The 2025 VLSP Task on Vietnamese Voice Conversion: Overview and Preliminary Results

Nguyen Thi Thu Trang[1*] **Huu Tuong Tu**[1,2]
Le Hoang Anh Tuan[1]
[1]Hanoi University of Science and Technology [2]VNPT AI, VNPT Group
trangntt@soict.hust.edu.vn, huutu12312vn@gmail.com, tuanbkak66@gmail.com

Abstract

The VLSP 2025 is the eleventh annual international workshop organized by the Vietnamese Language and Speech Processing community. This year, we introduce the first Vietnamese Voice Conversion (VC) shared task, aiming to establish a standardized benchmark for evaluating and developing voice conversion systems in the Vietnamese language. The task focuses on building systems capable of converting a source speaker's voice to a target speaker while preserving linguistic content and naturalness. To support this, we constructed and released a large-scale multi-genre Vietnamese speech dataset containing over 26 hours of data from 100 speakers across various speaking styles and recording conditions. A total of 18 teams registered to participate, and the best system, based on a multilingual diffusion-transformer architecture, achieved a MOS of 4.29, SMOS_TGT of 3.65, and WER of 9.83. The shared task results provide valuable insights and a foundation for future research on robust Vietnamese voice conversion.

1 Introduction

Voice Conversion is the process of transforming the speech of one speaker (source) into the voice of another speaker (target) while preserving the linguistic content. In recent years, VC has received growing attention due to its wide range of applications, including personalized speech synthesis, voice anonymization, and data augmentation for speech and language technologies.

Two main approaches exist for building Voice Conversion (VC) systems. Text-based methods (Hussain et al., 2023; Sun et al., 2016) rely on annotated corpora with speech-text pairs for training. In contrast, text-free methods (li et al., 2022; Chen et al., 2020; Tu et al., 2025) focus on techniques, such as data augmentation or bottleneck or adversarial training, in an attempt to disentangle linguistic and speaker information directly from audio without requiring transcript labels. Despite recent advances, these methods frequently suffer from issues such as speaker information leakage, low output naturalness, and the loss of information. Consequently, achieving robust and scalable VC remains an open research problem.

To encourage further progress and establish a benchmark for the community, we organize the Vietnamese Voice Conversion Shared Task 2025 under the framework of the eleventh Vietnamese Language and Speech Processing Workshop (VLSP 2025). This is the first time a VC task has been included in VLSP. The goal is to provide a Vietnamese dataset dedicated to VC, evaluate current approaches, and foster the development of robust solutions in Vietnamese scenarios.

The shared task consists of a single evaluation track, designated as VC-T1. Participants are allowed to use pretrained models and external datasets. However, any pretrained model must be publicly available and accessible to all without requiring special access or permission. Teams must disclose and share the pretrained models they intend to use with the organizers and other participants. Additionally, participants are required to inform the organizers in advance about the specific pretrained models they plan to use and their intended purpose, so that eligibility can be verified.

For practical relevance, the dataset has been designed to cover diverse Vietnamese accents and various recording conditions, making it possible to evaluate the robustness of submitted systems under real-world scenarios. We expect the challenge to inspire innovative approaches, improve the performance of VC in Vietnamese, and contribute to advancing research in multilingual and low-resource voice conversion.

The rest of this paper is organized as follows. Section 2 introduces the timeline and status at the

*Corresponding author

time of publication. Section 3 describes data preparation. Section 4 reports the evaluation results. Finally, Section 5 concludes the paper and discusses future directions.

2 Timeline and status at publication time

The VLSP 2025 Challenge on Vietnamese Voice Conversion was announced and the training data along with the public test sets were released on July 1st, 2025. Teams were required to report any external pretrained models or datasets they used by July 10th, 2025, ensuring transparency in the use of resources. Following this, the private test set was released on August 14th, 2025, with the deadline for private test submissions set on August 17th, 2025. Results from the private test phase were shared with participants on August 23rd, 2025. Technical reports were submitted by August 30th, 2025.

As of the publication date of this paper, the evaluation and verification processes remain underway. Official notifications of acceptance are scheduled for September 27th, 2025, and the camera-ready versions are due on October 3rd, 2025. The VLSP 2025 conference, where final results and discussions will be presented, will be held on October 29th-30th, 2025. Throughout the challenge duration, communication was facilitated via a dedicated Zalo group to support participant interaction and information exchange.

3 Overview of tasks

The VLSP 2025 Challenge on Vietnamese Voice Conversion focuses on advancing voice conversion technology for Vietnamese, emphasizing both practical performance and model robustness. The task permits the use of publicly available pretrained models and external datasets, allowing participants to leverage existing resources to enhance system quality by incorporating transferable knowledge from large-scale models. This task provides an opportunity to explore how pretrained knowledge can boost performance in Vietnamese voice conversion. The next provides detailed descriptions of the task, datasets, and evaluation protocols.

4 Data Building

This chapter describes the construction pipeline for a multi-genre corpus. In addition, the author also provides detailed explanations for each stage in the pipeline and describes related works.

To build a large-scale speech dataset, this project selects YouTube as the primary source of data collection, leveraging the convenience of diverse multimedia content spontaneously uploaded by users. To ensure diversity across categories, the author created a taxonomy of audio genres such as Talk show, Vlog, Sharing, etc., as summarized in Table 1. In total, 1,859 audio samples were collected.

Category	Duration (hours)	# audio
Sharing	211	799
Talk show	157	167
Review	83	317
TV show	68	137
Game	40	123
Lecture	36	129
Vlog	24	87
News	14	100
Total	642	1859

Table 1: Statistics of audio duration and number of audio samples by category

Currently, most language data pipelines (He et al., 2025; weon Jung et al., 2025) focus on utterances from videos with a single speaker or structured conversations where turns are respected. These tend to be formal or prepared, lacking the spontaneity of natural daily conversations.

Therefore, besides genre, speech delivery is also categorized into two groups: (1) Prepared speech, including monologues or arranged dialogs without interruptions, and (2) Spontaneous speech, found in talk shows or natural conversations with free-flowing, unprepared content. Details and statistics are shown in Table 2.

Speech Style	Duration (h)	Percentage (%)
Spontaneous	225	35.05
Prepared	417	64.95
Total	642	100

Table 2: Duration and percentage of speech style

4.1 Speaker diarization

Unlike the task of speaker verification, which aims to recognize an individual regardless of speaking style, this task defines a "speaker" as an individual associated with a specific speaking style. This means if a person intentionally changes their speaking style (e.g., impersonation, joking, acting), the system considers them as a different speaker from

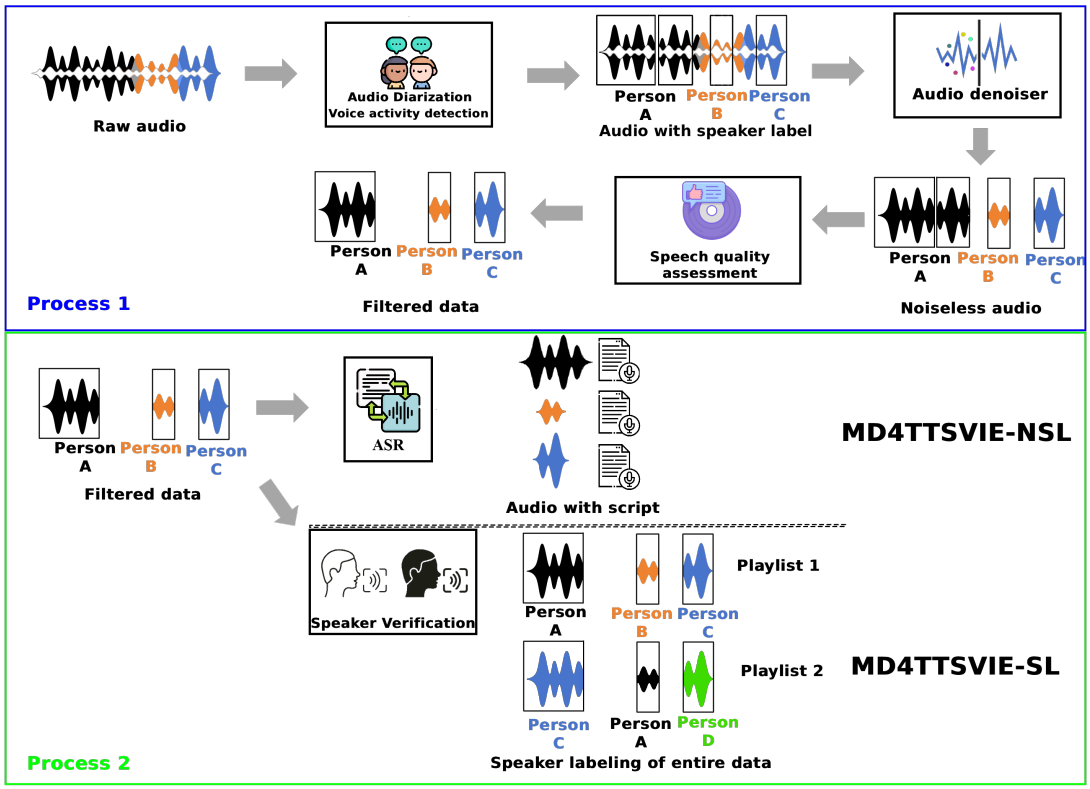

Figure 1: The entire process of building a dataset.

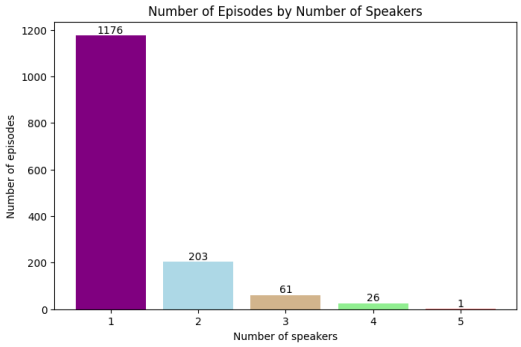

Figure 2: Speaker per video distribution

the original style.

To ensure each audio segment contains only one speaker at any given time-meeting the requirements for speech synthesis tasks - a speaker diarization pipeline based on pyannote speaker-diarization-3.1 [1] is used.

Real conversations often have speaker overlap and turn-taking, creating segments with multiple speakers. Unlike previous work assuming single-speaker audio, this approach accepts overlaps to include diverse conversational data from talk shows, news, interviews, and conferences.

This approach handles multiple simultaneous speakers, including overlaps and mislabeling caused by concurrent speech. This enables using diverse audio sources such as talk shows, TV programs, news, interviews, and conferences, which are rich in multi-speaker dialogue.

To ensure data quality, only single-speaker, clearly labeled, complete utterances are kept; overlapping or short segments are removed. Voice Activity Detection segments and cleans audio, and brief pauses between utterances from the same speaker are merged for natural flow.

4.2 Audio denoiseing

The collected YouTube data contains background noise that affects text-to-speech quality. While speaker diarization can handle noise, denoising is applied to improve audio quality and reduce processing time. Traditional denoisers (Defossez et al., 2020) often harm voice naturalness, so DeepFilterNet (Schröter et al., 2022) - a deep learning model that enhances speech by improving spectral envelopes and periodic components - is used. This approach effectively removes noise while preserving natural voice quality for speech synthesis.

[1] https://github.com/pyannote/pyannote-audio

4.3 Speech quality assessment

Audio quality evaluation is essential for building synthetic datasets and developing speech recognition systems, ensuring poor or noisy samples are removed to improve training efficiency and model reliability.

This process uses two advanced models - NISQA (Mittag et al., 2021) and WV-MOS (Andreev et al., 2023) - together to enhance evaluation accuracy and reduce bias. NISQA is a non-intrusive, multidimensional model combining CNN and self-attention to assess overall quality and specific factors like noisiness and distortion without needing a reference signal. WV-MOS predicts overall quality scores with high accuracy, trained on diverse, high-quality datasets.

Using both models allows combining NISQA's detailed analysis with WV-MOS's precise scoring, enabling better detection and removal of low-quality audio. A threshold of 3.2 is applied for both models Mean Opinion Scores to filter out substandard samples before further processing.

4.4 Auto speech recognition

Ensuring precise alignment between audio and transcripts is vital in TTS development. The author uses WhisperX (Bain et al., 2023), an ASR tool known for high accuracy in Vietnamese and precise time stamping. WhisperX transcribes audio and provides exact timestamps for sentences and words, enabling efficient data filtering, segment trimming, and normalization while reducing manual effort.

To maintain data quality, utterances with abnormal speech rates are filtered out. Based on studies and practical limits, a maximum threshold of 10 words per second is set to remove outliers, preserving the naturalness of the dataset.

4.5 Speaker labeling

The development of modern speech synthesis models increasingly integrates factors beyond text and audio, such as emotional information and speaker identity, to enhance expressiveness and naturalness (Cho et al., 2024; Shimizu et al., 2023). Motivated by these findings, the author proposes a data processing strategy to facilitate accurate speaker labeling.

Initially, speaker labeling is done independently for each episode, relying on the diarization model's reasonably accurate speaker identification. Visualizations using T-SNE Figure 3 confirm clear separation of speaker clusters, validating this approach. Utterances sharing the same speaker label are further clustered using similarity scores between speaker embeddings extracted by the ECAPA-TDNN model (Desplanques et al., 2020), pre-trained on VoxCeleb datasets (Nagrani et al., 2017; Chung et al., 2018).

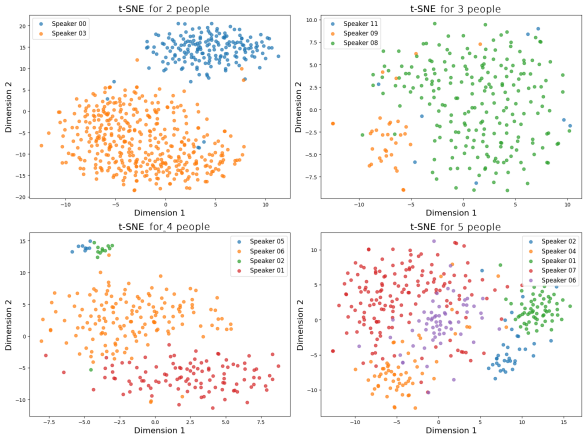

Figure 3: TSNE of speaker embeddings in each group

Similarity analysis Figure 4 reveals that utterances from the same speaker have notably higher similarity scores than those from different speakers. Based on this, a threshold of 0.75 similarity is chosen to group utterances belonging to the same speaker. A minimum of 30 utterances is required for a speaker to be considered valid, and 10% of these utterances are used to compute an average representative embedding.

In the next step, speaker embeddings from different sets are compared and clustered similarly, with a stricter similarity threshold of 97% to ensure accurate grouping. Statistical analysis supports these choices, balancing accuracy and efficiency in speaker clustering within the dataset.

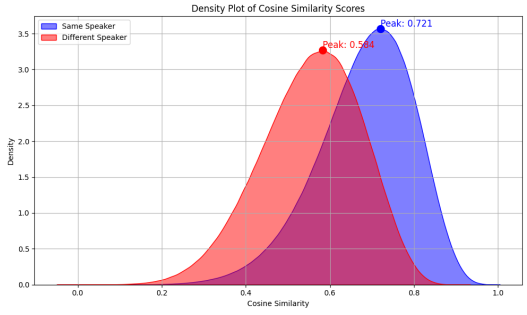

Figure 4: Similarity density table when comparing 2 utterances by the same/different speakers

4.6 Summary of Results

4.6.1 Results of Audio and Text Labeling

The audio and text data have been normalized and aligned with high quality following processing and labeling steps. Audio segments exceeding 50 seconds in duration were excluded to reduce computational burden and improve model efficiency. Additionally, segments with an average speaking rate exceeding 10 words per second were removed to maintain accuracy and naturalness in synthesized speech.

Duration (h)	# Utterances
26,35	**27,900**

Table 3: Final data for contest

4.6.2 Results of Speaker Labeling

The speaker labeling was developed from the dataset obtained in the Results of Audio and Text Labeling section. Through additional filtering and quality control, the 100 highest-quality speakers were selected for the final speaker labeling. Speaker labeling was conducted on this cleaned data by clustering utterances by individual speakers within each episode and merging speaker identities across the entire dataset. The outcomes demonstrate clear identification of speakers and utterances. This labeled data also serves as the **contribution dataset for the competition**. Each speaker in the dataset has between 100 to 200 utterances, ensuring sufficient examples per speaker for robust modeling and evaluation. This balanced distribution supports effective speaker-dependent tasks and speaker adaptation experiments within the challenge framework.

	Processed Data
# Speakers	**100**
# Utterances	**16,551**

Table 4: Statistics of speakers and utterances after labeling and cleaning

4.7 Testing dataset

The test dataset comprises a diverse collection of audio samples with varying file types and sources to ensure robust evaluation. It includes recordings data, student data, and retrieved data from YouTube (Table 5). The dataset is intentionally diversified in terms of regional accents and languages. Additionally, recordings were collected both from student voice sessions and publicly available YouTube content. During voice conversion testing, cross-regional voice conversion scenarios were included to assess model performance on speakers from different geographical regions. The dataset also features samples of singing to further increase diversity. This comprehensive and diverse test set aims to ensure that the competing teams' models have strong generalization capabilities on unseen data and across multiple speech variations. The evaluation setup strictly follows a zero-shot voice conversion scenario: test speakers do not overlap with training speakers, ensuring that systems are assessed on unseen identities. This setup challenges models to generalize conversion capabilities beyond familiar speakers.

Table 5: Composition of the test dataset

Audio Sample Type	Quantity
Internal recordings	8
Student Public Samples	11
Samples retrieved from YouTube	14

The dataset speakers are distributed by region as follows: 14 from the North, 11 from the Central region, and 8 from the South. This distribution ensures a diverse representation of regional accents in the testing data.

5 Evaluation

The composite scoring formula balances three crucial aspects of voice conversion quality: naturalness (MOS), speaker similarity (SMOS), and linguistic content preservation (WER). Each metric is essential to capture different facets of conversion performance. The weighting scheme assigns 40% importance to the difference between the speaker similarity of the reference and the output minus that of the source and the output (SMOS(ref, out) - SMOS(src, out)). This difference emphasizes the ability of a system to remove source speaker identity and accurately capture the target speaker's characteristics, a key challenge in voice conversion. If the output preserves too much source speaker information, it indicates speaker leakage, which compromises privacy and conversion fidelity. Meanwhile, MOS and WER are weighted at 30% each to balance speech naturalness and content intelligibility, both indispensable for practical VC applications.

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

$$\text{where} \quad \begin{cases} \text{SMOS}(\text{ref}, \text{out}): & \text{Speaker similarity between reference audio and output audio;} \\ \text{SMOS}(\text{src}, \text{out}): & \text{Speaker similarity between source audio and output audio;} \\ \text{MOS}: & \text{Naturalness rating;} \\ \text{WER}: & \text{Word Error Rate (calculated using ChunkFormer (Le et al., 2025)).} \end{cases}$$

Figure 5: Composite scoring formula used to evaluate voice conversion models.

5.1 Evaluation Metrics

Three main criteria were chosen to evaluate the voice conversion models, reflecting key aspects of model quality and effectiveness. **MOS** and **SMOS** represent subjective evaluation metrics. MOS rates the naturalness and overall quality of converted speech on a scale from 0 to 100, based on human listeners' judgments. SMOS measures how closely the converted speech resembles the reference speaker's voice, also rated by human perceptual judgments on the same scale. Grouping these together highlights their reliance on human evaluation. To ensure robust and generalizable MOS and SMOS evaluations, each contains 30 distinct pairs of audio samples. Each pair is independently labeled by 5 randomly assigned human raters to reduce bias and increase reliability. In total, 13 raters participated in the evaluation process; all were students from Hanoi University of Science and Technology.

Word Error Rate (WER): This metric evaluates content accuracy by comparing the converted speech to the source speech using a pretrained automatic speech recognition (ASR) model. WER serves as an objective measure to assess how well the converted audio preserves the original linguistic information, ensuring that important content is neither lost nor distorted during conversion.

These criteria are integrated into a composite scoring formula that balances speaker similarity, content accuracy, and perceptual quality-three fundamental factors for successful voice conversion (see Figure 5).

The submitted models are evaluated using three main criteria that reflect different aspects of voice conversion quality. First, the SMOS measures how similar the converted speech sounds compared to the reference speech, based on human perceptual ratings from 0 to 100. Second, the WER evaluates the accuracy of the spoken content by comparing the converted speech to the source speech using a pretrained ASR model, also scaled from 0 to 100. Finally, the MOS captures the naturalness and overall quality of the converted speech, rated directly by human listeners on the same scale.

The overall score is calculated by combining these measures as follows: 40% is given to the difference between the speaker similarity of the reference to the output and the speaker similarity of the source to the output, 30% to the naturalness score, and 30% to the complement of the WER (calculated as 100 minus the WER).

Here, SMOS(ref, out) refers to the speaker similarity between the reference audio and the converted output audio, SMOS(src, out) refers to the speaker similarity between the source audio and the converted output audio, MOS refers to the naturalness rating given by human listeners, WER refers to the Word Error Rate calculated using the ChunkFormer ASR model.

This multi-faceted evaluation approach ensures a balanced assessment of voice conversion systems, taking into account preservation of speaker identity, speech naturalness, and content accuracy.

To aid reproducibility and fair comparison, we provided participating teams with an objective evaluation toolkit implemented as a Gradio application. This toolkit enables automated computation of WER, facilitating consistent and convenient assessment across submissions.

5.1.1 Evaluation Results

5.2 Method summary

Two main types of system architectures were observed among the top teams: end-to-end and cascade approaches. Teams Twinkle and VCL utilized end-to-end systems, whereas ViettelRoar and ProfessorAgasa adopted cascade architectures.

Team Twinkle (Twinkle-VC): End-to-End Zero-Shot System (Seed-VC)

The Twinkle team proposed an end-to-end diffusion-transformer framework based on Seed-

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	3,53±0, 21	**3,66±0,21**	1,13±0, 08	12,95	67,53
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, 12	**1,11±0,07**	12,84	62,40

Table 6: Statistics of speakers and utterances after labeling and cleaning

Team	**Data / Augmentation**	**Approach Type**
Twinkle	Multilingual data (VCTK, JVS, Zeroth-Korean, PhoAudioBook, VLSP2025); *SR augmentation* for pitch and rhythm diversity	End-to-end system based on **Seed-VC** (Liu, 2024) with PhoWhisper-large semantic encoder
ViettelRoar	ViVoice (1000h Vietnamese) (Nguyen et al., 2024) + VCTK (English); phoneme-level training for cross-lingual robustness	Cascade system: **ChunkFormer** (Le et al., 2025) + **F5-TTS** (Chen et al., 2025)
VCL	VLSP + VNCeleb (Pham et al., 2023) datasets; no explicit augmentation	End-to-end systems: **MKL** (Lobashev et al., 2025)
ProfessorAgasa	PhoAudioBook + Vivoice, No augmentation reported	Cascade system: **ChunkFormer** (Le et al., 2025) + **ZipVoice** (Zhu et al., 2025)

Table 7: Summary of the Data and Methodological Choices of the Top 4 Teams in the Contest.

VC, incorporating Vietnamese-specific enhancements. The system integrates the *PhoWhisper-large* (Le et al., 2024) semantic encoder for robust linguistic representations, *CAM++* (Wang et al., 2023) for speaker embedding, and *BigVGAN-v2* (gil Lee et al., 2023) as a high-fidelity vocoder. During training, an *OpenVoiceV2 timbre shifter* was applied to reduce speaker leakage. To improve disentanglement, the team employed *SR augmentation (li et al., 2022)* along both time and frequency axes. Using multilingual datasets (VCTK (Veaux et al., 2017), JVS (Takamichi et al., 2019), Zeroth-Korean (Zeroth Project Contributors, 2017), PhoAudioBook (Vu et al., 2025), VLSP2025), Twinkle-VC achieved the best overall performance: **MOS 4.29**, **WER 9.83**, and high **SMOS_TGT 3.65**. The strong results indicate that Vietnamese-specific encoders and robust augmentation are critical to zero-shot generalization.

Team ViettelRoar: Cascade System (ChunkFormer + F5-TTS)

ViettelRoar adopted a two-stage ASR-TTS pipeline. The *ChunkFormer* ASR model handles long-form transcription efficiently through chunk-wise masked processing, while the customized *F5-TTS* module synthesizes natural, zero-shot speech using phoneme-level representations and flow-matching DiT architecture. Training was conducted sequentially on the *ViVoice* (1000h Vietnamese) and *VCTK* (English) datasets to enhance cross-lingual generalization. This system achieved the highest speaker similarity (**SMOS_TGT = 3.66**) with a moderate **WER = 12.95**. The clear modular structure improves interpretability and cross-lingual robustness, though minor ASR error propagation slightly affects naturalness.

Team VCL: Lightweight End-to-End Systems (MKL)

The VCL team experimented with an end-to-end model called MKL (Lobashev et al., 2025), a training-free voice conversion system based on optimal transport. The team used a retrieval-based method similar to KNN-VC (Baas et al., 2023) and fine-tuned WavLM to enhance its feature extraction capabilities using the VLSP and VNCeleb datasets. However, instead of using KNN as in KNN-VC to perform voice conversion, the team employed optimal transport to prevent information loss. Despite limited augmentation, the approach achieved **WER 10.98** and moderate naturalness (**MOS 3.72**).

The design emphasizes efficiency and adaptability, suitable for low-resource scenarios or edge deployment, though further data diversity could enhance robustness.

Team ProfessorAgasa: Cascade System (Chunkformer + ZipVoice)

This team also employed a cascade architecture combining *ChunkFormer* ASR with the *ZipVoice* TTS module. The system prioritizes compactness and efficiency, enabling fast inference. Although it achieved reasonable intelligibility (**WER 12.84**) and low source similarity (**SMOS_SRC 1.11**), its naturalness (**MOS 3.29**) remained behind other systems. The results suggest potential for lightweight applications but highlight the trade-off between quality and efficiency.

Comparative Discussion

End-to-end systems (Twinkle, VCL) excelled in capturing naturalness and minimizing ASR–TTS propagation errors, whereas cascade systems (ViettelRoar, ProfessorAgasa) offered clearer modularity and demonstrated better disentanglement performance, as evidenced by lower SMOS_SRC scores. Data augmentation and language-specific components proved decisive for superior performance. Cascade approaches, though interpretable, were more susceptible to the accuracy limits of ASR and TTS. Overall, Twinkle's integration of augmentation and Vietnamese-optimized semantic encoding achieved the best balance between *naturalness*, *content intelligibility*, and *speaker similarity*.

6 Conclusion

In this paper, we have presented an overview of Vietnamese Voice Conversion shared task at VLSP 2025. The task aimed to establish a common benchmark for evaluating and advancing voice conversion systems in the Vietnamese language. We have built and released a large-scale multi-genre Vietnamese speech dataset containing over 27 hours of recordings and 100 labeled speakers, covering diverse speaking styles, accents, and recording conditions.

A total of 18 teams registered to participate in this shared task. Two main system paradigms were observed: end-to-end diffusion-transformer models and cascade ASR-TTS architectures. The best system, proposed by **Team Twinkle**, achieved a **MOS of 4.29**, **SMOS_TGT of 3.65**, and **WER of 9.83**, demonstrating that multilingual training and robust augmentation strategies can significantly improve zero-shot voice conversion performance.

Looking ahead, we expect this shared task to serve as a valuable benchmark for future research and encourage further development on robust and scalable voice conversion systems in Vietnamese.

References

Pavel Andreev, Aibek Alanov, Oleg Ivanov, and Dmitry Vetrov. 2023. Hifi++: A unified framework for bandwidth extension and speech enhancement. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 1–5. IEEE.

Matthew Baas, Benjamin van Niekerk, and Herman Kamper. 2023. Voice conversion with just nearest neighbors. In *Interspeech*.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *Preprint*, arXiv:2303.00747.

Yen-Hao Chen, Da-Yi Wu, Tsung-Han Wu, and Hung yi Lee. 2020. Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization. *Preprint*, arXiv:2011.00316.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2025. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *Preprint*, arXiv:2410.06885.

Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, Sang-Hoon Lee, and Seong-Whan Lee. 2024. Emosphere-tts: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech. In *Interspeech 2024*, interspeech$_2$024, $page$1810˘1814.$ISCA$.

J. S. Chung, A. Nagrani, and A. Zisserman. 2018. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*.

Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. 2020. Real time speech enhancement in the waveform domain. *Preprint*, arXiv:2006.12847.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech 2020*, interspeech$_2$020.$ISCA$.

Sang gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2023. BigVGAN: A universal neural vocoder with large-scale training.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li,

Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. 2025. Emilia: A large-scale, extensive, multilingual, and diverse dataset for speech generation. *Preprint*, arXiv:2501.15907.

Shehzeen Hussain, Paarth Neekhara, Jocelyn Huang, Jason Li, and Boris Ginsburg. 2023. Ace-vc: Adaptive and controllable voice conversion using explicitly disentangled self-supervised speech representations. *Preprint*, arXiv:2302.08137.

Khanh Le, Tuan Vu Ho, Dung Tran, and Duc Thanh Chau. 2025. Chunkformer: Masked chunking conformer for long-form speech transcription. *Preprint*, arXiv:2502.14673.

Thanh-Thien Le, Linh The Nguyen, and Dat Quoc Nguyen. 2024. Phowhisper: Automatic speech recognition for vietnamese. In *The Second Tiny Papers Track at ICLR 2024*.

Jingyi li, Weiping tu, and Li xiao. 2022. Freevc: Towards high-quality text-free one-shot voice conversion. *Preprint*, arXiv:2210.15418.

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

Alexander Lobashev, Assel Yermekova, and Maria Larchenko. 2025. Training-Free Voice Conversion with Factorized Optimal Transport. In *Interspeech 2025*, pages 1373–1377.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Interspeech 2021*, interspeech$_2$021.$ISCA$.

A. Nagrani, J. S. Chung, and A. Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*.

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

Viet Thanh Pham, Xuan Thai Hoa Nguyen, Vu Hoang, and Thi Thu Trang Nguyen. 2023. Vietnam-Celeb: a large-scale dataset for Vietnamese speaker recognition. In *Proc. INTERSPEECH 2023*, pages 1918–1922.

Hendrik Schröter, Alberto N. Escalante-B., Tobias Rosenkranz, and Andreas Maier. 2022. Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering. *Preprint*, arXiv:2110.05588.

Reo Shimizu, Ryuichi Yamamoto, Masaya Kawamura, Yuma Shirahata, Hironori Doi, Tatsuya Komatsu, and Kentaro Tachibana. 2023. Prompttts++: Controlling speaker identity in prompt-based text-to-speech using natural language descriptions. *Preprint*, arXiv:2309.08140.

Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng. 2016. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Shinnosuke Takamichi, Kentaro Mitsui, Yuki Saito, Tomoki Koriyama, Naoko Tanji, and Hiroshi Saruwatari. 2019. Jvs corpus: free japanese multi-speaker voice corpus. *Preprint*, arXiv:1908.06248.

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

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

Thi Vu, Linh The Nguyen, and Dat Quoc Nguyen. 2025. Zero-shot text-to-speech for vietnamese. In *Proceedings of ACL*.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. 2023. Cam++: A fast and efficient network for speaker verification using context-aware masking. *Preprint*, arXiv:2303.00332.

Jee weon Jung, Wangyou Zhang, Soumi Maiti, Yihan Wu, Xin Wang, Ji-Hoon Kim, Yuta Matsunaga, Seyun Um, Jinchuan Tian, Hye jin Shim, Nicholas Evans, Joon Son Chung, Shinnosuke Takamichi, and Shinji Watanabe. 2025. Text-to-speech synthesis in the wild. *Preprint*, arXiv:2409.08711.

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

Han Zhu, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhaoqing Li, Weiji Zhuang, Long Lin, and Daniel Povey. 2025. Zipvoice: Fast and high-quality zero-shot text-to-speech with flow matching. *Preprint*, arXiv:2506.13053.