VNU Journal of Science: Comp. Science & Com. Eng., Vol. 41, No. 2 (2025) 68–85

VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

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

Nguyen Thi Thu Trang[1*], Huu Tuong Tu[2], Le Hoang Anh Tuan[1]

[1] *Hanoi University of Science and Technology, 1 Dai Co Viet, Hai Ba Trung, Hanoi, Vietnam*
[2] *VNPT AI, VNPT Group*

Received 05th December 2025;
Revised 18th December 2025; Accepted 22nd December 2025

Abstract: VLSP 2025 marks the eleventh annual workshop organized by the Vietnamese Language and Speech Processing community. This year, we introduce the inaugural Vietnamese Voice Conversion (VC) shared task, establishing a standardized benchmark for evaluating speech technologies in the Vietnamese language. The task focuses on developing systems capable of converting a source speaker's voice to a target identity while preserving linguistic integrity and naturalness. To support this initiative, we released a large-scale, multi-genre dataset comprising over 26 hours of speech from 100 speakers across diverse recording conditions. The challenge attracted 18 participating teams, with the top-performing system-based on a multilingual diffusion-transformer architecture-achieving a MOS of 4.29, an SMOS_TGT of 3.65, and a WER of 9.83. These results provide critical benchmarks and a robust foundation for future research in Vietnamese voice conversion.

Keywords: Voice conversion, Text to Speech, Multi-genre dataset, Benchmark evaluation.

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. Unlike non-tonal languages, Vietnamese presents unique challenges for Voice Conversion due to its complex tonal system and intricate phonetic nuances. A successful VC system for Vietnamese must not only transform the speaker's timbre but also accurately preserve the six tones, which are fundamental to the semantic meaning of the utterances. Any slight distortion in the fundamental frequency (F_0) trajectories during the conversion process can lead to a complete change in meaning or a significant drop in naturalness. This specific linguistic characteristic necessitates a more sophisticated modeling of prosodic features and fine-grained

*Corresponding author.
E-mail address: trangntt@soict.hust.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.6492

control over pitch contours to ensure that the converted speech remains both intelligible and culturally authentic. 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 [1, 2] rely on annotated corpora with speech-text pairs for training. In contrast, text-free methods [3–5] 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 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.

Currently, most language data pipelines [6, 7] 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.

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

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	1,859

Table 2. Duration and percentage of speech style

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

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 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-world conversations commonly include interruptions, overlaps, and turn-taking, meaning a segment attributed to one speaker may contain

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

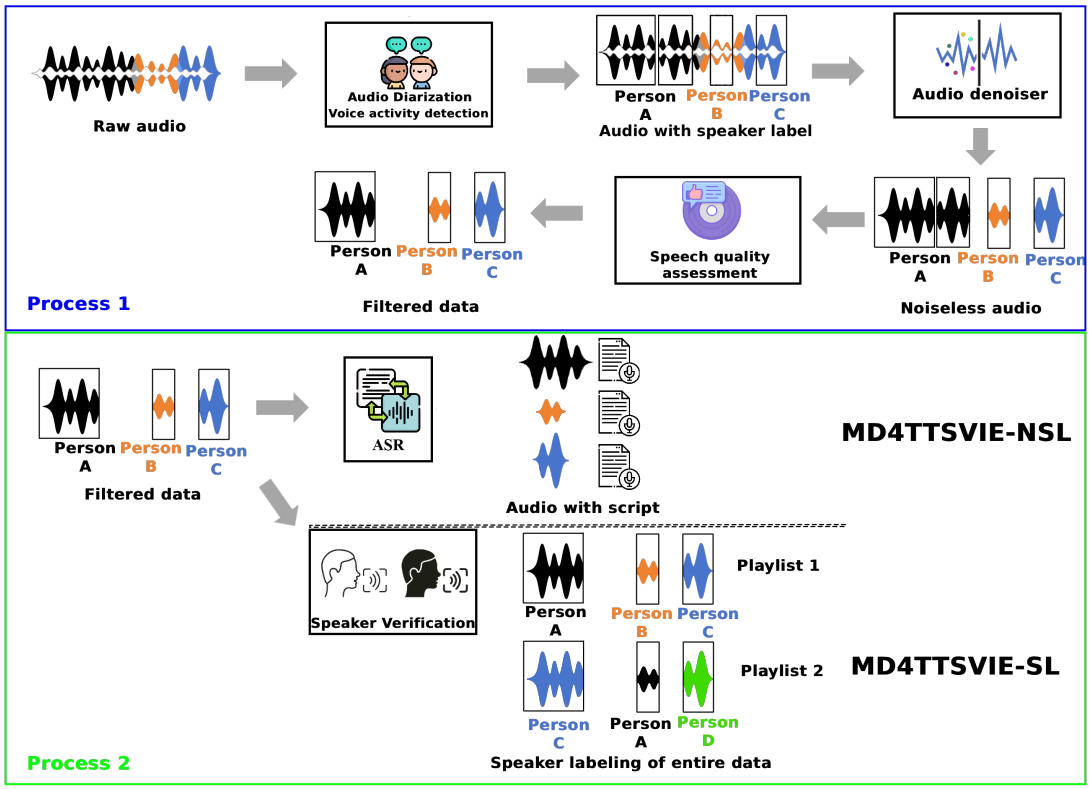

Figure 1. The entire process of building a dataset.

interleaving speech from another. While prior work often assumes single-speaker audio, this project embraces multi-speaker scenarios to enable the collection of more diverse conversational data from sources such as talk shows, TV programs, news broadcasts, interviews, and conferences.

Handling simultaneous speakers introduces challenges such as overlap, mislabeling, and fragmented utterances. However, effectively processing such cases opens access to large and diverse audio resources that traditional pipelines often discard due to their limitations. These multi-speaker environments provide valuable data for improving generalization and speaker variability in speech synthesis models.

To ensure high-quality data, this work defines three strict criteria for a valid data point: (i) each segment must contain exactly one speaker; (ii) at any given moment, only one speaker label is allowed; (iii) the utterance must be complete, with a clear start and end, without interruptions.

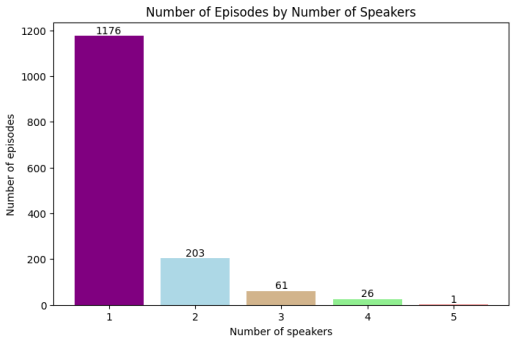

Figure 2. Speaker per video distribution.

To satisfy the first criterion, overlapped segments are removed by iterating through the diarization output in chronological order. Each data

point consists of the speaker label and segment boundaries; two additional fields, marking "conflict at start" and "conflict at end," are introduced to indicate overlap. As a result, each final segment includes five fields: start time, end time, speaker label, conflict-start flag, and conflict-end flag.

Next, to satisfy the second criterion, speakers with insufficient speaking duration are filtered out using a duration-based threshold:

$$T = \begin{cases} \dfrac{D \cdot 0.1}{\lfloor N/2 \rfloor}, & N > 1, \\ 0, & N \leq 1. \end{cases}$$

where T: threshold, D: sum duration of all speakers, N: number of speakers.

The rationale behind this thresholding mechanism is to prevent the inclusion of transient speakers or background voices that do not provide sufficient acoustic material for reliable model training. By dynamically adjusting T based on the total duration D and the number of detected speakers N, the pipeline maintains a flexible yet rigorous filtering criterion. In multi-party conversations, such as talk shows or games, the probability of overlapping speech and short interjections increases significantly. The denominator $\lfloor N/2 \rfloor$ acts as a penalty term to compensate for the fragmentation of speaker clusters, ensuring that only dominant and consistent voices are retained. This step is crucial for minimizing "label noise" which could otherwise degrade the performance of speaker-conditioned VC models.

This design assumes that in a two-speaker conversation, no more than 10% of total time involves simultaneous speech. As the number of speakers increases, the denominator accounts for potential duplicate or fragmented labels by assuming roughly one overlapping or redundant label per two additional speakers.

Finally, to ensure complete utterances, a pretrained Silero VAD model [8] is used to segment audio into coherent speech units and remove silence or non-speech portions. Segments marked with boundary conflicts (from overlap detection) are discarded to avoid utterances that were cut due to interruptions.

Although each VAD segment is a continuous utterance, overly short pauses may break natural phrasing. Following linguistic research by Kormos and Dénes [9], pauses shorter than 0.25 seconds between adjacent segments of the same speaker are treated as internal hesitations rather than true boundaries. Thus, such segments are merged to improve naturalness and discourse continuity.

With the complete processing pipeline-overlap removal, speaker filtering, VAD-based segmentation, and pause-aware merging-the resulting dataset contains only clean, single-speaker, complete utterances. This ensures suitability for downstream multi-speaker speech synthesis and speaker-conditioned modeling tasks.

4.2. Audio Denoising

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 [10] often harm voice naturalness, so DeepFilterNet [11] - 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.

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 [12] and WV-MOS [13] - 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 [14], 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

Recent advances in speech synthesis increasingly emphasize factors beyond raw text and acoustic signals-such as speaker identity and emotional cues-to improve naturalness and expressiveness [15, 16]. Building on these insights, this work introduces a data processing method designed to obtain reliable speaker labels.

The labeling process begins at the episode level. Each episode is diarized independently, leveraging the diarization model's strong ability to distinguish speakers. T-SNE visualizations Figure 3 reveal clearly separated embedding clusters, supporting this per-episode approach. Utterances tagged with the same initial label are then refined through clustering based on their pairwise similarity, computed from ECAPA-TDNN speaker embeddings [17], which are pretrained on the VoxCeleb datasets [18, 19].

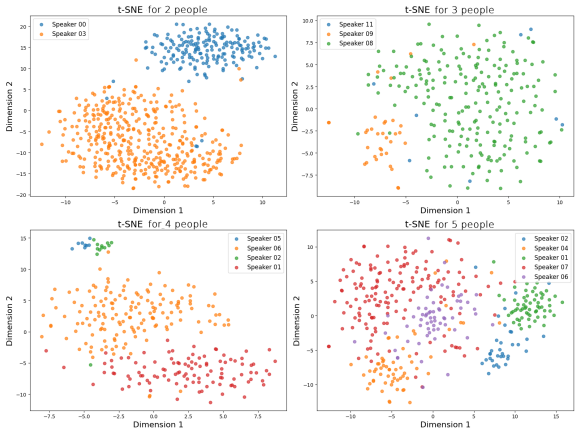

Figure 3. TSNE of speaker embeddings in each group.

Similarity distributions shown in Figure 4 indicate a near-Gaussian density for same-speaker pairs, peaking around 0.72. In contrast, embeddings from different speakers peak near 0.58. This separation motivates selecting a similarity threshold of 0.75 to consolidate utterances belonging to the same speaker within an episode.

To obtain stable cluster-level representations, each speaker is required to have at least 30 utterances. With an estimated same-speaker standard deviation of 0.09, the standard error of the mean embedding magnitude for 30 samples is:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{0.09}{\sqrt{30}} \approx 0.0164$$

which amounts to roughly 2.3% of the average similarity score (0.721). This ensures that the computed speaker embedding is statistically reliable. For each qualified cluster, 10% of the utterances (at least 30 samples) are randomly selected to compute the representative embedding.

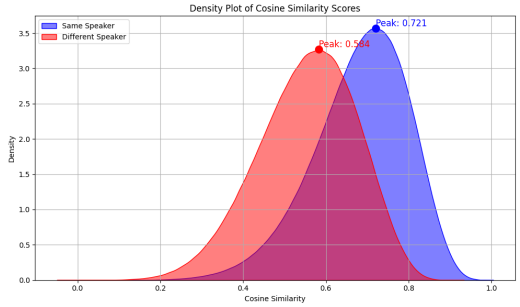

Figure 4. Similarity density table when comparing 2 utterances by the same/different speakers.

In the subsequent stage, these representative embeddings are compared across episodes to identify speakers who appear repeatedly within the same playlist. Similar to the earlier step, we construct a graph in which each node corresponds to a detected speaker, and edge weights are determined by embedding similarity.

Based on the statistical findings above, a stricter threshold of 97% similarity is adopted for merging speakers across episodes. This ensures that only highly consistent speaker representations-whose averaged embeddings differ within the expected 2–3% sampling variation-are combined. Such a cutoff provides a balance between minimizing erroneous merges and maintaining compact, coherent clusters, enabling robust cross-episode speaker identification for long-form content.

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.

Table 3. Final data for contest

Duration (h)	# Utterances
26.35	**27,900**

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.

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

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

Beyond the quantitative labeling results, we further analyze the qualitative performance regarding the balance between naturalness and identity preservation in the following subsection.

4.7. The Trade-off between Naturalness and Speaker Similarity

Analysis of the top-performing systems reveals a significant technical trade-off between speech naturalness (MOS) and target speaker similarity (SMOS_TGT). Team Twinkle, utilizing an end-to-end diffusion-based framework,

achieved the highest MOS of 4.29, but slightly lower similarity compared to ViettelRoar in specific categories. This phenomenon suggests that while diffusion models excel at generating high-fidelity, smooth acoustic trajectories, they may occasionally "oversmooth" the idiosyncratic vocal fry or micro-prosodic details that define a target speaker's unique identity. Conversely, the cascade system of ViettelRoar maintained high similarity by leveraging a flow-matching mechanism that preserves more aggressive speaker-specific features, though at the cost of a slightly higher WER due to potential error propagation between the ASR and TTS modules.

4.8. 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.

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.

Table 5. Composition of the test dataset

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

5. Evaluation

$$\begin{aligned}\text{Score} = 0.4 \times (\text{SMOS}(\text{ref}, \text{out}) - \text{SMOS}(\text{src}, \text{out})) \\ + 0.3 \times \text{MOS} + 0.3 \times (100 - \text{WER})\end{aligned} \tag{1}$$

where
SMOS(ref, out): Speaker similarity between reference audio and output audio;
SMOS(src, out): Speaker similarity between source audio and output audio;
MOS: Naturalness rating;
WER: Word Error Rate (calculated using ChunkFormer [20]).

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.

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 Equation 1).

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.

5.2. Evaluation Results

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.

5.3. Domain-wise Analysis

We perform a detailed evaluation of submitted systems using both subjective and domain-wise analyses to better understand system behavior under gender and dialect mismatch conditions. All subjective scores reported below are Mean Opinion Scores (MOS) on a five-point

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

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

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** [21] with PhoWhisper-large semantic encoder
ViettelRoar	ViVoice (1000h Vietnamese) + VCTK (English); phoneme-level training for cross-lingual robustness	Cascade system: **ChunkFormer** [20] + **F5-TTS** [22]
VCL	VLSP + VNCeleb [23] datasets; no explicit augmentation	End-to-end systems: **MKL** [24]
ProfessorAgasa	PhoAudioBook + Vivoice, No augmentation reported	Cascade system: **ChunkFormer** [20] + **ZipVoice** [25]

scale collected from human raters. Speaker similarity (SMOS) is also human-rated on the same 1-5 scale and reported separately for similarity between the output and the *target* reference (SMOS_TGT) and similarity between the output and the *source* (SMOS_SRC). The analysis is based on $n = 900$ test conversions.

5.3.1. Gender-domain Analysis

Table 8 summarizes MOS and SMOS by gender domain (same & cross).

The results reveal an interesting and somewhat counter-intuitive trend. Although one might expect same-gender conversion to be inherently easier due to closer vocal characteristics, the cross-gender condition performs competitively and in several cases even better. In terms of naturalness, cross-gender samples achieve a MOS of 3.40, essentially matching the same-gender score of 3.45. More notably, the target-speaker similarity (SMOS_TGT) is higher in the cross-gender setting (2.88 &. 2.65), suggesting that the model is more capable of capturing target-specific cues when the transformation requires a larger stylistic shift.

This pattern is further supported by the source-speaker similarity values (SMOS_SRC), where cross-gender conversion shows a greater reduction (1.30 & 1.79), indicating that the model more effectively suppresses residual characteristics of the source speaker in the cross-gender case. Taken together, these findings suggest that increased gender contrast may actually help the model disentangle speaker identity, leading to more distinct and perceptually clearer conversions.

Fine-grained gender-pair analysis To better understand how gender interactions affect perceptual performance, we further decompose the eval-

Gender-domain	#	MOS	SMOS_SRC	SMOS_TGT
same-gender	390	3.45	1.78	2.65
cross-gender	510	3.40	1.31	2.88

Table 8. MOS and SMOS_TGT by gender domain.

uation into four specific gender-pair directions. Table 9 summarizes MOS, SMOS_SRC, and SMOS_TGT for each pair.

Gender pair	MOS	SMOS_SRC	SMOS_TGT
M→M	3.46	1.83	2.56
F→F	3.44	1.76	2.70
F→M	3.44	1.33	2.89
M→F	3.36	1.28	2.87

Table 9. MOS and SMOS scores for fine-grained gender pairs.

Male-to-Male (M→M): This pair achieves the highest MOS among same-gender conversions (3.46), indicating relatively stable naturalness when both speakers share similar timbral and pitch characteristics. However, SMOS_TGT remains moderate (2.56), suggesting that although the system preserves general speech quality, capturing the fine-grained target identity within the same gender-where voices may share overlapping acoustic regions-remains challenging. The relatively high SMOS_SRC (1.83) further indicates stronger residual source leakage compared with cross-gender settings.

Female-to-Female (F→F): F→F conversion shows a similar pattern to M→M, with slightly lower MOS (3.44) but a marginally higher SMOS_TGT (2.70). This suggests that the system captures female vocal traits somewhat more accurately, potentially due to the broader pitch flexibility and clearer formant structure in female speech. Nonetheless, the SMOS_SRC score (1.76) indicates that, as with male–male conversion, residual source traits remain more prominent than in cross-gender conversions.

Female-to-Male (F→M): This direction achieves the best overall target similarity (SMOS_TGT = 2.89), showing that the system handles pitch lowering and timbre broadening effectively. The relatively low SMOS_SRC (1.33) suggests that the model suppresses source characteristics more successfully when shifting from a higher-pitched to a lower-pitched voice. MOS remains competitive (3.44), indicating that the significant timbral and pitch transformation required in this direction does not degrade perceived quality.

Male-to-Female (M→F): M→F conversion yields the lowest MOS among all pairs (3.36), consistent with the greater difficulty of expanding pitch range and modifying spectral envelopes when converting male voices to female voices. Despite this, SMOS_TGT (2.87) remains high, showing that listeners still perceive the target identity clearly. The lowest SMOS_SRC score across gender pairs (1.28) further supports the notion that large stylistic gaps-such as male to female-help the model reduce source leakage more effectively.

5.3.2. Dialect-domain Analysis

The linguistic diversity of Vietnam, primarily categorized into Northern (N), Central (C), and Southern (S) dialects, poses a non-trivial challenge for zero-shot voice conversion. Each dialect is distinguished not only by its phonetic inventory but also by its unique tonal contours and rhythmic patterns. Table 10 summarizes the aggregated performance metrics across same-dialect and cross-dialect scenarios.

The macro-level results reveal a consistent performance gap: same-dialect conversion consistently outperforms cross-dialect conversion across all perceptual dimensions. Specifically, same-dialect samples achieve a higher MOS

Table 10. MOS and SMOS_TGT by dialect domain

Dialect domain	#	MOS	SMOS_SRC	SMOS_TGT
same-dialect	420	3.51	1.56	2.84
cross-dialect	480	3.35	1.47	2.73

(3.51 vs. 3.35), suggesting that the diffusion-transformer architecture finds it inherently easier to maintain naturalness when the source and target share congruent prosodic structures. Interestingly, the lower SMOS_SRC in cross-dialect tasks (1.47) compared to same-dialect (1.56) indicates that the model is more "aggressive" in suppressing source identity when the dialectal mismatch is high, whereas in same-dialect cases, the acoustic similarity between source and target makes perfect disentanglement more difficult to achieve.

Although the differences are not large, the results imply that dialectal mismatch introduces additional variability in prosody and phonetic realization, making the target identity slightly harder to reproduce faithfully. Nonetheless, the cross-dialect performance remains strong overall, demonstrating that the model generalizes reasonably well across dialect boundaries.

Fine-grained dialect-pair analysis: To more precisely characterize the impact of dialect mismatch, we further analyze all major source–target dialect combinations. Table 11 reports MOS and speaker-similarity scores for each direction.

North→South: This direction achieves relatively strong naturalness (MOS = 3.57), indicating that mapping Northern speech characteristics onto Southern prosody is handled effectively by the system. The SMOS_TGT score (2.64) is moderate, suggesting that while the target identity is captured reasonably well, the relaxed prosody and more open vowel patterns of Southern speech still present challenges.

South→Central: This pair exhibits consistently strong performance, especially in SMOS_TGT (3.20), the highest across all dialect directions. This suggests that the system captures the Central dialect's distinctive tonal and prosodic patterns particularly well when starting from Southern speech, which provides a rhythmically flexible baseline for transformation. The MOS score (3.69) further indicates stable naturalness.

Table 11. MOS and SMOS scores for fine-grained dialect pairs. N = Northern, S = Southern, C = Central

Dialect pair	MOS	SMOS_SRC	SMOS_TGT
N→N	3.35	1.66	2.82
N→S	3.57	1.47	2.64
S→N	3.54	1.76	2.63
S→S	3.93	1.32	3.12
S→C	3.69	1.44	3.20
C→N	2.89	1.34	2.58
C→S	2.47	1.23	2.63
C→C	3.27	1.65	2.37

Central→South: This is one of the lowest-performing directions (MOS = 2.47), reflecting the difficulty of reducing the rich tonal inventory of Central Vietnamese into the comparatively flatter Southern contour. Although SMOS_TGT (2.63) remains reasonable, the drop in MOS indicates audible artifacts likely caused by tone simplification and timing adjustments.

South→North: Performance in this direction is strong overall (MOS = 3.54), though the elevated SMOS_SRC (1.76) suggests that Northern tonal precision is challenging to reproduce and may result in increased residual source characteristics. SMOS_TGT (2.63) is moderate and consistent with the complexity of the Northern dialect's tone system.

North→North: As expected for a same-dialect setting, this pair performs robustly, with

MOS = 3.35. However, it also shows one of the highest SMOS_SRC scores (1.66), indicating that removing subtle source-specific tonal cues is difficult when source and target speakers share very similar acoustic structures.

Central→North: This direction yields moderate performance (MOS = 2.89), reflecting challenges in mapping Central tones into the analytically structured Northern tone system. Although SMOS_TGT (2.58) remains acceptable, the reduction in MOS suggests perceptual artifacts related to tonal alignment.

South→South: This pair achieves the highest MOS overall (3.93), illustrating that conversions within the Southern dialect are particularly stable. SMOS_TGT is also high (3.12), indicating that intra-dialectal identity features are captured effectively. The relatively low SMOS_SRC (1.32) further suggests that the system suppresses residual source traits more effectively when dialectal patterns are consistent.

Central→Central: Although this is also a same-dialect condition, it performs less favorably than other intra-dialect conversions (MOS = 3.27). The Central dialect exhibits high tonal complexity and greater inter-speaker variability, which likely increases the difficulty of reconstructing target-specific identity cues, as reflected in the lowest SMOS_TGT score among same-dialect pairs (2.37).

6. Methodology Summary of Top Participants

In this section, we provide a comprehensive technical analysis of the systems developed by the top-performing teams. The diversity in architectures, ranging from end-to-end diffusion models to modular cascade pipelines, offers valuable insights into the current state of Vietnamese voice conversion.

6.1. Team Twinkle (Twinkle-VC): End-to-End Diffusion-Transformer

The Twinkle team proposed an advanced end-to-end framework based on the Seed-VC [21] architecture, which utilizes a diffusion-transformer backbone. To tailor the system for the Vietnamese language, they integrated the *PhoWhisper-large* [26] semantic encoder, providing a rich linguistic foundation. The speaker identity is captured using *CAM++* [27], while speech reconstruction is handled by the *BigVGAN-v2* [28] vocoder.

To address the critical challenge of speaker leakage, the team applied an *OpenVoiceV2* timbre shifter during the training phase. Furthermore, they employed *SR augmentation* [3] along both temporal and frequency axes to improve the disentanglement of content and style. Leveraging a multilingual corpus including VCTK, JVS [29], Zeroth-Korean [30], and PhoAudioBook [31], Twinkle-VC achieved the highest overall naturalness with a MOS of 4.29. This suggests that the combination of self-supervised semantic features and robust augmentation is highly effective for zero-shot scenarios.

6.2. Team ViettelRoar: Modular Cascade Architecture

ViettelRoar adopted a robust two-stage cascade paradigm. The first stage involves the *ChunkFormer* [20] ASR model, which produces high-accuracy transcriptions through masked chunk-wise processing. These linguistic tokens are then passed to a customized *F5-TTS* [22] module. This synthesis component utilizes a flow-matching Diffusion Transformer (DiT) to generate natural, target-conditioned speech.

The team utilized an extensive 1000-hour Vietnamese dataset (ViVoice) combined with the English VCTK dataset to enhance cross-lingual generalization. This modular strategy allowed ViettelRoar to achieve the highest speaker similarity score (SMOS_TGT of 3.66). While cascade systems may suffer from error propagation between the ASR and TTS stages, the use of a large-scale pre-trained synthesis backbone proved superior in accurately mimicking the target speaker's unique vocal characteristics.

6.3. Team VCL: Optimal Transport-based Training-Free System

The VCL team introduced a training-free approach based on the MKL [24] framework. Unlike traditional neural conversion methods that require gradient-based fine-tuning, this system employs the principles of optimal transport to map source features to the target speaker's distribution. The team utilized fine-tuned *WavLM* features to enhance acoustic representation, drawing data from the VLSP and VNCeleb [23] datasets.

By substituting the standard K-Nearest Neighbors approach used in KNN-VC [32] with an optimal transport solver, the VCL system effectively minimized information loss during the conversion process. This lightweight and efficient design is particularly promising for real-world applications where low latency and minimal computational overhead are required. Despite its efficiency, the model maintained a competitive WER of 10.98, demonstrating the viability of non-parametric methods in voice conversion.

6.4. Team ProfessorAgasa: Efficiency-Oriented Cascade System

Similar to the second-place team, ProfessorAgasa implemented a cascade architecture combining *ChunkFormer* [20] for transcription and *ZipVoice* [25] for synthesis. The primary design goal of this system was to achieve high-speed inference and model compactness.

The *ZipVoice* module enables fast, high-quality zero-shot conversion, making it suitable for deployment on edge devices. Although the naturalness score (MOS 3.29) was lower than the diffusion-based counterparts, the system demonstrated excellent speaker disentanglement, as evidenced by a low SMOS_SRC of 1.11. This indicates that the source speaker's identity was successfully suppressed, a key requirement for privacy-preserving voice conversion applications.

6.5. Comparative Discussion

The competition results reveal a fascinating trade-off between the two dominant paradigms: End-to-End (E2E) and Cascade architectures. End-to-End systems, notably exemplified by Team Twinkle, leveraged the latent representations of self-supervised models like PhoWhisper to bypass the explicit phoneme bottleneck. This approach allowed for a more fluid transfer of prosody and emotion, as evidenced by their superior MOS scores. However, E2E models are typically more computationally intensive and require vast amounts of high-quality data to avoid speaker leakage.

On the other hand, Cascade systems such as ViettelRoar demonstrated high robustness in terms of speaker similarity. By separating the task into ASR and TTS modules, these systems can leverage massive external TTS datasets to ensure that the target voice identity is reproduced with high fidelity. The "price" for this modularity is often a slight increase in WER due to error propagation: if the ASR module misinterprets a word, the TTS will faithfully synthesize the incorrect term. Our analysis suggests that for Vietnamese, hybrid approaches that combine the linguistic stability of cascades with the prosodic richness of E2E models-potentially through diffusion-based refinement-might be the most promising path for future research.

6.6. Self-Observed Qualitative Evaluation

Beyond the quantitative metrics provided by MOS and SMOS, a rigorous qualitative audit of the synthesized samples offers deeper insights into the model's capabilities in extreme and diverse scenarios. This manual inspection focus specifically on prosodic stability, temporal alignment, and the preservation of emotional intensity across varying acoustic environments.

In terms of prosody and temporal stability, the model demonstrates a remarkable capacity for handling non-standard speech rhythms and

atypical vocal deliveries. A particularly notable observation was made during the analysis of samples involving rapid-tempo utterances or even singing voices. In cases where the target reference contains fast-paced singing or rhythmic speech, the diffusion-transformer architecture successfully synchronized the semantic content with the target's temporal structure without compromising the clarity of consonants or the integrity of vowel durations. This indicates that the semantic tokens derived from the ASR encoder are effectively decoupled from the temporal information of the target speaker, allowing the vocoder to reconstruct high-fidelity speech even under challenging prosodic constraints that would typically cause alignment failures in traditional cascade systems.

The robustness of the system was further tested against the consistency of emotional intensity across three distinct audio domains: studio-quality recordings, internal device captures, and wild audio extracted from YouTube. In controlled environments, such as studio or clean internal recordings, the conversion remains exceptionally stable, faithfully preserving the subtle emotional nuances-such as excitement, hesitation, or calmness-inherent in the target speaker's profile. The model effectively maps the source linguistic content onto the target's affective space, ensuring that the prosodic "color" of the emotion is maintained throughout the utterance.

However, a slight "emotion weakening" effect was observed in samples sourced from noisy YouTube environments. When the input audio contains significant background interference, non-stationary noise, or heavy reverberation, the model's inherent denoising mechanisms-while successful in cleaning the signal-occasionally "smooth out" the micro-prosodic variations that carry emotional weight. This results in a phenomenon we term spectral blurring, where the converted voice loses a degree of its crispness and emotional "edge" compared to conversions originating from clean sources. Despite these minor acoustic artifacts, the overall intelligibility and speaker identity remain intact, suggesting that while the model is robust to dialectal and rhythmic variations, future iterations should incorporate more advanced noise-robust training objectives to better preserve emotional high-frequency components in "in-the-wild" scenarios.

6.7. Robustness Across Variable Acoustic Environments

The variability in recording conditions-ranging from studio-quality student recordings to noisy YouTube "in-the-wild" audio-served as a rigorous test for model robustness. We observed that systems employing advanced denoisers like DeepFilterNet [11] were able to maintain higher MOS scores even when the source audio contained background music or environmental interference.

However, for the YouTube "Vlog" and "Game" categories, which often feature spontaneous speech with high emotional arousal, most systems showed a noticeable drop in SMOS_TGT. This suggests that current zero-shot models still struggle to decouple the target's identity from the source's emotional state, leading to a "prosody leakage" where the output voice sounds like the target speaker but inherits the source speaker's unintended stress and tempo patterns.

6.8. Discussion and Error Analysis

The experimental results obtained from our evaluation framework provide a comprehensive overview of the current state of Vietnamese Voice Conversion. A critical observation is the inherent trade-off between the perceptual naturalness of the synthesized speech and the fidelity of the speaker's identity preservation. While diffusion-based architectures demonstrated a superior ability to generate fluid and high-fidelity acoustic trajectories, resulting in a leading MOS of 4.29, they are not without significant limitations. A rigorous spectral analysis reveals a persistent "oversmoothing" effect within the gener-

ated Mel-spectrograms, where the stochastic denoising process inadvertently suppresses the fine-grained micro-prosodic variations that constitute a speaker's unique vocal signature. This loss of spectral detail often manifests as a reduction in "vocal fry" or breathiness, which are essential cues for human listeners to distinguish between similar timbres.

Furthermore, our findings underscore the extreme sensitivity of Vietnamese phonology to fundamental frequency (F_0) modeling. Unlike non-tonal languages where pitch primarily conveys intonation or emotion, in Vietnamese, the six-tone system is the primary vehicle for lexical distinction. Our error analysis highlighted several failure cases where the model failed to accurately reconstruct the sharp inflections required for "thanh ngã" or the deep glottalization of "thanh hỏi". These subtle distortions in the F_0 trajectory do not merely affect the prosody; they frequently result in a complete shift in semantic meaning, which is reflected in the elevated Word Error Rate (WER) across certain test samples. This suggests that for tonal languages, the evaluation of VC systems must prioritize tonal integrity as a core component of linguistic intelligibility.

6.9. Future Work

Building upon the insights gained from this study, several critical avenues for future research are identified to bridge the gap between experimental results and practical deployment. Primarily, we intend to investigate the integration of explicit tonal-aware loss functions and multi-task learning frameworks that jointly optimize for speaker identity and tonal accuracy. By incorporating a dedicated F_0 conditioning module, we aim to provide the model with a more granular control over the pitch contours, ensuring that the subtle nuances of Vietnamese regional dialects-each with their own tonal variations-are preserved with higher fidelity.

Additionally, to address the "oversmoothing" phenomenon identified in diffusion models, we plan to explore the use of hybrid architectures that combine the stability of diffusion-based generation with the high-frequency detail preservation of adversarial training. This could involve the development of specialized vocoders capable of reconstructing fine spectral details from latent representations, thereby restoring the "natural roughness" of human speech. On the operational side, future work will also focus on model compression and low-latency optimization techniques. Our goal is to enable these sophisticated VC systems to operate efficiently on edge devices, facilitating real-time applications such as personalized assistive technologies for individuals with speech impairments or secure voice anonymization in sensitive communication environments.

7. Conclusion

This paper has presented a comprehensive synthesis of the inaugural Vietnamese Voice Conversion (VC) shared task, organized as a cornerstone of the VLSP 2025 workshop. By curating and disseminating a high-quality, multi-genre dataset consisting of 26 hours of speech from 100 diverse speakers, we have effectively addressed the long-standing void of standardized benchmarks for tonal language voice conversion. The task has not only provided a rigorous framework for performance evaluation but also catalyzed the development of innovative architectures tailored to the phonetic and prosodic intricacies of the Vietnamese language.

The collective contributions from 18 participating teams have significantly redefined the performance envelope of Vietnamese VC systems. A critical analysis of the results suggests that diffusion-based transformer architectures, particularly when integrated with advanced semantic representations like PhoWhisper, represent the current state-of-the-art, achieving a remarkable MOS of 4.29. Furthermore, our findings reveal a fundamental technical dichotomy: while end-to-end models demonstrate superior fluidness in

prosodic generation, cascaded frameworks maintain higher fidelity in speaker identity preservation and linguistic accuracy, as evidenced by lower Word Error Rates (WER). These outcomes confirm that for tonal languages, the precise modeling of F_0 trajectories is not merely a technical requirement but a prerequisite for preserving semantic integrity.

Despite these advancements, the shared task has unveiled persistent challenges that define the roadmap for future research. Issues such as the residual leakage of source speaker characteristics and the performance degradation under extreme dialectal variations remain primary bottlenecks for real-world application. Moving forward, our research will prioritize the exploration of zero-shot adaptation techniques and the improvement of model robustness in heterogeneous, low-resource environments. We believe that the methodologies and benchmarks established through the VLSP 2025 VC task will serve as a vital catalyst for the community, driving the evolution of voice conversion technologies towards greater cultural authenticity and practical utility.

References

[1] S. Hussain, P. Neekhara, J. Huang, J. Li, B. Ginsburg, Ace-vc: Adaptive and controllable voice conversion using explicitly disentangled self-supervised speech representations (2023).

[2] L. Sun, K. Li, H. Wang, S. Kang, H. Meng, Phonetic posteriorgrams for many-to-one voice conversion without parallel data training, in: Proceedings Of 2016 IEEE International Conference On Multimedia And Expo (ICME), 2016, pp. 1–6.

[3] J. Li, W. Tu, L. Xiao, Freevc: Towards high-quality text-free one-shot voice conversion (2022).

[4] Y. H. Chen, D. Y. Wu, T. H. Wu, H. Lee, Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization (2020).

[5] H. T. Tu, L. T. Long, V. Huan, N. T. P. Thao, N. V. Thang, N. T. Cuong, N. T. T. Trang, Voice conversion for low-resource languages via knowledge transfer and domain-adversarial training, in: Proceedings Of ICASSP 2025 - 2025 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP), 2025, pp. 1–5.

[6] H. He, Z. Shang, C. Wang, X. Li, Y. Gu, H. Hua, L. Liu, C. Yang, J. Li, P. Shi, Y. Wang, K. Chen, P. Zhang, Z. Wu, Emilia: A large-scale, extensive, multilingual, and diverse dataset for speech generation (2025).

[7] J. W. Jung, W. Zhang, S. Maiti, Y. Wu, X. Wang, J. H. Kim, Y. Matsunaga, S. Um, J. Tian, H. J. Shim, N. Evans, J. S. Chung, S. Takamichi, S. Watanabe, Text-to-speech synthesis in the wild (2025).

[8] Silero Team, Silero vad: Pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier, gitHub Repository (2024).

[9] J. Kormos, M. Dénes, Exploring measures and perceptions of fluency in the speech of second language learners, System, Vol. 32, No. 2, 2004, pp. 145–164.

[10] A. Defossez, G. Synnaeve, Y. Adi, Real time speech enhancement in the waveform domain (2020).

[11] H. Schröter, A. N. Escalante-B., T. Rosenkranz, A. Maier, Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering (2022).

[12] G. Mittag, B. Naderi, A. Chehadi, S. Möller, Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets, in: Proceedings Of Interspeech 2021, 2021, pp. 2127–2131.

[13] P. Andreev, A. Alanov, O. Ivanov, D. Vetrov, Hifi++: A unified framework for bandwidth extension and speech enhancement, in: Proceedings Of ICASSP 2023 - 2023 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP), 2023, pp. 1–5.

[14] M. Bain, J. Huh, T. Han, A. Zisserman, Whisperx: Time-accurate speech transcription of long-form audio (2023).

[15] D. H. Cho, H. S. Oh, S. B. Kim, S. H. Lee, S. W. Lee, Emosphere-tts: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech, in: Proceedings Of Interspeech 2024, 2024, pp. 1810–1814.

[16] R. Shimizu, R. Yamamoto, M. Kawamura, Y. Shirahata, H. Doi, T. Komatsu, K. Tachibana, Prompttts++: Controlling speaker identity in prompt-based text-to-speech using natural language descriptions (2023).

[17] B. Desplanques, J. Thienpondt, K. Demuynck, Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification, in: Proceedings Of Interspeech 2020, 2020, pp. 3830–3834.

[18] A. Nagrani, J. S. Chung, A. Zisserman, Voxceleb: A large-scale speaker identification dataset, in: Proceedings Of Interspeech 2017, 2017, pp. 2616–2620.

[19] J. S. Chung, A. Nagrani, A. Zisserman, Voxceleb2: Deep speaker recognition, in: Proceedings Of Interspeech 2018, 2018, pp. 1086–1090.

[20] K. Le, T. V. Ho, D. Tran, D. T. Chau, Chunkformer: Masked chunking conformer for long-form speech transcription (2025).

[21] S. Liu, Zero-shot voice conversion with diffusion transformers, arXiv Preprint arXiv:2411.09943 (2024).

[22] Y. Chen, Z. Niu, Z. M. sneak, K. Deng, C. Wang, J. Zhao, K. Yu, X. Chen, F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching (2025).

[23] V. T. Pham, X. T. H. Nguyen, V. Hoang, T. T. T. Nguyen, Vietnam-celeb: A large-scale dataset for vietnamese speaker recognition, in: Proceedings Of Interspeech 2023, 2023, pp. 1918–1922, (in Vietnamese).

[24] A. Lobashev, A. Yermekova, M. Larchenko, Training-free voice conversion with factorized optimal transport, in: Proceedings Of Interspeech 2025, 2025, pp. 1373–1377.

[25] H. Zhu, W. Kang, Z. Yao, L. Guo, F. Kuang, Z. Li, W. Zhuang, L. Lin, D. Povey, Zipvoice: Fast and high-quality zero-shot text-to-speech with flow matching (2025).

[26] T. T. Le, L. T. Nguyen, D. Q. Nguyen, Phowhisper: Automatic speech recognition for vietnamese, in: Proceedings Of The Second Tiny Papers Track At ICLR 2024, 2024, (in Vietnamese).

[27] H. Wang, S. Zheng, Y. Chen, L. Cheng, Q. Chen, Cam++: A fast and efficient network for speaker verification using context-aware masking (2023).

[28] S. G. Lee, W. Ping, B. Ginsburg, B. Catanzaro, S. Yoon, Bigvgan: A universal neural vocoder with large-scale training, in: Proceedings Of The Eleventh International Conference On Learning Representations (ICLR 2023), 2023.

[29] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, H. Saruwatari, Jvs corpus: Free japanese multi-speaker voice corpus (2019).

[30] Zeroth Project Contributors, Zeroth-korean: Korean open speech dataset, open-source Korean Speech Corpus For ASR Research (2017).

[31] T. Vu, L. T. Nguyen, D. Q. Nguyen, Zero-shot text-to-speech for vietnamese, in: Proceedings Of ACL 2025, 2025, (in Vietnamese).

[32] M. Baas, B. v. Niekerk, H. Kamper, Voice conversion with just nearest neighbors, in: Proceedings Of Interspeech 2023, 2023, pp. 1–5.