Twinkle-ASR: A lightweight and sufficient framework for the ASR shared task at VLSP 2025

Dang Nguyen Pham
Independence Researcher
dangnguyen667201@gmail.com

Viet Tien Pham
Independence Researcher
phamviettien130102@gmail.com

Minh Nghia Vu
Independence Researcher
vuminhnghia.work@gmail.com

Minh Nguyen Le
Independence Researcher
leminhnguyen.mywork@gmail.com

Abstract

The VLSP community brings together research groups from both academia and industry in the field of Vietnamese language and speech processing. One of the key shared tasks organized at the 11th Workshop on Vietnamese Language and Speech Processing was Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER). The dataset provided for this task included audio recordings of speakers along with corresponding text and labels. A major challenge faced by most existing ASR-SER systems is that ASR models often misrecognize or skip segments with strong emotional expressions (e.g., speech during crying or anger). At the same time, emotion models also frequently mislabel emotions due to the complex nature of human feelings (for example, crying out of happiness versus crying out of frustration). In this paper, we present our approach to addressing these challenges. We first preprocess the training data and then use it to train a FastConformer-based ASR model. For emotion recognition, we explore multiple models and incorporate additional data to establish a strong rule-based foundation for the emotion task. According to the official evaluation by VLSP in the 2025 ASR-SER challenge, our approach achieved a Word Error Rate (WER) of 12.66% for ASR and an SER accuracy of 80.84%, resulting in an overall score of 85.39 and securing a Top-3 position on the VLSP leaderboard.

Key words: VLSP-2025, Speech-to-text, speech emotion recognition, Automatic Speech Recognition.

1 Introduction

Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER) have seen remarkable progress in recent years, driven by advances in deep learning. Traditional hybrid ASR systems (based on hidden Markov models and deep neural networks) have largely been supplanted by end-to-end approaches that learn to map audio directly to text. Notably, attention-based sequence-to-sequence models were among the first successful end-to-end ASR frameworks. For example, the Listen, Attend and Spell model demonstrated that an attention-enabled encoder-decoder can jointly learn acoustic and language modeling without the need for explicit alignment or pronunciation dictionaries (Chan et al., 2016). This paradigm shift opened the door to training ASR entirely from data, simplifying the pipeline and improving accuracy.

The introduction of the Transformer architecture (Vaswani et al., 2017) further accelerated progress in speech recognition. Transformers, which rely on self-attention in place of recurrence, achieved state-of-the-art results in natural language processing and were soon adopted in ASR, outperforming prior recurrent models. By the late 2010s, Transformer-based ASR had become a de facto standard approach. For instance, a Transformer-based system for Vietnamese ASR (VLSP 2021 Task) achieved a strong syllable error rate of 6.72% on the VLSP evaluation set (Truong, 2022), indicating the effectiveness of self-attention models for speech.

While Transformers capture long-range dependencies well, researchers have sought to improve their ability to model local speech patterns. This led to the development of the Conformer architecture, which combines self-attention with convolutional layers to jointly model global and local features (Gulati et al., 2020). Conformer has delivered superior accuracy on ASR benchmarks, significantly outperforming plain Transformers. In the Vietnamese ASR research community, Conformer-based models also proved highly successful. The winning system of VLSP 2021's ASR challenge leveraged a Conformer with advanced training techniques (gradient mask and pseudo-labeling) and achieved the best performance (Syllable Error Rate of 8.28%) in the competition (Son et al., 2022).

This exemplifies the trend of leveraging state-of-the-art architectures from the global research in local Vietnamese ASR tasks.

Speech Emotion Recognition has likewise benefitted from the general progress in deep learning for speech. SER systems historically relied on handcrafted acoustic features, but modern approaches use learned representations and often borrow architectures from ASR and other speech tasks (Latif et al., 2021). Indeed, representation learning with deep models has significantly improved SER accuracy and robustness. Recent studies show that ASR-oriented models can be adapted to recognize emotion: for example, fine-tuning an ASR model (with appropriate auxiliary information) yielded notable gains in SER performance (Ta et al., 2022). These findings suggest a synergy between ASR and SER, where a common acoustic model can serve both transcription and paralinguistic recognition.

Given these developments, the VLSP 2023 challenge even combined ASR and SER into a single evaluation task under low-resource conditions (VLS, 2023). This joint task setting underlines the need for efficient, high-capacity models that can handle both speech recognition and emotion classification simultaneously. Building on the lessons from VLSP 2021–2023, we select the Fast Conformer architecture as our primary model for the VLSP 2025 ASR & SER challenge. Fast Conformer is an improved Conformer that employs a linearly scalable self-attention mechanism and a novel downsampling schema for efficiency. It is reported to be 2.8× faster than the original Conformer while still achieving state-of-the-art accuracy on ASR benchmarks (Rekesh et al., 2023a). This makes it highly suited for our task: the faster training and inference speed facilitate experimentation under limited data and compute, and the model's strong accuracy provides a solid foundation for both recognition and emotion detection. Moreover, Fast Conformer's ability to handle long speech sequences (through limited-context attention and global token integration) is advantageous for real-world applications where utterances may be lengthy.

In summary, our system adopts Fast Conformer as the backbone for both speech recognition and emotion classification. By leveraging the latest advancements in end-to-end modeling – from attention mechanisms and Transformers through Conformers and their fast variants – we aim to build a unified model that excels in transcribing Vietnamese speech and identifying speaker emotions. In the following sections, we detail our Fast Conformer-based architecture and training strategies for the VLSP 2025 ASR & SER tasks, and evaluate its performance against the challenge criteria.

In this paper, we present our approach to addressing these challenges in the context of the VLSP 2025 Vietnamese ASR–SER challenge. We first perform thorough preprocessing on the training data (e.g., speech normalization and augmentation) and then use it to train a FastConformer-based ASR model, a fast implementation of the Conformer architecture tailored for Vietnamese to better handle the acoustic variability introduced by emotional speech. For the emotion recognition sub-task, we explore multiple model architectures and incorporate additional approved external data, then employ a simple rule-based fusion to establish a strong, robust predictor for the emotion labels. According to the official evaluation on the VLSP 2025 test set, our integrated system achieved a word error rate (WER) of 12.66% for ASR and an SER accuracy of 80.84%, which corresponds to an overall challenge score of 85.39. This performance secured our submission a Top-3 position on the VLSP leaderboard, demonstrating the effectiveness of our combined ASR–SER approach.

2 Relate Work

2.1 SER model

Emotion2Vec is a universal speech emotion representation model pre-trained with self-supervised online distillation on large-scale, open-source, unlabeled emotional speech corpora. Its pre-training jointly optimizes both utterance-level and frame-level objectives, enabling the model to capture not only the global emotional state of an utterance but also the fine-grained temporal dynamics of emotional expression. This dual-level representation is particularly powerful for Speech Emotion Recognition (SER), since holistic utterance embeddings provide stable cues for categorical emotion classification, while frame-level features (at 50 Hz resolution) enhance sensitivity to local prosodic variations such as pitch, intensity, or rhythm shifts that strongly correlate with emotional intensity.

As a result, Emotion2Vec yields highly transferable features across languages, datasets, and tasks; with only a lightweight linear probe, it achieves strong results on benchmarks such as IEMOCAP,

and demonstrates consistent gains across 10 languages (Ma et al., 2024a). In our system, we adopt the *emotion2vec_base* checkpoint as a frozen feature extractor (no fine-tuning). This choice is motivated by its proven robustness in SER, where the availability of richly pre-trained, emotion-sensitive embeddings reduces reliance on large labeled datasets and delivers strong generalization across diverse speech conditions (emotion2vec Team, 2024).

2.2 ASR model

After data preprocessing to ensure balance, diversity, and quality, selecting the right ASR architecture becomes critical for achieving both high accuracy and low-latency inference. The growing demand for deployment on edge devices emphasizes the need for lightweight models that can run efficiently on CPUs and resource-constrained environments. FastConformer was chosen for this task due to its superior inference speed and competitive accuracy, ranking among the top models on the Hugging Face ASR leaderboard. Its design builds upon Conformer, which integrates convolutional and transformer components to capture both local acoustic features and long-range dependencies. FastConformer enhances this architecture with 8× subsampling, refined convolutional kernels, and a combination of local and global attention mechanisms to achieve higher efficiency. As a result, it provides up to 2.7× faster inference with minimal accuracy trade-offs, making it well suited for real-time and large-scale ASR applications.

3 Data preprocessing

To ensure that the selected FastConformer architecture can achieve optimal performance, it is crucial to construct a well-prepared dataset. In this section, we present a comprehensive description of our data pre-processing pipeline and justify the strategies adopted. At the initial stage, we conducted an in-depth analysis of publicly available datasets to determine the typical utterance durations, which informed the time range prioritized during data collection. Furthermore, we examined the distribution of emotional categories and their associated labels to maintain class balance throughout training and to design an effective pre-processing strategy. Additionally, linguistic and prosodic features across different emotional states—such as happiness, sadness, and anger—were analyzed to guide

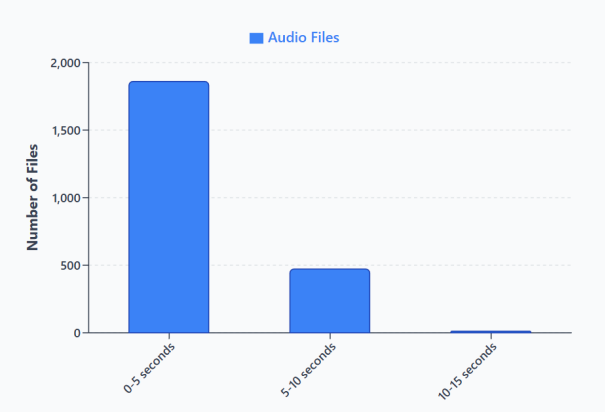

Figure 1: The distribution of audio files and their duration ranges in the public dataset.

the development of a suitable data augmentation approach. This systematic data preparation process constitutes a core contribution of our proposed framework, as it ensures robustness and reliability in subsequent model training.

Based on the analysis of the public test set (Fig. 1), we observed that the majority of speech segments fall within the range of 0-5 seconds (1860 of the 2344 samples, accounting for approximately 79%), while segments between 5-10 seconds represent approximately 20% (472 of the 2344 samples). Consequently, during data processing, we prioritize the selection of 0-5 second samples first and then incorporate 5–10 second samples to maintain an approximate 4:1 ratio, consistent with the distribution in the test set.

To accurately simulate the diversity of data so that it closely resembles the competition dataset, it is essential to first analyze and approximate the temporal distribution of the public data. Starting with the initial set of publicly available audio samples, we performed a filtering step based on utterance duration, categorizing the data into two subsets: short-duration segments (0–5 seconds) and medium-duration segments (5–10 seconds). This division reflects natural speech patterns, as most human utterances typically fall within the 0–5 second range.

Once the filtering process was complete, we aimed to reconstruct the original distribution of the competition data by proportionally mixing the two subsets. Specifically, we adopted a 4:1 ratio, meaning that four samples from the subset containing utterances of five seconds or less were combined with one sample from the subset containing utterances longer than five seconds. This approach

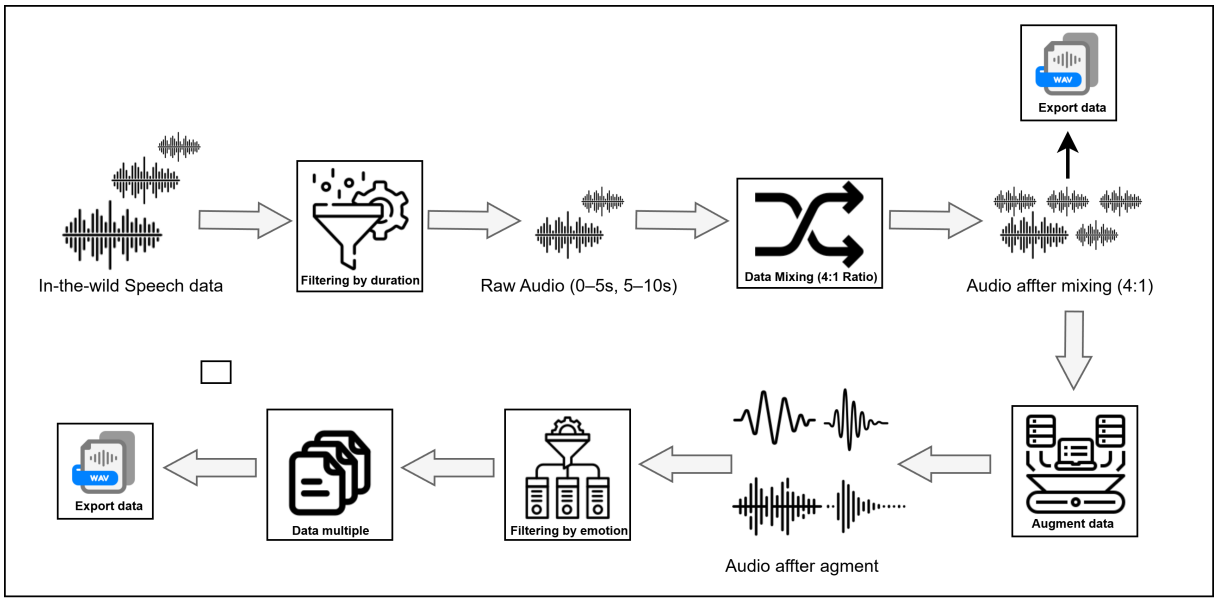

Figure 2: Comprehensive Data Processing Pipeline.

ensured a realistic representation of the temporal structure observed in the target dataset.

However, this fixed-ratio mixing created an imbalance, with a shortage of short utterances (0–5 seconds) and a surplus of longer segments. To address this issue, we applied Voice Activity Detection (VAD) to segment long utterances into shorter ones. These newly segmented utterances were then pre-labeled using Erax on top of Whisper (Nguyen Anh Nguyen, 2025) for training in Phase 1, while approximately 30,000 audio samples were reserved for manual annotation in Phase 2. This strategy increased the number of short samples. The newly segmented data were subsequently reintegrated using the same proportional mixing strategy, ultimately forming the initial version of the raw dataset.

After uniformly mixing the data, we proceeded to enrich the dataset using various techniques informed by human emotional characteristics. Based on research on emotional expression in speech, we observed that when individuals are angry, they tend to speak faster and louder, whereas sadness or crying is often associated with slower and softer speech (Juslin and Laukka, 2003).

Specifically, we applied the following augmentation techniques: speed adjustment to simulate hurried or rushed speech; volume scaling to reflect the intensity of emotions from whispering to shouting; pitch shifting to replicate frequency variations occurring when someone yells or speaks with a trembling voice; and various sound effects, such as echo to convey a sense of emptiness during sadness, or tremolo to emulate a quivering voice associated with anxiety or fear.

In addition to emotion-driven augmentation, we also performed augmentations based on real-world environmental factors. We incorporated diverse background noises, including bustling market sounds with vendors and customers, traffic noise from streets, machinery in factories, as well as natural sounds like wind and rain. These augmentations created a richer dataset closely resembling real recording conditions that the system may encounter during deployment.

Following the data augmentation process, it remained challenging to obtain a sufficient number of emotionally rich samples, as such data are inherently rare in natural speech. The publicly available dataset provided by the organizers also contained only a limited number of instances with high emotional intensity. To address this limitation, we employed additional techniques, including leveraging an emotion detection model combined with manual verification through auditory inspection, to identify segments exhibiting strong emotional expressions. Given that the quantity of these samples was still insufficient, we oversampled them by a factor of three to achieve better class balance in the training set. This approach enables the model to learn more effectively from emotionally expressive samples throughout the training process.

Following comprehensive processing and augmentation (Fig. 2), we obtained two complemen-

tary datasets: a original raw dataset that preserves the inherent distributional characteristics of human speech in real-world environments, and an enhanced dataset enriched through the integration of multiple advanced audio transformation techniques, vividly capturing the diverse emotional spectrum of human expression.

3.1 SER preprocessing

We standardize audio to **16 kHz** mono PCM; apply **VAD/trim silence** to remove long pauses and non-speech; choose feature granularity: **utterance**-level (pooled) vs. **frame**-level **50 Hz** embeddings; enforce **speaker-independent** splits; harmonize labels when mixing corpora; and mitigate class imbalance via **class weights/sampling**. In practice, our pipeline is: *resample* → *VAD/trim* → *loudness norm* → *extract* `frame` *(50 Hz) or pooled* `utterance` *features* → *balance classes*. (Ma et al., 2024a; emotion2vec Team, 2024; Ayadi et al., 2011) As shown in Table 1, the VLSP 2022 and

Dataset	# Utterances
VLSP 2022	19673
VLSP 2023	42040
Total	61713

Table 1: Statistics of the number of utterances in the datasets

VLSP 2023 datasets contain a total of 61,713 utterances after the processing stage.

3.2 ASR preprocessing

ASR (Automatic Speech Recognition) is a specialized task that involves converting speech into text. In this process, the duration of audio segments plays a critical role as it directly influences the model's ability to capture contextual information, maintain data balance, and ensure generalization. Specifically, if the dataset predominantly consists of short utterances, the model tends to become biased toward predicting short sentences and struggles with longer ones during inference. Conversely, if the dataset contains many long utterances, the model learns to preserve context effectively but may produce redundant or incorrect predictions when dealing with shorter sentences.

In the preparation phase for model training, tokenizer processing is a critical component for ensuring robust and efficient model learning. For the Automatic Speech Recognition (ASR) task, we adopted a structured text preprocessing pipeline comprising the following steps: normalizing all text to lowercase, removing punctuation and extraneous special characters, and applying digit normalization prior to tokenizer training. This preprocessing strategy is intended to enhance token consistency and, consequently, improve the overall accuracy of the speech recognition system.

4 Methodology

4.1 Unified Pipeline (Twinkle ASR)

Our proposed system, Twinkle ASR, is designed to simultaneously exploit the strengths of emotion representation learning and efficient automatic speech recognition within a unified framework. Here, unified framework means coupling the two models together into a single framework for inference, where the input is speech and the outputs are both the emotion label and the transcribed text. Subsequently, the results are passed through a post-processing module (rubase) before generating the final output, further enhancing overall accuracy.

4.1.1 ASR-SER Framework

At the upstream stage, we adopt the Emotion2Vec paradigm (Ma et al., 2024a), a universal speech emotion representation model trained with self-supervised online distillation from a teacher–student setup. As illustrated in Figure 3, the teacher network guides the student through both utterance-level and frame-level objectives, ensuring that the extracted features capture both global emotional context and fine-grained temporal variations. We use the emotion2vec_base checkpoint as a frozen feature extractor, producing robust utterance-level embeddings and optional 50 Hz frame-level representations. These representations are then passed to the downstream ASR module, where they provide rich emotional and acoustic cues beneficial for recognition robustness in diverse and emotionally expressive speech.

For the downstream architecture, Twinkle ASR incorporates FastConformer, a recent variant of the Conformer model that offers an excellent trade-off between accuracy and efficiency. While Conformer (Gulati et al., 2020) harmonizes convolutional layers to model local acoustic patterns and Transformer layers to capture long-range dependencies, FastConformer (Rekesh et al., 2023b) introduces several enhancements. These include 8× convolutional subsampling, depthwise-separable convolutions, and a hybrid local–global attention mechanism, which together yield up to 2.7× faster

inference without significant loss in recognition accuracy. Importantly, this design supports scalability to long input sequences (up to several hours), making it well suited for real-world ASR deployments.

The Twinkle ASR pipeline proceeds as follows. Raw input waveforms undergo spectrogram augmentation for robustness against acoustic variability. Next, an 8× convolutional subsampling layer reduces the temporal resolution, feeding into a linear projection layer and dropout regularization to stabilize training. The processed features are then passed through stacked FastConformer blocks, each consisting of feed-forward modules, multi-head self-attention, and convolutional modules with residual connections and normalization. Finally, the output layer maps to the target label set for recognition. To align with competition evaluation protocols, we fine-tune using the original labels of the datasets (Table 1), before mapping them into two high-level categories (neutral and negative) for downstream emotion evaluation.

Overall, As shown in Fig. 3 Our model Twinkle ASR combines the representational power of Emotion2Vec with the speed and scalability of FastConformer, resulting in a system that is accurate, efficient, and emotionally aware. This hybrid design not only achieves strong recognition performance under real-time constraints but also enables robust handling of emotionally rich speech, making it highly suitable for next-generation human–computer interaction systems.

4.1.2 Rule-base postprocessing

The rubase technique we applied is conceptually simple and specifically tailored for the emotion recognition task. By examining the dataset, we observed that in most cases of anger, speakers frequently use offensive or profane words such as "tao", "may", ... and similar expressions. Leveraging this characteristic, we developed an additional rubase module as a post-processing step, designed to detect negative emotions purely based on the presence of such profane words in the transcribed text.

As a result, the emotion label can be reassigned after the ASR + SER pipeline if any offensive words are detected in the recognized text. This lightweight but effective strategy significantly enhanced emotion recognition performance, improving accuracy from approximately 80 to 82 points in our experiments with the private test set.

4.2 Loss function

4.2.1 Loss function for our ASR model

For the ASR component, the loss computation mechanism is designed to optimize training efficiency while maintaining stability across variable-length audio sequences. The primary component is Connectionist Temporal Classification (CTC) Loss (Graves et al., 2006), applied on sequences of BPE (Byte-Pair Encoding) subword units. The decoder is implemented as ConvASRDecoder, which maps the encoder output to the corresponding number of classes:

```
num_classes=vocab+blank
```

prior to CTC loss computation. To ensure stability, the parameter `ctc_reduction` is configured as `mean_volume`, meaning that the loss is averaged over the total number of valid frames in the batch (excluding padding) rather than simply based on batch size.

Additionally, the model supports an optional InterCTC mechanism, which introduces auxiliary CTC losses at one or more intermediate encoder layers to improve convergence and act as a regularization technique. These components are configured via `interctc.loss_weights` and `interctc.apply_at_layers`. The overall loss is computed as:

$$\begin{aligned} \mathcal{L}_{\text{total}} &= \alpha_0 \, \mathcal{L}_{\text{CTC}}^{(\text{final})} + \sum_i \alpha_i \, \mathcal{L}_{\text{CTC}}^{(\text{layer } i)} \\ &\text{with } \alpha_0 = 1 - \sum_i \alpha_i. \end{aligned} \tag{1}$$

4.2.2 Loss function for our EMO model

Our emotion recognition (EMO) module is fine-tuned downstream from the universal speech emotion representation model emotion2vec (Ma et al., 2024b), which is pre-trained via self-supervised online distillation combining both utterance-level and frame-level losses. To further adapt the model to our task, we use a lightweight classifier head on top of frozen emotion2vec features.

We train this downstream classifier using the cross-entropy loss:

$$\mathcal{L}_{\text{EMO}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log \hat{p}_{i,c}, \tag{2}$$

where N is the number of training examples, C is the number of emotion categories, $y_{i,c}$ is the one-

hot label, and $\hat{p}_{i,c}$ is the predicted probability for class c.

In line with the original emotion2vec framework, we freeze the backbone and only update the parameters of the classifier during fine-tuning. This strategy preserves the generalization ability of the pre-trained emotion representations while minimizing the risk of overfitting on limited labeled emotion data. Label smoothing is optionally applied to regularize the model, particularly when class distributions are imbalanced.

5 Experiment

5.1 Training metric

5.1.1 ASR training

The FastConformer-CTC-BPE model was trained from scratch for a total of 300 epochs, with Phase 1 consisting of 200 epochs on the public dataset, followed by Phase 2 with 100 epochs on the carefully annotated dataset. The entire training process took approximately one month on a multi-GPU server equipped with 2 × NVIDIA RTX 4090 GPUs (24 GB each) and 100 GB of RAM, using a batch size of 32 and 16 parallel data-loading workers (`num_workers`).

To mitigate overfitting, a dropout rate of 0.1 was applied to the core network components. The training procedure followed the NeMo configuration for large-scale FastConformer models (120 million parameters), utilizing the CTC loss over BPE subword sequences in combination with a ConvASRDecoder to map encoder outputs into the vocabulary logits space.

Key optimization strategies included mixed-precision training (BF16) to accelerate convergence, reduce memory footprint, and allow stable training with large batch sizes; in case of instability, the model could fall back to FP32. The AdamW optimizer was used with weight decay = 1e-3 and an initial learning rate of 1e-3, coupled with a CosineAnnealing scheduler and 15,000-step warmup, ensuring smooth learning rate adjustments. Regularization and data augmentation strategies incorporated SpecAugment with time masking = 10 and frequency masking = 2, along with feature-wise normalization, enhancing model generalization.

Architecturally, the FastConformer Encoder comprises 18 layers, with `d_model` = 512, 8 attention heads, depthwise convolutional subsampling at an 8× factor with kernel size = 9, balancing performance and accuracy. The optional InterCTC Loss provides auxiliary supervision at intermediate encoder layers, promoting faster convergence and improved alignment. Additionally, gradient clipping, synchronized batch normalization, and checkpointing were applied to maintain stability during large-scale training.

This setup adheres to NVIDIA's recommendations for FastConformer in large ASR systems, effectively balancing accuracy, speed, and scalability, and is particularly suitable for long-sequence speech recognition tasks.

5.1.2 SER training

We employed a pretrained distilled student model derived from the teacher model of emotion2vec. On top of this representation, we trained a linear classification layer with the label sets provided in the two datasets listed in Table 1. During fine-tuning, we preserved the original labels from the papers without applying any mapping. After training, the labels were mapped into two categories, neutral and negative, for the purpose of competition evaluation.

5.2 Result

The FastConformer-CTC-BPE model was trained from scratch for a total of 300 epochs, with Phase 1 consisting of 200 epochs on the public dataset, followed by Phase 2 with 100 epochs on the carefully annotated dataset. The entire training process took approximately one month on a multi-GPU server equipped with 2 × NVIDIA RTX 4090 GPUs (24 GB each) and 100 GB of RAM, using a batch size of 32 and 16 parallel data-loading workers. This rigorous training procedure enabled comprehensive optimization in both performance and generalization. As a result, our final system achieved an overall score of 85.4 across both ASR and SER tasks on the VLSP public dataset, with a word error rate (WER) of 0.074 specifically on the ASR test set. These results highlight the effectiveness of our architectural choices, optimization strategies, and data augmentation techniques in building an advanced Vietnamese speech processing system.

The final evaluation of the competition was conducted on a private test set provided by the organizers, ensuring an objective and rigorous assessment of all submitted systems. Among the generated checkpoints, Checkpoint 299 demonstrated superior performance as a result of extensive fine-tuning and optimization during training.

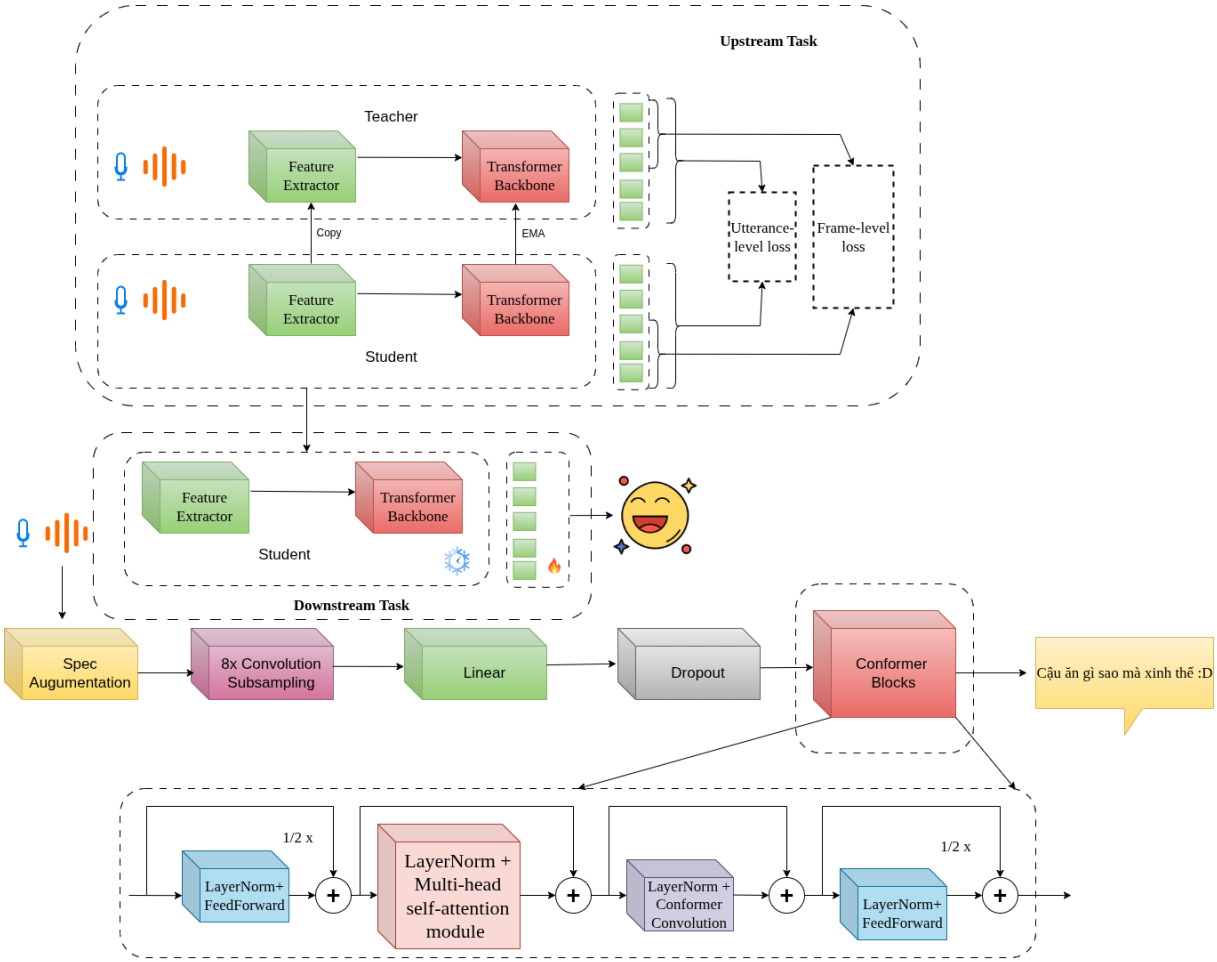

Figure 3: Twinkle ASR.

This model achieved a Word Error Rate (WER) of approximately 0.07 on the internal test set, indicating a substantial reduction in recognition errors compared to earlier iterations. Such a low WER not only confirms the model's high accuracy in Vietnamese speech recognition but also highlights its robustness under diverse acoustic conditions. These findings underscore the effectiveness of our architectural design, training strategies, and data processing pipeline in developing a high-performance ASR system. In the final leaderboard, our team ranked third overall, achieving WER = 12.66%, SER Accuracy = 80.84%, and a Final Score of 85.39, as shown in Table 2.

Rank	Team	WER	SER Acc	Final Score
1	nguyenhythien	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

Table 2: Performance comparison across teams.

6 Conclusion

Acknowledgments

This work is done outside of office hours, and our team was established to bring the best Vietnamese speech processing models.

References

2023. VLSP 2023 Challenge on Automatic Speech Recognition and Speech Emotion Recognition. https://vlsp.org.vn/vlsp2023/eval/asr. Online; accessed Aug. 28, 2025.

Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. 2016. Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964.

emotion2vec Team. 2024. emotion2vec_base — hugging face model card. https://huggingface.co/emotion2vec/emotion2vec_base. Accessed: 2025-08-24.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proc. 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, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. Interspeech*, pages 5036–5040.

Patrik N. Juslin and Petri Laukka. 2003. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129(5):770–814.

Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Junaid Qadir, and Björn Schuller. 2021. Survey of Deep Representation Learning for Speech Emotion Recognition. *TechRxiv Preprint*.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. 2024a. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760, Bangkok, Thailand. Association for Computational Linguistics.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024b. emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation. *Proceedings of ACL 2024 Findings*.

cty bao hiem AAA Nguyen Anh Nguyen, Pham Huynh Nhat. 2025. EraX-WoW-Turbo-V1.1-CT2: Lang nghe de Yeu thuong.

Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. 2023a. Fast Conformer with Linearly Scalable Attention for Efficient Speech Recognition. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Dima Rekesh, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Ankur Kumar, and Boris Ginsburg. 2023b. Fast conformer with linearly scalable attention for efficient speech recognition. https://research.nvidia.com/labs/conv-ai/blogs/2023/2023-06-07-fast-conformer/. Accessed: 2025-08-24.

Dang Dinh Son, Le Dang Linh, Dang Xuan Vuong, Duong Quang Tien, and Ta Bao Thang. 2022. ASR - VLSP 2021: Conformer with Gradient Mask and Stochastic Weight Averaging for Vietnamese Automatic Speech Recognition. *VNU Journal of Science: Computer Science and Communication Engineering*, 38(1):15–21.

Bao Thang Ta, Tung Lam Nguyen, Dinh Son Dang, Nhat Minh Le, and Van Hai Do. 2022. Improving Speech Emotion Recognition via Fine-Tuning ASR with Speaker Information. In *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.

Tien Toan Truong. 2022. ASR - VLSP 2021: An Efficient Transformer-based Approach for Vietnamese ASR Task. *VNU Journal of Science: Computer Science and Communication Engineering*, 38(1).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.