DFAT: Dual-stage Fusion of Acoustic and Text feature for Speech Emotion Recognition

Nhi Nguyen Yen Truong[1], Sang Le Quang[1], Huy Tran Quang[1], Tri Pham Xuan[1], Duong Tran Ham[1], Binh Tran Le Hai[1], Tin Huynh[1], Kiem Hoang [1]

[1]The Saigon International University
{truongnguyenyennhi, lequangsang, tranquanghuyk15, phamxuantri, tranhamduong, tranlehaibinhk12, huynhngoctin, hoangkiem}@siu.edu.vn

Abstract

Speech Emotion Recognition (SER) is an important task in affective computing and human-computer interaction, with applications in virtual assistants, customer service, education, and healthcare. Most existing approaches use early or late fusion, but they are complex and require large labeled datasets, which limits practical use, especially for low-resource languages like Vietnamese. We propose a hybrid fusion pipeline that concatenates acoustic features with ASR-based text features and processes them using Logistic Regression, Random Forest, and XGBoost with ensemble weighting. On the VLSP 2025 private test set, hybrid fusion achieves 0.8438 WA, outperforming early fusion (0.8131 WA), late fusion (0.8140 WA), and both acoustic-only (0.7458 WA) and text-only (0.7463 WA) approaches. This demonstrates that hybrid fusion is the most effective method for SER in Vietnamese.

1 Introduction

Emotion in communication is the process through which humans express, convey, and perceive affective states via both linguistic and non-linguistic channels. It goes beyond the semantic meaning of speech and encompasses vocal attributes such as prosody, pitch, rhythm, and intensity (Scherer, 2003). This diversity enables humans to easily recognize emotions but poses significant challenges for machines, thereby driving the need to develop automatic systems capable of emotion recognition.

In this context, SER has emerged as a central task in affective computing and human-computer interaction. SER refers to the process of analyzing speech signals to infer the emotional state of the speaker. This capability unlocks a wide range of applications, from customer experience analysis to online education support.

However, emotion recognition based only on acoustics is not sufficient, since emotions are also expressed through linguistic content. Acoustics reflect paralinguistic cues, while text conveys explicit meaning; the two channels complement each other in emotion recognition. Therefore, the key challenge is how to effectively combine both sources of information, leveraging their strengths and addressing their weaknesses through fusion methods, in order to build a lightweight pipeline.

Fusion refers to integrating multiple information sources to predict an output variable. There are three main approaches: early fusion (combining features right after extraction, capturing inter-channel correlations but requiring parallel data), late fusion (combining predictions from each channel via averaging, voting, or weighting-flexible with missing data but ignoring low-level interactions), and hybrid fusion aka dual-stage fusion (leveraging both approaches for better performance) (Baltrušaitis et al., 2018).

Recently, in the line of fusion-based approaches for SER, researchers have mainly explored early fusion (Thi et al., 2025) or late fusion (Gómez-Sirvent et al., 2025) strategies, and in some cases, hybrid fusion approaches have incorporated mechanisms such as attention or ensemble learning (Resende Faria et al., 2024; He et al., 2024) to improve accuracy. However, these hybrid methods often depend on complex architectures, which reduce their practicality, and suggest a need for lightweight fusion models that can still achieve good performance (Chowdhury et al., 2025). In addition, their reliance on large amounts of labeled data makes them less suitable for low-resource languages such as Vietnamese (Anh et al., 2024).

In this study, we address the challenge of SER in low-resource languages such as Vietnamese, where large-scale labeled datasets and complex deep architectures are impractical. Prior work has shown that lightweight classifiers combined through ensemble learning can deliver accuracy and robustness comparable to deep models when applied to

multimodal features (Sahu et al., 2019; Guo et al.). Motivated by the limitations of traditional fusion methods, we design a lightweight hybrid fusion pipeline that integrates acoustic features and textual features extracted from Automatic Speech Recognition (ASR) outputs into a unified representation. This joint representation is processed by multiple classifiers to exploit their diverse characteristics. Their outputs are combined through a weighted ensemble mechanism, with the weights optimized using Optuna, yielding final predictions that are efficient, easy to deploy, and reliably accurate.

On the VLSP 2025 private test set, our hybrid fusion approach achieves 0.8438 WA, outperforming early fusion (0.8131 WA) and late fusion (0.8140 WA). It also surpasses the acoustic-only (0.7458 WA) and text-only (0.7463 WA) methods. This shows that hybrid fusion not only improves upon unimodal baselines but also works better than traditional fusion techniques, making it an effective solution for SER in Vietnamese.

The main contributions of this work are as follows:

1. We propose DFAT, a lightweight dual-stage fusion pipeline that unifies acoustic and textual features (early fusion) and combines classifiers through algorithmically optimized ensemble weighting (late fusion), addressing data scarcity and complexity in low-resource SER.

2. This practice includes extensive experiments comparing our method with early fusion, late fusion, text-only, and acoustic-only baselines, showing that the proposed approach consistently outperforms these alternatives.

3. Presented a comprehensive technical report offering the evaluation of Vietnamese multimodal SER on the VLSP 2025 private test set.

Our implementation is publicly available.[1]

2 Related Work

In SER, three main feature fusion strategies are commonly employed: early fusion, late fusion, and hybrid fusion. Early fusion enables the model to capture low-level interactions between prosody and semantics but requires highly synchronized data, suffers from *high dimensionality*, and remains sensitive to noise propagation (Zadeh et al., 2020). Late fusion *works better when one modality is unavailable*; however, it fails to exploit the complementary information between transcript and speech, making performance heavily dependent on the quality of each sub-module (Mai et al., 2024). To address these limitations, hybrid fusion has been proposed and shown to improve accuracy (Mai et al., 2024).

Within hybrid fusion, the text branch often relies on large pre-trained large language models such as BERT, RoBERTa, or GPT-2 (Chen et al., 2021). While effective, these models require extensive labeled data and computational resources, which poses challenges for low-resource languages like Vietnamese. On the speech branch, self-supervised learning backbones such as Wav2Vec 2.0 and HuBERT have become standard since 2020. Prior studies demonstrated that Wav2Vec 2.0 embeddings outperform handcrafted features (Pepino et al., 2021), HuBERT achieves up to 79.6% WA on IEMOCAP (Wang et al., 2021), and two-stage fine-tuning improves the emotional expressiveness of Wav2Vec 2.0 embeddings (Gao et al., 2023). More recently, Yu et al. (Yu et al., 2024) provided a comprehensive benchmark of Wav2Vec 2.0 across SER, SLU, and speaker verification. Nevertheless, these backbones are mainly trained on English data and are not optimized for end-to-end pipelines in low-resource languages.

3 Method

3.1 Overview of the Pipeline

The system takes raw audio as input and extracts two feature streams: *speech emotion features* from the Speech Emotion Feature Extraction (SEFE) block, and *text emotion features* derived via Automatic Speech Recognition and encoded with the Text Emotion Feature Extraction (TEFE) block. These features are concatenated and fed into three classifiers: XGBoost, Logistic Regression, and Random Forest. XGBoost is fine-tuned with Optuna, while the others serve as baselines. Their probability outputs are then combined through an Optuna-tuned ensemble to produce the final emotion prediction (Figure 1).

3.2 ASR

Among the many ASR models available for Vietnamese, we benchmarked three representative

[1] https://github.com/nhitny/DFAT

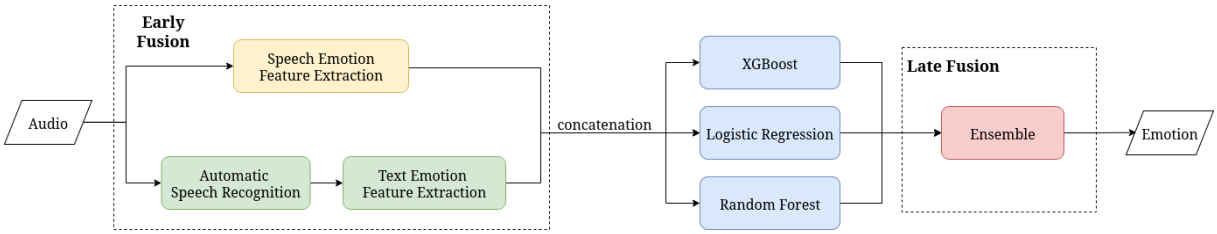

Figure 1: Overview of the hybrid fusion pipeline. Audio is processed through both speech-based and text-based feature extraction streams, then merged and classified using XGBoost, Logistic Regression, and Random Forest. Their outputs fed into a weighted ensemble to produce the final emotion label.

backbones, following the design principles discussed in (Radford et al., 2023), where models are selected based on their accessibility, representativeness, and feasibility for integration into downstream pipelines:

- *wav2vec2-base-vietnamese 250h*[2]: a language-specific model trained on 250 hours of Vietnamese speech, but it often produces spelling errors and, lacking a decoder, generates less coherent transcripts.
- *wav2vec2-xls-r-300m*[3]: a large multilingual model supporting 436 languages (including Vietnamese), with good cross-lingual generalization but not fully optimized for Vietnamese.
- *whisper-small*[4]: an encoder-decoder model trained on 680k hours of multilingual data, more robust and easier to fine-tune.

Among the whisper variants officially released (tiny, base, small), we selected whisper-small because it provides the best trade-off between accuracy and computational efficiency. As reported in the original whisper paper (Radford et al., 2023), the tiny and base variants are considerably faster but yield substantially higher WER, especially for non-English languages such as Vietnamese. In contrast, whisper-small achieves much lower error rates while remaining computationally feasible, which aligns with the competition's resource constraints. The model was fine-tuned on two large-scale Vietnamese datasets: *28k Vietnamese Voice Augmented of VinBigData* (28k-vn) and *PhoAudioBook*, using early stopping based on Word Error Rate (WER) and saving the best checkpoints. The resulting transcripts were then fed into the TEFE to obtain textual emotion representations, which were subsequently fused with acoustic features in the hybrid fusion pipeline.

[2]https://huggingface.co/nguyenvulebinh/wav2vec2-base-vietnamese-250h

[3]https://huggingface.co/facebook/wav2vec2-xls-r-300m

[4]https://huggingface.co/openai/whisper-small

3.3 TEFE

The TEFE block takes ASR-generated transcripts and encodes them into a fixed-length feature that captures linguistic emotional cues. We experimented with four architectures-LSTM, BiLSTM, CNN, and BiLSTM_CNN-chosen for their complementary strengths in modeling sequential dependencies and local patterns. All models were trained on the ViSEC dataset, and the one achieving the highest accuracy was selected as the TEFE backbone. The resulting representation, denoted as $\mathbf{f}_{\text{TEFE}} \in \mathbb{R}^{1024}$, is passed into the classifiers and the fusion module.

3.4 SEFE

The SEFE block takes raw audio and outputs a 1024-dimensional feature that captures emotion-related cues for classification and fusion. Under VLSP 2025 constraints, we evaluated two pre-trained encoders - WavLM and Emotion2Vec - and selected the one with higher accuracy on the public test set. The pipeline includes resampling to 16kHz, amplitude normalization - where the waveform is scaled by dividing all samples by the maximum absolute amplitude (peak normalization), and trimming/padding before the backbone generates the feature vector. The resulting representation, $\mathbf{f}_{\text{SEFE}} \in \mathbb{R}^{1024}$, is passed to the classifiers and fusion module.

3.5 Fusion

3.5.1 Overview

The fusion stage concatenates acoustic and textual features from the SEFE and TEFE blocks into a

single feature vector, which is then used for classification. The purpose is to capture complementary information from both modalities. To make the prediction more reliable, we use an ensemble of three classifiers - Logistic Regression, Random Forest, and XGBoost. Each model produces a probability score, and we combine these scores with learned weights so that the final decision benefits from the strengths of all three models. The input is the combined SEFE-TEFE representation, and the output is the predicted emotion label.

3.5.2 Classifiers

We picked three classifiers with different characteristics:

- **Logistic Regression (LR):** a linear baseline that provides an interpretable decision boundary. It is efficient and suitable when data is approximately linearly separable (Hastie et al., 2009).
- **Random Forest (RF):** an ensemble of decision trees that models non-linear relationships and is robust to noise. By averaging predictions from multiple trees, RF reduces variance and mitigates overfitting (Breiman, 2001).
- **XGBoost (XGB):** a boosting-based method that sequentially improves weak learners to form a strong predictor. Hyperparameters such as learning rate, maximum depth, and regularization terms were optimized with stratified cross-validation, ensuring strong performance on complex decision boundaries (Freund and Schapire, 1997).

These models were selected because they represent diverse inductive biases-linear (LR), bagging (RF), and boosting (XGB)-which makes their combination more robust (Sagi and Rokach, 2018).

3.5.3 Ensemble Fusion

Instead of relying on a single classifier, we aggregate their probability outputs using a weighted average:

$$P_{\text{final}} = w_1 P_{\text{XGB}} + w_2 P_{\text{RF}} + w_3 P_{\text{LR}}, \quad (1)$$

subject to the constraint:

$$w_1 + w_2 + w_3 = 1. \quad (2)$$

The ensemble weights (w1, w2, w3) were selected based on validation results. This allows the ensemble to balance the strengths of each model-LR's simplicity, RF's stability, and XGB's predictive power while avoiding reliance on a single classifier.

4 Experiment

4.1 Dataset

We employ three publicly available Vietnamese datasets for training and evaluation.

(i) *ViSEC* (Nguyen et al., 2022) is a text-based emotion corpus, which we use for the textual branch of SER. It contains 5,280 utterances labeled with four emotion categories: *neutral*, *angry*, *happy*, and *sad*. To align with the binary classification setup of the competition, we map *neutral* and *happy* into a single neutral class, while *angry* and *sad* are merged into the negative class. After this mapping, the neutral class accounts for 51.7% of the data and the negative class for 48.3%, resulting in a fairly balanced dataset that simplifies the emotion space while preserving key distinctions relevant to sentiment polarity.

(ii) *28k-vn* (Vingroup Big Data Institute, 2020) is a speech corpus containing approximately 28,000 augmented utterances, used for ASR fine-tuning.

(iii) *PhoAudiobook* (Thi Vu and Nguyen, 2025) is a large-scale Vietnamese audiobook dataset, which provides long-form speech data to further improve ASR performance.

In the context of the competition, where the use of training data is restricted to publicly available Vietnamese datasets, we carefully selected these three corpora. They are diverse and representative of different aspects of the Vietnamese language: *ViSEC* provides rich textual emotional expressions, *28k-vn* offers a wide range of augmented speech utterances for robust ASR training, and *PhoAudiobook* contributes large-scale, long-form speech data that captures natural prosody and speaking styles. Together, they ensure both diversity and linguistic characteristics specific to Vietnamese, which are crucial for building effective SER and ASR systems.

The statistics of these datasets are summarized in Table 1.

[5] https://huggingface.co/datasets/hustep-lab/ViSEC

[6] https://huggingface.co/datasets/natmin322/28k_vietnamese_voice_augmented_of_VinBigData

[7] https://huggingface.co/datasets/thivux/phoaudiobook

Table 1: Summary of Vietnamese datasets used in our experiments

Dataset	Task	Size	Unit
ViSEC[5]	SER	5,280	utterances
28k-vn[6]	ASR	28,000	utterances
PhoAudiobook[7]	ASR	1,494	hours

4.2 Evaluation Metrics

4.2.1 ASR Metrics

To measure the accuracy of ASR models, we adopted the Word Error Rate (WER), which is the most widely used evaluation metric in the ASR literature.

4.2.2 SEFE, TEFE and Fusion Metrics

We evaluate model performance using three standard metrics widely adopted in SER research: Weighted Accuracy (WA), Unweighted Accuracy (UA), and Weighted F1-score (WF1). These metrics capture both overall correctness and class-wise balance, which are crucial for datasets with imbalanced emotion distributions.

WA measures the overall proportion of correctly predicted samples:

$$\text{WA} = \frac{\sum_{i=1}^{N} \mathbf{1}(y_i = \hat{y}_i)}{N} \times 100, \tag{3}$$

where N is the total number of samples, y_i is the ground-truth label, and $\hat{y}_i$ is the predicted label.

UA computes the average of recall scores across all K emotion classes, treating each class equally, regardless of frequency:

$$\text{UA} = \frac{1}{K} \sum_{k=1}^{K} \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k} \times 100, \tag{4}$$

where TP_k and FN_k denote true positives and false negatives for class k.

WF1 evaluates the harmonic mean of precision and recall across all classes, weighted by the class support w_k:

$$\text{WF1} = \frac{\sum_{k=1}^{K} w_k \cdot \text{F1}_k}{\sum_{k=1}^{K} w_k} \times 100, \tag{5}$$

where F1_k is the per-class F1-score:

$$\text{F1}_k = \frac{2 \cdot \text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}. \tag{6}$$

Higher values of WA, UA, and WF1 indicate better classification performance.

4.3 Experimental Setup

4.3.1 Experiment Setup ASR

We conduct ASR experiments for Vietnamese using three models: wav2vec2-base-vietnamese-250h, wav2vec2-xls-r-300m, and whisper-small. The two wav2vec2 models use an encoder architecture with a CTC head, while whisper applies a Transformer encoder-decoder architecture.

We evaluated models on the public and private test sets of VLSP 2025 to compare their initial performance. Subsequently, we fine-tuned whisper-small on two large-scale Vietnamese datasets: PhoAudiobook (1,494 hours of audiobook data) and 28k-vn (28,000 augmented sentences). Fine-tuning was performed using early stopping based on WER, and the best checkpoint was saved for evaluation.

The audio in all datasets is normalized to 16 kHz. The text is normalized (correcting abbreviations, spelling, special characters) before being tokenized using WhisperProcessor. The training process uses AdamW with a learning rate ranging from 1×10^{-5} to 3×10^{-5}, a warmup of 500 steps, a maximum of 20 epochs, an effective batch size of 16-20 (through gradient accumulation), and training with fp16 mixed precision.

Evaluation: The models are compared using WER on both the public and private test sets.

4.3.2 Experiment Setup TEFE

The goal of this experiment is to compare different text embedding feature extraction (TEFE) models for emotion recognition from text. We evaluate CNN, LSTM, BiLSTM, and BiLSTM_CNN architectures to determine which model best captures textual emotional features.

We use the ViSEC dataset, which contains Vietnamese text annotated with emotion labels. The preprocessing includes:

- Tokenizing text into word sequences.
- Mapping tokens to a frozen 1024-dimensional embedding space, where the 1024 dimension was chosen to align with SEFE outputs (also 1024-d) to enable direct fusion.
- Padding or truncating sequences to a fixed maximum length of 200 tokens.
- Handling out-of-vocabulary (OOV) tokens by mapping them to a special <UNK> symbol.

- Text normalization by converting all characters to lowercase while preserving Vietnamese diacritics to maintain semantic distinctions.

We evaluate four different TEFE architectures:

- **CNN:** A frozen 1024-d *Embedding*, four parallel *Conv1D* branches with 256 filters each, followed by *BatchNorm*, *ReLU*, and *GlobalMaxPooling1D*. Outputs are concatenated into a 1024-d vector, followed by *Dropout* and a final *Dense(1)*. Total: 8.56M parameters (4.72M trainable, 3.83M non-trainable).

- **LSTM:** A frozen 1024-d *Embedding*, a *LSTM* layer with 1024 units, followed by *Dropout* and a final *Dense(1)*. Total: 12.23M parameters (8.39M trainable, 3.83M non-trainable).

- **BiLSTM:** A frozen 1024-d *Embedding*, a *BiLSTM* layer with 512 units per direction (1024 total), followed by *Dropout* and a final *Dense(1)*. Total: 10.13M parameters (6.30M trainable, 3.83M non-trainable).

- **BiLSTM_CNN:** A frozen 1024-d *Embedding*, a *BiLSTM* layer with 512 units per direction, followed by four parallel *Conv1D* branches (256 filters each, with different kernel sizes). Each branch applies *BatchNorm*, *ReLU*, and *GlobalMaxPooling1D*; outputs are concatenated into a 1024-d vector, followed by *Dropout* and a final *Dense(1)*. Total: 8.82M parameters (4.99M trainable, 3.83M non-trainable).

In all models, the textual feature is represented as $\mathbf{f}_{\text{TEFE}} \in \mathbb{R}^{1024}$.

For each input sentence:

- The sentence is tokenized and mapped into a frozen 1024-d embedding.

- A TEFE model (CNN, LSTM, BiLSTM, or BiLSTM_CNN) processes the embedding to extract textual emotion features.

- A final *Dense(1)* layer outputs the predicted emotion.

All models are trained for 30 epochs on the ViSEC dataset with early stopping to select the best checkpoint. The models are compared using Weighted Accuracy (WA), Unweighted Accuracy (UA), and Weighted F1 (WF1) scores.

We expect to determine which architecture (CNN, LSTM, BiLSTM, or BiLSTM_CNN) best captures emotional features from text. By evaluating these models directly, we will identify the most effective TEFE design for Vietnamese text-based emotion recognition.

This setup allows us to determine which TEFE architecture best captures textual emotional features in Vietnamese, while ensuring alignment with SEFE for seamless multimodal fusion.

4.3.3 Experiment Setup SEFE

The goal of this experiment is to compare the performance of two pretrained acoustic models, WavLM and Emotion2Vec, when used as end-to-end classifiers for speech emotion recognition. Both models are pretrained specifically for speech emotion tasks and are capable of mapping an input waveform directly to an emotion label without requiring any additional downstream classifiers.

We use the VLSP 2025 public test set, which contains Vietnamese emotional speech. The audio files undergo the following preprocessing steps:

- Resampling to 16 kHz to match the model input requirements.

- Converting the audio to mono and applying peak amplitude normalization.

- Padding or trimming each audio clip to a fixed length of 5 seconds.

We evaluate two pretrained models in the end-to-end setting:

- **WavLM (Wav2Vec2-based):** A model fine-tuned for speech emotion recognition that predicts emotion labels directly from raw speech.

- **Emotion2Vec:** A model trained for speech emotion recognition that outputs emotion predictions directly from input waveforms.

For each audio file, the following steps are performed:

- Load the waveform using the librosa library.

- Pass the waveform through the chosen encoder (WavLM or Emotion2Vec).

- Obtain a probability distribution over emotion classes and select the label with maximum probability.

The two models are compared using Weighted Accuracy (WA), Unweighted Accuracy (UA), and Weighted F1 (WF1) scores.

This experiment provides a benchmark of end-to-end acoustic classifiers and serves as a reference point for the subsequent fusion experiments, where acoustic features are combined with textual features for improved performance.

4.3.4 Experiment Setup Fusion

The goal of this experiment is to evaluate the effectiveness of hybrid fusion of acoustic (SEFE) and textual (TEFE) features. After feature extraction, we compute frame-level statistics (mean, variance, maximum, minimum) for each modality and concatenate them into a joint feature vector.

Three classifiers are trained on these fused representations: Logistic Regression (with balanced class weights and maximum 500 iterations), Random Forest (300 trees, maximum depth = 8), and XGBoost. (3) XGBoost hyperparameters are tuned with Optuna over the following search space: $n\_estimators \in [200, 500]$, $max\_depth \in [3, 8]$, $learning\_rate \in [0.01, 0.2]$, $subsample \in [0.6, 1.0]$, $colsample\_bytree \in [0.6, 1.0]$, $min\_child\_weight \in [1, 10]$, and $\gamma \in [0, 2.0]$.

To combine classifiers, we adopt an ensemble fusion strategy where the final probability is a weighted sum of the three outputs. Ensemble weights (w_1, w_2, w_3) are optimized using Optuna with stratified cross-validation. Unlike uniform averaging, this approach allows the ensemble to exploit the complementary strengths of Logistic Regression (linear), Random Forest (bagging), and XGBoost (boosting).

Fusion optimization. The search space for the ensemble weights was defined as $w_1 \in [0, 1]$, $w_2 \in [0, 1 - w_1]$, and $w_3 = 1 - w_1 - w_2$. We conducted 50 Optuna trials with 2-fold stratified cross-validation and a fixed random seed (42) to ensure.

5 Result and discussion

5.1 Result

We evaluated three Vietnamese ASR backbones: wav2vec2-base-vi-250h, XLS-R-300M, and whisper-small. As shown in Table 3, whisper-small achieved the best performance after fine-tuning and was selected as the final ASR backbone. For acoustic features (SEFE), Emotion2Vec outperformed WavLM, while for textual features (TEFE), LSTM gave the best results among four candidates. Finally, fusion experiments with Logistic Regression, Random Forest, and XGBoost showed that the Optuna-tuned ensemble consistently outperformed unimodal baselines (Table 2).

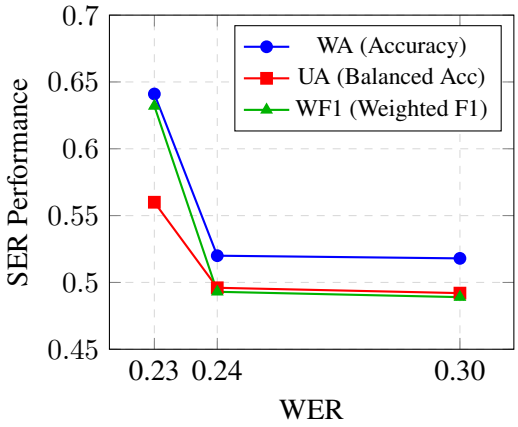

Figure 2: Relationship between ASR quality (WER) and SER performance (WA/UA/WF1). As WER decreases, all SER metrics improve, confirming the coupling between ASR and SER.

After the VLSP 2025 competition results were announced, we reproduced our experiments following the described pipeline and achieved 2nd place with SER Accuracy = 0.8438 (equivalent to Weighted Accuracy – WA), as shown in (Table 4).

5.2 Discussion

The experimental results highlight that while whisper-small is the most reliable ASR backbone and Emotion2Vec and LSTM provide the strongest unimodal features for speech and text, the largest gains come from the fusion stage. By combining SEFE and TEFE features, and optimizing classifier weights with Optuna, the ensemble consistently surpassed all experiment unimodal baselines. This confirms that integrating acoustic and textual cues through fusion is essential for robust emotion recognition.

The best weights were found to be:

$$w_1 = 0.562 \quad w_2 = 0.160 \quad w_3 = 0.279$$

where w_1, w_2, and w_3 correspond to XGBoost, Random Forest, and Logistic Regression, respectively. These results indicate that XGBoost contributed most strongly to the final decision, while the other classifiers provided complementary improvements.

Our model performed well in SER thanks to the

Table 2: Performance on Public and Private test sets. WA = overall accuracy; UA = balanced accuracy; WF1 = weighted F1

Model	Public Test			Private Test		
	WA	**UA**	**WF1**	**WA**	**UA**	**WF1**
Audio encoders (SEFE)						
WavLM	0.5627	0.5627	0.4053	0.5709	0.5709	0.4149
Emotion2Vec	**0.7458**	**0.7458**	**0.7354**	**0.7376**	**0.7376**	**0.7253**
Text encoders (TEFE)						
LSTM	**0.7542**	**0.7419**	**0.7510**	**0.7462**	**0.7315**	**0.7426**
BiLSTM	0.7495	0.7374	0.7464	0.7300	0.7145	0.7259
CNN	0.7261	0.7060	0.7172	0.7105	0.6861	0.6997
BiLSTM_CNN	0.7359	0.7157	0.7271	0.7215	0.6981	0.7118
Early Fusion						
LSTM + Emotion2Vec + XGB	**0.8106**	**0.8106**	**0.7616**	**0.8251**	**0.8251**	**0.7878**
LSTM + Emotion2Vec + RF	0.7569	0.7569	0.6256	0.7850	0.7850	0.6870
LSTM + Emotion2Vec + LR	0.4606	0.4606	0.6141	0.4654	0.4654	0.6206
Late Fusion						
LR + XGB + RF	**0.7900**	**0.7900**	**0.7930**	**0.8140**	**0.8040**	**0.8130**
Hybrid Fusion						
LSTM + Emotion2Vec + LR + XGB + RF	**0.8439**	**0.8529**	**0.8412**	**0.8438**	**0.8438**	**0.8436**

Table 3: WER results of four ASR models on public and private test sets

Model	Public Test	Private Test
wav2vec2-base-vietnamese-250h	0.26	0.26
wav2vec2-xls-r-300m	0.68	0.67
whisper-small	0.60	0.56
whisper-small (ours)	**0.22**	**0.23**

Table 4: VLSP 2025 SER Results

Rank	SER Accuracy
1	0.8579
2	**0.8438**
3	0.8221
4	0.8084
5	0.7950
6	0.7913
7	0.6650

hybrid fusion method, which combines both acoustic and textual features for emotion recognition. As shown in Figure 2, when the ASR WER goes down, all SER scores (WA, UA, WF1) go up. This shows that ASR quality has a clear effect on how well the emotion recognition system works. In other words, changes in WER can directly affect SER results.

6 Conclusion and Future Work

This paper introduced a hybrid fusion pipeline for Vietnamese Speech Emotion Recognition, leveraging acoustic features from Emotion2Vec and textual features from LSTM applied to whisper-small ASR transcripts. The two modalities were combined in a fusion stage where Logistic Regression, Random Forest, and XGBoost were trained, and their outputs aggregated via an Optuna-tuned ensemble. Experiments on the VLSP 2025 benchmark demonstrated that this approach consistently outperforms unimodal systems, highlighting the effectiveness of modality integration and ensemble fusion.

Future work will focus on two directions. First, to further improve accuracy, we plan to explore larger-scale pretraining and advanced fusion architectures such as Bayesian neural networks and test-time augmentation for better uncertainty handling. Second, to enable deployment in real-world applications with limited resources, we will investigate model compression techniques-including knowledge distillation, pruning, and quantization-to reduce computational and memory overhead while maintaining performance. These directions will help extend the proposed pipeline toward both higher accuracy and broader usability in affective computing scenarios.

References

ND Quang Anh, Manh-Hung Ha, Quynh Chi Nguyen, Thu Hien Nguyen Thi, Quan Vu, DX Minh-Duc, Duc-Chinh Nguyen, and Thai Kim Dinh. 2024. Vnemos: Vietnamese speech emotion inference us-

ing deep neural networks. In *2024 9th International Conference on Integrated Circuits, Design, and Verification (ICDV)*, pages 97–101. IEEE.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Xuankai Chen, Wei-Ning Hsu, and James Glass. 2021. Exploring pre-trained language models for speech emotion recognition. In *ICASSP*.

Jaher Hassan Chowdhury, Sheela Ramanna, and Ketan Kotecha. 2025. Speech emotion recognition with light weight deep neural ensemble model using hand crafted features. *Scientific Reports*, 15(1):11824.

Yoav Freund and Robert E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Yujie Gao, Ming Li, and Wei Xu. 2023. Improving speech emotion recognition with two-stage fine-tuning of wav2vec 2.0. *Computer Speech & Language*.

José L Gómez-Sirvent, Francisco López de la Rosa, Daniel Sánchez-Reolid, Roberto Sánchez-Reolid, and Antonio Fernández-Caballero. 2025. Small language models for speech emotion recognition in text and audio modalities. *Applied Sciences*, 15(14):7730.

Meng Guo, Xinyu Li, Hong Xu, and Tianyu Wu. An ensemble learning approach for speech emotion recognition. *PLOS ONE*, 17(4):e02.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2 edition. Springer, New York.

Jiajun He, Xiaohan Shi, Xingfeng Li, and Tomoki Toda. 2024. Mf-aed-aec: Speech emotion recognition by leveraging multimodal fusion, asr error detection, and asr error correction. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11066–11070. IEEE.

Thanh Mai, Bao Nguyen, and Tuan Le. 2024. A survey on speech emotion recognition: Datasets, methods, and trends. *IEEE Transactions on Affective Computing*.

Van-Huy Nguyen, Minh-Trung Le, and 1 others. 2022. Visec: A vietnamese text-based emotion corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA). Vietnamese text-based emotion corpus for SER.

Laura Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. In *Interspeech*.

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.

Diego Resende Faria, Abraham Itzhak Weinberg, and Pedro Paulo Ayrosa. 2024. Multimodal affective communication analysis: Fusing speech emotion and text sentiment using machine learning. *Applied Sciences*, 14(15):6631.

Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249.

Samiksha Sahu, Ruchika Gupta, Kalika Bali, and Monojit Choudhury. 2019. Multi-modal learning for speech emotion recognition: A comparative analysis. In *Proceedings of the 19th International Conference on Multimodal Interaction*, pages 222–226. ACM.

Klaus R Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1-2):227–256.

Ngoc Tram Huynh Thi, Minh Dzuy Pham, Son Thanh Le, Duc Dat Pham, Kha-Tu Huynh, Nguyen Tan Viet Tuyen, and Tan Duy Le. 2025. Vietnamese emotion recognition from voice and text: A confidence-based approach. In *2025 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6. IEEE.

Xuan Thi Vu and et al. Nguyen. 2025. Phoaudiobook: A large-scale vietnamese audiobook dataset. https://huggingface.co/datasets/thivux/phoaudiobook. 1,494 hours of Vietnamese audiobook speech for ASR training and evaluation.

Vingroup Big Data Institute. 2020. 28k vietnamese voice augmented dataset (vigbigdata). https://huggingface.co/datasets/natmin322/28k_vietnamese_voice_augmented_of_VigBigData. Speech corpus containing ∼28k augmented utterances for ASR.

Sanyuan Wang, Zhuo Chen, Yu Wu, and 1 others. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. In *NeurIPS*.

Chen Yu, Hao Zhang, and Xiaodong Li. 2024. A comprehensive benchmark of wav2vec 2.0 across speech emotion recognition, spoken language understanding, and speaker verification. In *ICASSP*.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2020. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of ACL*.