A Comprehensive Pipeline for Vietnamese Speech Recognition and Emotion Recognition

Hy Nguyen Thien
AI Engineer
hynguyenthien@gmail.com

Abstract

We present a comprehensive pipeline for Vietnamese Speech Recognition (ASR) and Speech Emotion Recognition (SER). Our contributions include: (1) a large-scale data curation and augmentation strategy combining multiple corpora, (2) a robust filtering pipeline using multi-model voting and n-gram scoring to construct an additional high-quality dataset, (3) a Zipformer-based ASR model trained from scratch with joint CTC and RNN-T loss on 4000 hours of augmented speech, and (4) a feature-fusion approach for SER leveraging Wav2Vec and Emotion2Vec embeddings with SpeechFormer++. Our systems achieve state-of-the-art performance on VLSP 2025 benchmarks, demonstrating the effectiveness of our methods for both ASR and SER in Vietnamese.

1 Introduction

Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER) are fundamental components of spoken language understanding and play a critical role in applications such as conversational agents, call centers, and human-computer interaction. Recent advances in self-supervised learning (SSL) frameworks, such as Wav2Vec (Baevski et al., 2019), Whisper (Radford et al., 2023) or WavLM (Chen et al., 2022), have demonstrated remarkable improvements in ASR across multiple languages. However, Vietnamese remains under-resourced compared to high-resource languages like English and Mandarin, which results in challenges related to limited vocabulary coverage, out-of-vocabulary (OOV) occurrences, and domain mismatch across datasets.

On the other hand, Speech Emotion Recognition has become increasingly important for affective computing and empathetic conversational systems. Despite progress using deep learning architectures and multimodal fusion (He et al., 2023), SER still suffers from noisy or weakly supervised emotion labels, especially in low-resource languages like Vietnamese. These limitations hinder the generalization and robustness of deployed systems in real-world environments such as noisy call centers and spontaneous dialogues.

In this paper, we present a comprehensive pipeline for Vietnamese ASR and SER. For ASR, we introduce dataset filtering and augmentation strategies that prioritize rare OOV tokens to improve lexical coverage, combined with multi-model voting to refine transcription quality. We then train a 30M-parameters Zipformer (Yao et al., 2024) model on 4000 hours of augmented speech data. For SER, we investigate the joint Wav2Vec and Emotion2Vec (Ma et al., 2023) representations within a shared feature space, combined with the SpeechFormer++ (Chen et al., 2023) architecture, aiming to improve robustness against label noise and variability in emotional expressions. Our approach advances spoken language understanding in Vietnamese and offers insights that may extend to other low-resource languages.

2 Data Statistics Overview

We curate and filter 8 Vietnamese corpora. Table 1 summarizes dataset statistics. The **8th dataset**, **VLSP2023-D1+3+4-Voting**, was constructed based on a model voting mechanism. The details of this procedure will be elaborated in Section 3. Figure 1 visualizes OOV and vocabulary size per dataset. Figure 2 and 3 visualizes top 20 vi-words, OOV per dataset.

3 Data Curating and Augmentation

Voting-based Pseudo Labeling. We employed a model voting mechanism to generate pseudo-labels for the **VLSP2023-D1+3+4** dataset, which originally contained 245 hours of unlabeled speech. First, we developed a text normalization module that converts written forms into spoken forms. For

Dataset	Hours	OOV	Vocab (Vi)
PhoAudioBook	1494	4980	5640
VietBud500	500	43	5322
ViMD	100	764	3927
ViVoice	1000	33867	8332
28k-Vietnamese	50	1022	4867
VIVOS	15	5	4924
VLSP2023-D2	60	602	3170
VLSP2023-D1+3+4-Voting	180	1745	5750

Table 1: Statistics of curated datasets.

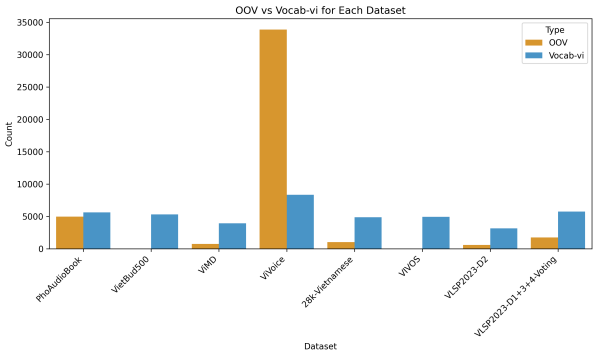

Figure 1: OOV and Vietnamese vocabulary counts per dataset.

Figure 2: Top 20 vi-words per dataset.

example, *"1"* is normalized to *"một"*, *"chương iii"* to *"chương 3"*, and *"3km"* to *"ba km"*. Such written forms frequently appear in the ViVoice dataset. Due to the inherent ambiguity of spoken language—for instance, *"113"* can be read as either *"một một ba"* or *"một trăm mười ba"*—we allowed for a certain degree of error. We refer to this module as **TN1**. In the following step, we sampled **300 hours** from the initial 7 datasets (was normalized through the TN1 module) to ensure sufficient coverage of both the top 20 most frequent Vi-words and the OOV set. This subset was then used to fine-tune two ASR models for about **10 epochs**: **Wav2Vec-250h** and **Whisper-Small**.

For Wav2Vec, we additionally constructed a 6-gram language model for decoding and a 4-gram language model for scoring, both trained on transcriptions from the first seven datasets. After fine-tuning, the resulting systems were denoted as **STT1** (Wav2Vec-250h-ft-300h) and **STT2** (Whisper-Small-ft-300h). During our experiments, we found that both **STT1** and **STT2** often missed words in audio with relatively fast speaking rates. To address this, we slowed down all audios in the **VLSP2023-D1+3+4** dataset using a speed factor of 0.9, resulting in **VLSP2023-D1+3+4-Speed0.9**. We then performed voting between **STT1** and **STT2** outputs on the modified dataset, applying a WER threshold of 10% together with 4-gram LM scoring (threshold -5.0). We prioritized the selection of audio–transcript pairs after voting, using the outputs of the Whisper-Small-ft-300h model as the criterion. This process yielded the **8th VLSP2023-D1+3+4-Voting** dataset, comprising **180 hours** of pseudo-labeled speech.

Augmentation. We employed multiple augmentation strategies, including noise injection, low-bitrate simulation, pitch shifting, and speech permutation. To ensure fairness, the proportion of augmented speech was balanced across the different augmentation methods. Augmentation was selectively applied to audio samples containing rare OOV tokens in seven datasets, while the 8th dataset (VLSP2023-D1+3+4-Voting) remained unaltered. Through this process, we obtained approximately **4000 hours** of training data.

4 ASR Model Training

The Wav2Vec model demonstrates two primary limitations. First, its reliance on character-level tokens for CTC decoding renders transcription accuracy highly dependent on the performance of the accompanying n-gram language model. Second, fine-tuning a pretrained Wav2Vec model without explicit knowledge of the pretraining data distribu-

Figure 3: Top 20 OOV per dataset.

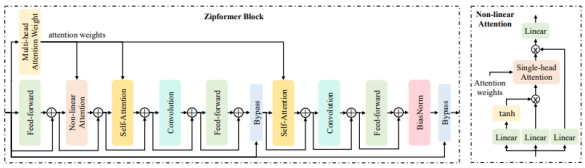

Figure 4: (Left): Zipformer block structure. (Right): Non-Linear Attention module structure.

tion often results in overfitting to the target fine-tuning domain and reduced robustness when applied across diverse domains. Similarly, the Whisper model exhibits several weaknesses. It tends to produce a high rate of hallucinations in noisy audio conditions, thereby compromising the reliability of transcriptions. Moreover, training Whisper necessitates extremely large-scale computational resources, which poses a barrier to its broader applicability. Motivated by these reasons, we chose to train a **30M parameters Zipformer** model from scratch. Compared with the conventional Conformer (Gulati et al., 2020), Zipformer (Figure 4) achieves higher recognition accuracy while using fewer parameters, thanks to its gated linear attention and hierarchical time-reduction mechanism. This design not only reduces computational cost and decoding latency but also makes the model more robust to noisy and spontaneous speech. Recent studies show that combining CTC (Graves et al., 2006) with RNN-T (Kuang et al., 2023) loss stabilizes encoder training and accelerates convergence, while simultaneously improving recognition accuracy compared to RNN-T alone. The experiments conducted by the authors in (Yao et al., 2023) show that, this joint objective leverages the strong alignment signal from CTC and the streaming capability of RNN-T, making it highly effective for large-scale ASR systems. The basic training configurations of our model are as follows:

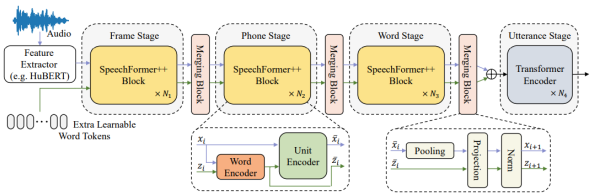

Figure 5: SpeechFormer++ Arch.

- BPE tokenizer (Kozma and Voderholzer, 2024) with a vocabulary size of 2048.
- Joint CTC and RNN-T loss.
- Trained on **4000h augmented data** for 50 epochs using 4 × RTX 3090 GPUs.
- Decoding is performed using the **modified beam search** (Kang et al., 2023) algorithm with a beam size of 15.

5 SER Model Training

We filtered **VLSP2023-Dataset4** with low quality emotion label using emotion2vec-finetuning-large, obtaining a cleaner subset. Final SER training data included **VLSP2023-Dataset3 + VLSP2023-Dataset4-filtered**.

Feature fusion. We concatenated Wav2Vec-250h-ft-300h (dim-512) and emotion2vec-base (dim-768) embeddings into a **dim-1280** feature vector.

Classifier. SpeechFormer++ (Figure 5) leverages the hierarchical structure of speech with unit encoders and merging blocks, effectively capturing both fine and coarse-grained information. On SER tasks (IEMOCAP & MELD), it outperforms the standard Transformer while significantly reducing computational cost. In this study, SpeechFormer++ (SF2) was trained for 25 epochs using binary classification labels (negative vs. neutral). The training employed a batch size of 32, a learning rate of 0.0001, and cross-entropy loss as the optimization objective. The dataset was divided into training and validation sets with a 9:1 ratio.

6 Results

6.1 ASR

As shown in Table 2, the Wav2Vec-250h models combined with a 6-gram LM achieve moderate performance with WERs ranging from 17.8% to 22.7%. Fine-tuning the Wav2Vec trained on 250h with an additional 300h of in-domain data significantly improves the results, reducing the WER by around 4–5 points compared to the baseline Wav2Vec-250h model.

For Whisper-Small, the performance is considerably worse, with WERs over 50%, even after fine-tuning with 300h of data (WER $\approx$ 20%). This degradation mainly comes from hallucinations, a common issue for Whisper in low-resource and domain-mismatched settings. For example, in some audios Whisper-Small produces irrelevant phrases such as *"hãy subscribe cho kênh ghiền mì gõ để không bỏ lỡ..."* or repeats words and phrases multiple times (e.g., *"mày cứt đặt chân đất phải dương mấy mấy mấy mấy mấy mấy mấy ..."*). These errors lead to extremely high WER compared with other models.

The best results are obtained with the **Zipformer + CTC-RNN-T** trained from scratch on 4000 hours of data, achieving WERs below 10% across all test sets. This demonstrates the importance of large-scale and well-augmented training data, where the Zipformer model with only 30M parameters can outperform both large pretrained models (Wav2Vec and Whisper) and fine-tuned versions. The results highlight that careful data curation and augmentation strategies, together with an appropriate architecture and training objective (CTC+RNNT), are critical to achieving state-of-the-art performance in VLSP 2025 benchmarks ASR.

Model	Pr23	Pb25	Pr25
W2V-250h + 6-gram LM	22.7	22.7	21.9
W2V-250h-ft-300h + 6-gram LM	18.6	18.4	17.8
Whisper-S	50.20	57.60	55.32
Whisper-S-ft-300h	22.2	21.8	20.05
Zipformer + CTC-RNNT + 4000h	**9.62**	**9.54**	**9.07**

Table 2: ASR results (WER%).

6.2 SER

As shown in Table 3, using only mel spectrogram features (Mel+SF2) or a pretrained Wav2Vec model (W2V-Base+SF2) yields comparable accuracies in the range of 76–87%, with limited improvements when fine-tuning Wav2Vec-250h on 300 hours data. In contrast, incorporating **Emotion2Vec** representations into the W2V-250h-ft-300h system achieves a substantial performance boost, reaching 90.24% on Pr23, 82.83% on Pb25, and 82.21% on Pr25. These results highlight the effectiveness of combining self-supervised speech representations (Wav2Vec) with emotion-oriented embeddings (Emotion2Vec), which capture prosodic and affective cues that conventional spectral or linguistic features often miss. The synergy between the two feature types enables more robust modeling of emotional expressions in speech, leading to significantly higher SER accuracy across all test sets.

Model	Pr23	Pb25	Pr25
Mel+SF2	86.76	76.36	77.64
W2V-Base+SF2	86.68	76.79	76.67
W2V-250h-ft-300h+SF2	87.20	78.93	78.33
W2V-250h-ft-300h+emo2vec+SF2	**90.24**	**82.83**	**82.21**

Table 3: SER results (accuracy%).

6.3 ASR + SER

According to the Final Score definition

$$Score = 0.7 \times (1 - WER_{ASR}) + 0.3 \times ACC_{SER}$$

ASR accuracy (reflected by $1 - WER$) contributes 70% to the final ranking, while SER accuracy contributes 30%. As shown in Table 4, the top-ranked system (*hynguyenthien - ours*) achieved the lowest WER (9.07%) together with high SER accuracy (82.21%), resulting in the best Final Score of 88.31. Although *CodeSERSai* obtained the highest SER accuracy (85.79%), its much higher WER (25.22%) significantly reduced the Final Score (78.08). This contrast clearly demonstrates that lowering WER is more influential for the overall evaluation metric than maximizing SER alone. The systems *ishowspeech* and *dangnguyen-VLSP* also illustrate this balance, achieving competitive rankings with moderate SER accuracy but relatively low WER values. Overall, these results confirm that improvements in ASR (WER reduction) play the dominant role in boosting the Final Score, while SER accuracy provides an additional but smaller contribution. Notably, the system of *hynguyenthien* consistently optimized both WER and SER, leading to a clear margin over the second and third ranked teams.

Team	WER	SER Acc	Final Score
hynguyenthien	**9.07**	82.21	**88.31**
ishowpeech	11.38	79.13	85.77
dangnguyen-VLSP	12.66	80.84	85.39
SoFarSoGood	19.12	79.50	80.47
CodeSERSai	25.22	**85.79**	78.08
SoulSound	20.87	66.50	75.34
nhitny	23.56	71.76	75.04

Table 4: Leaderboard results on VLSP-2025 private test dataset.

7 Conclusion

We proposed a comprehensive pipeline for Vietnamese ASR and SER. Our Zipformer-30M-parameters model achieved state-of-the-art WER on the VLSP 2025 benchmarks, while the integration of Wav2Vec and Emotion2Vec representations significantly improved SER accuracy. Looking ahead, we plan to explore multilingual expansion and to unify ASR and SER into a single, multimodal framework, potentially integrating additional components such as gender recognition, regional accent identification, and other paralinguistic features. This approach aims to enhance the efficiency and robustness of spoken language understanding systems, with potential applications in conversational AI and other low-resource language scenarios.

References

Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Proceedings of Interspeech*.

Sanyuan Chen, Chengyi Wang, Zhuo Chen, Yu Wu, Jia Jia, Xie Chen, Zejun Ma, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2970–2984.

Wei Chen, Xinxin Xing, Xiaoxiao Xu, Jiale Pang, and Lei Du. 2023. Speechformer++: A hierarchical efficient framework for paralinguistic speech processing. In *Proceedings of Interspeech*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 369–376.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhifeng Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of Interspeech*.

Jing He, Xinyuan Shi, Xiaohai Li, and Tomoki Toda. 2023. Mf-aed-aec: Speech emotion recognition by leveraging multimodal fusion, asr error detection, and asr error correction. In *Proceedings of ICASSP*.

Wenxin Kang, Long Guo, Fangjie Kuang, Liyong Lin, Min Luo, Zengwei Yao, Xiaoyi Yang, Piotr Żelasko, and Daniel Povey. 2023. Fast and parallel decoding for transducer. In *Proceedings of ICASSP*.

László Kozma and Johannes Voderholzer. 2024. Theoretical analysis of byte-pair encoding. *arXiv preprint arXiv:2402.10234*.

Fangjie Kuang, Long Guo, Wenxin Kang, Liyong Lin, Min Luo, Zengwei Yao, and Daniel Povey. 2023. Pruned rnn-t for fast, memory-efficient asr training. In *Proceedings of ICASSP*.

Zeyu Ma, Zhaoyu Zheng, Jun Ye, Jialin Li, Zhen Gao, Shiyu Zhang, and Xun Chen. 2023. Emotion2vec: Self-supervised pre-training for speech emotion representation. In *Proceedings of ICASSP*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Whisper: Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Zengwei Yao, Long Guo, Xiaoyi Yang, Wenxin Kang, Fangjie Kuang, Yifan Yang, Zhiyong Jin, Liyong Lin, and Daniel Povey. 2024. Zipformer: A faster and better encoder for automatic speech recognition. In *Proceedings of ICASSP*.

Zengwei Yao, Wenxin Kang, Xiaoyi Yang, Fangjie Kuang, Long Guo, Hao Zhu, Zhiyong Jin, Zheng Li, Liyong Lin, and Daniel Povey. 2023. Cr-ctc: Consistency regularization on ctc for improved speech recognition. In *Proceedings of Interspeech*.