bc06775d02b7406880b93fc-Paper.pdf
5. Kim, J., Kong, J., Son, J.: Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 5530–5540. PMLR (2021). https://proceedings.mlr.press/v139/kim21f.html
6. van den Oord, A., et al.: WaveNet: a generative model for raw audio (2016). https://arxiv.org/abs/1609.03499

7. Ren, Y., et al.: Fastspeech 2: fast and high-quality end-to-end text to speech (2020). https://arxiv.org/abs/2006.04558. Accepted by ICLR 2021; Latest version: v8, August 2022
8. Shen, J., et al.: Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4779–4783 (2018). https://doi.org/10.1109/ICASSP.2018.8461368
9. svc-develop-team: SoftVC VITS Singing Voice Conversion. https://github.com/svc-develop-team/so-vits-svc (2023). Archived repository. Accessed 04 Oct 2025
10. Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T.: Speech parameter generation algorithms for HMM-based speech synthesis. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3, pp. 1315–1318 (2000). https://doi.org/10.1109/ICASSP.2000.861820
11. Łańcucki, A.: Fastpitch: parallel text-to-speech with pitch prediction. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6588–6592 (2021). https://doi.org/10.1109/ICASSP39728.2021.9413889