VLSP 2025 ASR/SER: Vietnamese Speech Recognition and Speech Emotion Recognition Challenge: Technical Analysis and Insights

Manh Hai Cao[1], Chi Dung Hoang[1], Quang Trung Le[2], Van Hai Do[3],
[1]NamiTech, [2]Torilab, [3]Thuyloi University
Correspondence: haidv@tlu.edu.vn

Abstract

The VLSP 2025 ASR/SER Challenge introduced a joint evaluation framework requiring systems to simultaneously perform Vietnamese speech recognition and emotion classification. This paper analyzes the technical approaches of the top seven teams, examining innovations in data curation, model architecture, and fusion strategies. The winning approach achieved 9.07% Syllable Error Rate for ASR and 82.21% accuracy for SER through custom Zipformer architecture and sophisticated data augmentation. Key findings highlight the importance of data quality over quantity, effectiveness of multi-modal fusion, and Vietnamese-specific challenges including tonal characteristics and dialectal variations.

1 Introduction

Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER) have emerged as fundamental components of modern human-computer interaction systems (Schuller et al., 2018). The VLSP 2025 challenge marked a significant advancement in Vietnamese speech processing by requiring unified systems capable of simultaneous transcription and emotion classification, reflecting real-world applications where understanding both semantic content and emotional context is crucial (Busso et al., 2008).

Unlike previous competitions focusing on individual tasks, this challenge employed a joint evaluation framework that encouraged participants to develop holistic approaches. The competition allowed participants to leverage external resources including pre-trained models and open datasets, reflecting current trends toward foundation model adaptation (Radford et al., 2023).

The evaluation framework employed Syllable Error Rate ($SyER$) for speech recognition and classification accuracy (ACC) for binary emotion recognition (neutral vs. negative), with a composite score balancing both tasks. Seven teams achieved competitive performance through diverse technical approaches, providing valuable insights for Vietnamese speech processing advancement.

2 Competition Results

Table 1 presents the final competition results. The performance distribution reveals several interesting patterns: while hynguyenthien achieved the best overall score through balanced performance across both tasks, CodeSERSai obtained the highest SER accuracy (85.79%) but ranked fifth overall due to significantly higher ASR error rate (25.22%).

Rank	Team	SyER	SER	Score
1	hynguyenthien	**9.07**	82.21	**88.31**
2	ishowspeech	11.38	79.13	85.77
3	dangnguyen-VLSP	12.66	80.84	85.39
4	SoFarSoGood	19.12	79.50	80.47
5	CodeSERSai	25.22	**85.79**	78.08
6	SoulSound	20.87	66.50	75.34
7	nhitny	23.56	71.76	75.04

Table 1: VLSP 2025 ASR-SER Challenge Results (%)

The results demonstrate a clear trade-off between ASR and SER performance. Teams achieving sub-13% SyER (top 3) maintained SER accuracy between 79-82%, while teams with higher SER accuracy often struggled with ASR performance. This suggests that joint optimization remains challenging, requiring careful balance between competing objectives.

3 Technical Approaches

The competition received submissions from 7 teams in total, with 5 teams providing technical reports. Below is a detailed analysis of each team's solution approach.

3.1 Winner: hynguyenthien - Custom Zipformer Architecture

The winning team developed a comprehensive pipeline combining sophisticated data curation, custom architecture design, and multi-modal fusion. Their approach achieved 9.07% SyER for ASR and 82.21% accuracy for SER through three key innovations.

Voting-based Data Curation: The team curated 8 Vietnamese corpora totaling 4,000 hours, with a novel pseudo-labeling mechanism for VLSP2023-D1+3+4 dataset. They developed a text normalization module converting written forms to spoken forms (e.g., "1" $\rightarrow$ "mt", "3km" $\rightarrow$ "ba km") and fine-tuned two ASR models (Wav2Vec-250h and Whisper-Small) on 300 hours of sampled data. To address fast speech issues, they applied 0.9× speed factor and performed voting between models with WER threshold of 10% and 4-gram language model scoring, yielding 180 hours of high-quality pseudo-labeled data.

Custom Zipformer Training: Unlike other teams using pre-trained models, they trained a 30-million parameter Zipformer (Yao et al., 2023) from scratch to avoid domain mismatch. The architecture employs gated linear attention and hierarchical time-reduction for efficiency. Training used joint CTC-RNN-T loss on 4,000 hours for 50 epochs with BPE tokenizer (vocabulary 2,048) and modified beam search decoding (beam size 15).

Multi-modal SER Fusion: For emotion recognition, they filtered VLSP2023-Dataset4 using emotion2vec-finetuning-large and concatenated Wav2Vec-250h-ft-300h (512-dim) with emotion2vec-base (768-dim) embeddings into 1,280-dimensional features. SpeechFormer++ (Chen et al., 2022) classifier with hierarchical structure captured both fine and coarse-grained emotional patterns through 25 epochs of binary classification training.

This integrated approach demonstrated that custom architectures with careful data engineering and feature fusion significantly outperform fine-tuned foundation models for Vietnamese speech processing.

3.2 Runner-up: ishowspeech - Multi-Stage Training Framework

The second-place team developed a comprehensive three-stage training framework that systematically exploits diverse open-source datasets with varying quality conditions. Their approach achieved 11.38% SyER and 79.13% SER accuracy through strategic multi-stage optimization and advanced inference techniques.

Stage 1 - ASR Training with Quality Filtering: The team began with comprehensive ASR model training using all available Vietnamese datasets totaling over 3,000 hours: VLSP2023 (300h), phoaudiobook (1,494h), vivoice (1,000h), vietbud500 (500h), 28kVigBigData (460k utterances), ViMD (100h), and VIVOS (15h). They employed joint CTC-RNN-T training with combined loss: $L_{ASR} = \lambda L_{RNN-T} + (1 - \lambda)L_{CTC}$.

A key innovation was their training-loop data filtering strategy inspired by OWSM approaches (Peng et al., 2025). They iteratively trained ASR models, evaluated transcription quality, and filtered samples using dual criteria: CTC confidence scores and WER thresholds. Samples were retained only if they satisfied: $0.3 < CTC_{conf} < 0.95$ and $5\% < WER < 40\%$. This approach removed both overly simplistic samples (low WER + high confidence) and corrupted data (high WER + low confidence), retaining moderately challenging samples for effective learning.

Stage 2 - SER Training with Hybrid Loss: For emotion recognition, they leveraged both Vietnamese (ViSEC with 5,400 utterances) and international datasets (IEMOCAP, EMODB, RAVDESS, CREMA-D, EmoV-DB) totaling over 30,000 emotional utterances. They systematically evaluated various pre-trained models including WavLM, Emotion2Vec (Ma et al., 2023), and Wav2Vec2.0, finding optimal performance using layer 9/12 features from Wav2Vec2.0.

Their hybrid loss function combined cross-entropy with supervised contrastive learning (Khosla et al., 2020). This formulation encouraged similar emotional states to have similar representations while pushing different emotions apart in the embedding space, improving discriminative capabilities across diverse data sources.

Stage 3 - Advanced Inference Optimization: The final stage implemented sophisticated inference strategies combining multiple prediction sources. They employed cross-attention fusion mechanisms to integrate acoustic and semantic information from multiple pre-trained models (WavLM, Wav2Vec2.0, Whisper, Emotion2Vec, SenseVoice, HuBERT), addressing the challenge of ASR errors degrading SER performance.

For robust inference, they implemented k-

NN interpolation strategy creating an embedding database from training data: $(K, V) = \{(x_i, y_i), i \in D\}$. The final prediction interpolated between neural model and k-NN predictions: $p(y|x) = \beta p_{model}(y|x) + (1-\beta)p_{knn}(y|x)$, where β was optimized based on validation performance.

This multi-stage methodology effectively handled heterogeneous data quality while maintaining model robustness, demonstrating that systematic training strategies can achieve competitive performance even with diverse, lower-quality datasets.

3.3 Third Place: dangnguyen-VLSP - FastConformer Efficiency

The third-place team developed "Twinkle ASR," prioritizing deployment efficiency through FastConformer architecture (Rekesh et al., 2023), achieving 12.66% WER and 80.84% SER accuracy with 2.8× speedup over standard Conformer.

Efficient Architecture: FastConformer employs 8× convolutional subsampling, depthwise-separable convolutions, and hybrid local-global attention, delivering 2.7× faster inference while supporting long sequences up to several hours. The unified pipeline combines frozen Emotion2Vec features with FastConformer processing through stacked blocks containing feed-forward modules, multi-head self-attention, and convolutional components.

Emotion-Aware Processing: Their preprocessing maintained 4:1 ratio of short (0-5s) to medium (5-10s) utterances matching test distribution. Emotion-driven augmentation applied speed adjustment for angry speech, volume scaling for intensity, and pitch shifting for emotional frequency variations. They incorporated real-world noise including traffic and market sounds for robustness.

Rule-Based Fusion: A key innovation was linguistic post-processing analyzing transcribed text for Vietnamese emotion cues. By detecting offensive words ("tao", "may") associated with anger, they improved emotion accuracy from 80% to 82% through simple rule-based reassignment after initial ASR/SER processing.

This efficiency-focused approach demonstrated that lightweight architectures with strategic processing can achieve competitive performance suitable for resource-constrained deployment scenarios.

3.4 Fourth Place: SoFarSoGood - Data-Centric Methodology

The fourth-place team achieved 19.12% WER and 79.5% SER accuracy through systematic data preprocessing and progressive Whisper fine-tuning.

Data-Centric Approach: They assembled 2,500+ hours from five Vietnamese datasets (VLSP 2023, phoaudiobook, Viet_Bud500, viVoice, VinBigData) and identified numerous quality issues through exploratory data analysis. Conservative preprocessing prioritized quality over quantity: removing samples with digits/non-Latin characters, filtering by duration (shorter than 15s and less than 85 words), and requiring Vietnamese tone marks.

Two-Phase Training: Whisper Small underwent progressive adaptation - Phase 1 on large-scale general data (1.2M samples from viVoice/phoaudiobook, then Viet_Bud500/VinBigData) for broad Vietnamese knowledge, followed by Phase 2 in-domain training on VLSP 2023 for target characteristics.

Hybrid SER: Used frozen Emotion2Vec_plus_large (768-dim embeddings) with lightweight neural classifier, decoupling representation learning from task-specific classification for scarce labeled emotion data.

This methodology demonstrated effective foundation model adaptation through systematic data cleaning and progressive domain specialization.

3.5 Seventh Place: NhiTNY - DFAT Hybrid Fusion Pipeline

The seventh-place team achieved 84.38% SER accuracy through DFAT (Dual-stage Fusion of Acoustic and Text features), combining early and late fusion with lightweight ensemble learning.

Dual-Stream Features: Their pipeline extracted 1024-dimensional acoustic features via SEFE (WavLM/Emotion2Vec) and 1024-dimensional textual features via TEFE (neural networks processing Whisper-small ASR transcripts trained on ViSEC dataset).

Hybrid Fusion: Early fusion concatenated SEFE-TEFE features (2048-dim), processed by three diverse classifiers: Logistic Regression, Random Forest, and Optuna-optimized XGBoost. Late fusion used weighted ensemble: $P_{final} = w_1 P_{XGB} + w_2 P_{RF} + w_3 P_{LR}$ with optimized weights.

This approach outperformed early fusion (81.31%), late fusion (81.40%), and unimodal base-

lines (acoustic: 74.58%, text: 74.63%), demonstrating effective multimodal integration for Vietnamese SER.

4 Technical Innovations and Contributions

The VLSP 2025 competition showcased several significant technical innovations that advance the state-of-the-art in Vietnamese speech processing. These contributions span multiple aspects of system design, from data handling strategies to novel architectural approaches and training methodologies.

Data-centric methodologies emerged as a dominant theme across multiple successful teams. The winning team's large-scale data curation approach with strict filtering criteria demonstrates the critical importance of data quality in Vietnamese speech processing. These approaches recognize that in low-resource language scenarios, careful attention to data selection, cleaning, and augmentation can provide substantial performance improvements that may exceed those achieved through architectural innovations alone.

Multi-stage training frameworks represented another significant innovation area. The ishowspeech team's multi-stage framework and SoFarSoGood's progressive two-phase training both address the challenge of effectively leveraging diverse datasets with varying quality conditions. These approaches demonstrate sophisticated understanding of how to structure training procedures to maximize knowledge transfer while minimizing negative interference from heterogeneous data sources.

Fusion strategy innovations, particularly NhiTNY's DFAT approach, represent notable advances in multimodal speech processing. Their hybrid fusion methodology demonstrates that combining early and late fusion strategies can significantly outperform traditional approaches, particularly important for low-resource languages where training data limitations make robust fusion approaches especially valuable. The quantitative results provide strong empirical evidence for the effectiveness of sophisticated multimodal integration strategies.

Efficiency and lightweight design considerations, exemplified by the dangnguyen team's Twinkle-ASR framework, address practical deployment concerns that are often overlooked in research-focused competitions. Their approach demonstrates that efficiency-focused designs can remain competitive while providing significant advantages for practical deployment scenarios.

5 Discussion and Future Directions

The VLSP 2025 competition results reveal several important trends and insights that have broader implications for Vietnamese speech processing research and development. The diversity of successful approaches suggests that multiple viable paths exist for advancing Vietnamese speech technology, each with distinct advantages and suitable application scenarios.

The emphasis on data quality over quantity across multiple teams suggests a maturation in the field's understanding of effective training strategies for low-resource languages. Rather than simply seeking larger datasets, successful teams focused on careful curation, quality control, and strategic utilization of available data resources. This trend indicates that future research should continue to emphasize sophisticated data handling strategies rather than relying solely on scale.

Strategic model adaptation emerged as a crucial factor for success, with multiple teams developing specialized approaches for addressing Vietnamese speech characteristics. This finding suggests that language-specific adaptation strategies are essential for achieving optimal performance, rather than simply applying general-purpose models without modification. Future research should continue to explore Vietnamese-specific adaptation techniques that address dialectal diversity, tonal characteristics, and cultural factors.

The effectiveness of multimodal integration, particularly demonstrated by NhiTNY's hybrid fusion approach, indicates significant potential for advanced multimodal techniques in Vietnamese speech emotion recognition. Future research should explore more sophisticated fusion strategies that can effectively combine acoustic, linguistic, and potentially visual modalities for comprehensive speech understanding.

Practical deployment considerations, highlighted by the success of lightweight approaches, reflect the growing importance of developing solutions that can be effectively deployed in real-world scenarios. Future research should continue to balance performance optimization with efficiency requirements, ensuring that advances in Vietnamese speech processing can be translated into practical

applications.

6 Conclusion

The VLSP 2025 shared task demonstrated diverse and innovative approaches to Vietnamese speech processing challenges. Key contributions include data-centric methodologies, multi-stage training frameworks, hybrid fusion strategies, and efficiency-focused designs. These findings highlight the importance of data quality, language-specific adaptation, and practical deployment considerations. Future research should continue to explore these directions to advance Vietnamese speech technology.

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. 2022. Speechformer: A hierarchical efficient framework incorporating the characteristics of speech. *arXiv preprint arXiv:2203.03812*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2023. Emotion2vec: Self-supervised pre-training for speech emotion recognition. *Computing Research Repository*, arXiv:2312.15185. Version 1.

Yifan Peng, Shakeel Muhammad, Yui Sudo, William Chen, Jinchuan Tian, Chyi-Jiunn Lin, and Shinji Watanabe. 2025. Owsm v4: Improving open whisper-style speech models via data scaling and cleaning. *arXiv preprint arXiv:2506.00338*.

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

Dima Rekesh, Somshubra Majumdar, Vahid Noroozi, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. *Computing Research Repository*, arXiv:2305.05084. Version 1.

Björn Schuller, Stefan Steidl, Anton Batliner, Simone Hantke, Florian Hönig, Juan Rafael Orozco-Arroyave, Elmar Nöth, Yue Zhang, and Felix Weninger. 2018. The interspeech 2018 computational paralinguistics challenge: Atypical and self-assessed affect, crying and heart beats. In *Interspeech*, pages 122–126.

Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey. 2023. Zipformer: A faster and better encoder for automatic speech recognition. *Computing Research Repository*, arXiv:2310.11230. Version 1.