# Speech Emotion Recognition via Parallel Dual-Branch Fusion Model

Zhongliang Wei[1]*, Chang Ge[2], Lijun Zhu[3], Jinmin Ye[4]
The First Affiliated Hospital, Anhui University of Science and Technology, Huainan, China[1]
School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, China[1,2,3,4]

***Abstract*—Speech Emotion Recognition (SER) has become a pivotal topic within affective computing and human–computer interaction, where the core challenge lies in jointly capturing both the time–frequency structure and the semantic context of speech. To overcome the shortcomings of current approaches—including single-view feature representation, the lack of emotional discriminability in self-supervised models, and suboptimal complementarity among fusion strategies—this study proposes a parallel dual-branch fusion architecture for SER. The framework consists of a wav2vec 2.0 branch and a CNN–Transformer spectrogram branch, which respectively extract contextual semantic representations from raw waveforms and explicit time–frequency features from spectrograms. A logistic regression fusion mechanism is further introduced at the decision level to achieve adaptive weighting in the probability space, thereby fully leveraging the complementary strengths of the two feature types. Experiments carried out on the RAVDESS audio subset showed that the proposed model surpassed several mainstream baselines (e.g., CNN-n-GRU and RELUEM), achieving 92.7% accuracy and 92.2% Macro-F1, with an average improvement of about 3.2 percentage points. The layer unfreezing studies confirmed the effectiveness of partial fine-tuning for transferring pretrained features, while the comparative experiments on fusion strategies validated the superiority of probability-space fusion in both performance and stability. Overall, the proposed framework achieves simultaneous gains in accuracy and robustness through feature complementarity, branch decoupling, and lightweight fusion. Future work will explore cross-lingual generalization, multimodal extensions, lightweight deployment, and dynamic emotion modeling, contributing to more efficient affective computing and intelligent interaction systems.**

***Keywords*—RAVDESS; Speech Emotion Recognition; spectrogram modeling; probability-space fusion; wav2vec 2.0 fine-tuning***

## I. Introduction

Speech Emotion Recognition (SER) is a fundamental field of study within affective computing and human–computer interaction, focused on automatically identifying a speaker's emotional state from vocal signals. With its ability to enhance natural communication between humans and machines, SER has demonstrated broad application potential in healthcare monitoring, in-vehicle safety systems, intelligent customer service, educational assistance, and virtual human interaction [1, 2].

Early approaches primarily utilized manually engineered audio attributes, including Mel-Frequency Cepstral Coefficients (MFCCs), pitch, energy, and resonance peaks, alongside traditional classifiers such as Support Vector Machines (SVMs) and Gaussian Mixture Models (GMMs) [3,4]. Although effective in constrained scenarios, these methods struggle with cross-speaker, cross-lingual, and noise-robust emotion recognition. The advent of deep learning has revolutionized the field: architectures including Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), as well as Transformers have significantly boosted performance on SER tasks [5-7]. Moreover, integrating residual learning and attention mechanisms further enhances the modeling of emotional cues [8, 9].

In recent years, self-supervised learning (SSL) frameworks, exemplified by wav2vec 2.0 [10] and HuBERT [11], have enabled pretraining on large-scale unlabeled speech corpora, producing robust and transferable speech representations. These models achieve remarkable performance in low-resource SER scenarios, mitigating data scarcity issues. Meanwhile, end-to-end architectures directly model raw waveforms [12,13], eliminating complex feature engineering and improving model generalization. Beyond representation learning, fusion strategies have emerged as another critical direction to boost emotion recognition performance. By integrating features across multiple modalities or hierarchical levels, fusion models can exploit complementary information. For instance, multimodal fusion networks, feature excitation–aggregation models [14], and Aural Transformers [15] have demonstrated substantial gains in affective speech understanding.

Despite these advances, existing SER methods still face three major challenges:

First, single-path feature modeling often fails to jointly capture both time–frequency structures and contextual dependencies, leading to incomplete emotion representation.

Second, while SSL models like wav2vec 2.0 provide generalizable speech embeddings, they frequently struggle to extract emotion-specific fine-grained cues from limited labeled corpora, thus limiting recognition accuracy.

Finally, most fusion techniques remain confined to simple feature-level concatenation or independent decision-level integration, without fully leveraging the complementarity among heterogeneous representations.

To address the aforementioned limitations, this study proposes a parallel dual-branch fusion architecture for Speech Emotion Recognition. One branch takes spectrograms as input and employs a CNN–Transformer hybrid network to model

*Corresponding author.

time–frequency dependencies, while the other branch processes raw waveforms using the pretrained wav2vec 2.0 model to extract contextualized speech representations. The outputs of both branches are then integrated at the model level through a decision-fusion mechanism, followed by an emotion classifier.

The key contributions of this study can be summarized as follows:

*1) Parallel dual-branch modeling framework*: A unified architecture is designed, where the spectrogram branch and wav2vec 2.0 branch operate in parallel to capture complementary time–frequency and contextual cues, enabling multi-level emotional representation learning.

*2) Enhanced emotion-related feature representation*: The introduction of the spectrogram branch provides explicit time–frequency information, compensating for the limited emotional sensitivity of wav2vec 2.0 when fine-tuned on small-scale datasets. This enables the model to more comprehensively identify detailed features specific to emotions.

*3) Improved decision-level fusion mechanism*: An optimized model-level fusion strategy designed to integrate the probabilistic outputs of both branches, fully leveraging their representational complementarity. Compared with single-path or conventional fusion schemes, the proposed approach demonstrates superior recognition accuracy and robustness.

The rest of this study is structured as follows: Section II reviews related studies on speech emotion recognition. Section III presents the proposed parallel dual-branch architecture, including the spectrogram branch, wav2vec 2.0 branch, and the fusion strategy. Section IV describes the experimental setup and performance analysis, covering dataset description, data augmentation, comparative methods, and ablation experiments. Finally, Section V offers the conclusion and outlines potential directions for future work.

## II. Related Work

### A. Handcrafted Feature-Based Methods

Initial studies on Speech Emotion Recognition (SER) mainly depended on manually designed acoustic features such as MFCCs, energy, pitch, and formants [3]. Huang et al. combined differential MFCC features with a BiLSTM–CNN hybrid network, attaining an accuracy exceeding 81% on the RAVDESS dataset [16]. Similarly, Gu et al. introduced a multi-feature fusion strategy that performed well on both Tibetan and multilingual datasets [17]. Mishra et al. further enhanced performance using MFCC entropy features [3]. Additionally, sinusoidal model [18] and cross-lingual feature extraction [19] have also been explored for SER.

Although these handcrafted approaches achieved promising results on small-scale datasets, they rely heavily on manual feature design and exhibit limited representational capacity, making it difficult to capture complex time–frequency variations and contextual dependencies. Consequently, their generalization ability under noisy or cross-lingual conditions remains insufficient.

### B. Deep Learning-Based Methods

Deep learning methodologies have significantly advanced SER. CNNs excel at capturing local spectro-temporal features [5,20]; RNNs/LSTMs are well-suited for sequential modeling [7]; and Transformers effectively learn long-range dependencies [6]. Liu et al. proposed Res-Transformer, which integrates residual connections into a Transformer encoder, achieving 84.89% accuracy on RAVDESS [8]. Wei et al. developed a parallel CNN–Transformer architecture that further improved SER performance [9]. Similarly, Zhang et al. [21] and Issa et al. [5] confirmed the efficacy of deep CNNs for emotional speech analysis.

Despite these advances, deep learning models often require large-scale labeled datasets and tend to overfit in low-resource scenarios. Moreover, different architectures have complementary strengths—CNNs for local detail modeling and Transformers for global dependency learning—yet it remains challenging to balance both aspects within a single network.

### C. Self-Supervised and End-to-End Methods

Trigeorgis et al. introduced an end-to-end SER framework that directly models raw waveforms without handcrafted preprocessing [12]. Nfissi et al. proposed the CNN-n-GRU model, which achieved outstanding results in waveform-based SER [13]. Recently, self-supervised learning models such as wav2vec 2.0 [10] and HuBERT [11] have demonstrated strong potential by pretraining on massive unlabeled corpora to learn robust and transferable speech representations. Pepino et al. combined wav2vec 2.0 with Transformer encoders for emotion recognition, further validating the advantages of self-supervised representations [22].

These methods avoid complex feature engineering and achieve improved generalization, yet they still struggle to extract emotion-specific fine-grained cues in small-sample settings. Moreover, purely end-to-end approaches often neglect the explicit time–frequency patterns inherent in spectrograms, which are essential for accurate emotional inference.

### D. Feature and Multimodal Fusion Methods

Feature fusion and multimodal integration have recently drawn increasing attention.

Qi et al. proposed MFGCN [1] and AFEA-Net [14], demonstrating the advantages of combining acoustic, visual, and linguistic cues. Luna-Jiménez et al. introduced a multimodal Aural Transformer framework [15], while Mustaqeem et al. proposed AAD-Net [23], achieving excellent end-to-end recognition performance. Smietanka et al. [24] enhanced the embedding representations by leveraging low-level features, and Zhang et al. [25] proposed RELUEM, a model grounded in reinforcement learning, which improved emotional feature discrimination.

Nevertheless, most existing fusion methods are limited to simple feature-level concatenation or independent decision-level fusion, lacking the ability to fully exploit inter-feature complementarity. Furthermore, some approaches involve high computational complexity, which makes it challenging to achieve a balance between effectiveness and efficiency.

In summary, existing SER approaches face several inherent limitations. It remains highly valuable to develop a fusion architecture capable of jointly integrating contextual information and time–frequency representations, thereby achieving comprehensive and robust emotion understanding.

## III. METHODS

### A. Overall Architecture

The proposed parallel dual-branch fusion architecture consists of two independent yet complementary feature extraction pathways:

- wav2vec 2.0 branch: This branch fine-tunes a pretrained wav2vec 2.0 model to obtain context-aware deep speech representations, effectively capturing semantic and prosodic cues directly from raw audio waveforms.
- Spectrogram branch: Based on 48 kHz Mel-spectrogram inputs, this branch employs a CNN–Transformer hybrid network to extract explicit time–frequency features, enhancing the model's ability to detect local changes in the spectrum and variations over time.

Each branch independently produces an emotion prediction, and their probabilistic outputs are subsequently integrated through a decision-level fusion mechanism to yield the final classification result. The comprehensive framework of the proposed model is depicted in Fig. 1.

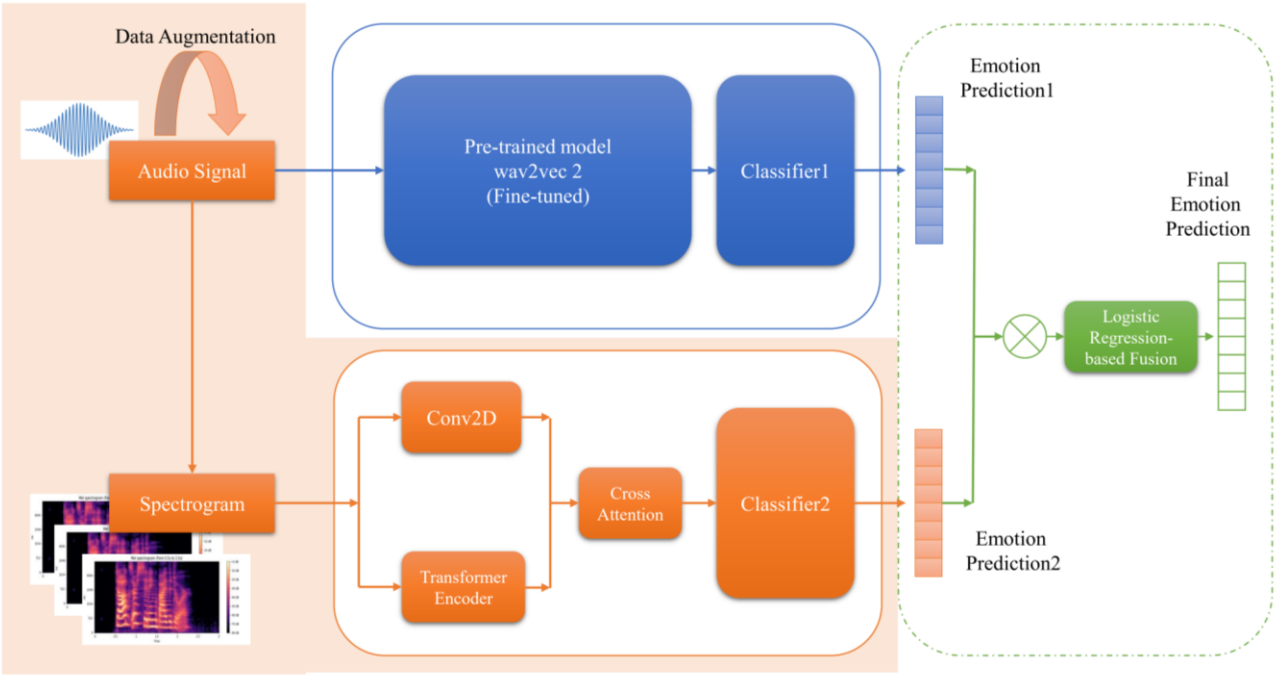



Fig. 1. General architecture of the proposed model.

### B. Input Preprocessing

To ensure the complementarity between the signal-level and time–frequency-level representations, we designed a systematic data preprocessing pipeline. This process aims to generate dual-path inputs compatible with both the wav2vec 2.0 and spectrogram branches, while enhancing the model's ability to generalize and its robustness through noise augmentation and feature normalization.

Specifically, all raw speech recordings were first trimmed or padded to a uniform duration of three seconds, ensuring consistent input length across samples. To satisfy the requirements of the two branches, a dual-sampling strategy was adopted: on one hand, each waveform was downsampled to 16 kHz for the wav2vec 2.0 branch to match its pretrained sampling rate; on the other hand, the original 48 kHz high-resolution audio was retained to compute 128-dimensional log-Mel spectrograms, providing richer acoustic details for the spectrogram branch.

During data augmentation, additive white Gaussian noise (AWGN) was injected into the 48 kHz waveforms, with random signal-to-noise ratios (SNRs) applied to generate multiple augmented samples. Each augmented waveform was subsequently converted into both a 16 kHz version and its corresponding 48 kHz Mel-spectrogram, ensuring dual-path consistency while effectively expanding the training set. This strategy enhances the model's adaptability and resilience in acoustically demanding conditions.

Finally, to mitigate amplitude variations and dynamic range imbalances across recordings, the spectrogram features were standardized before being fed into the network. The normalization parameters were estimated from the training dataset and consistently applied uniformly to both the validation and test datasets, maintaining a stable feature distribution. Such normalization facilitates smoother gradient propagation and faster convergence during optimization.

Through this preprocessing pipeline, the speech data are jointly aligned in temporal and spectral domains, establishing a unified and robust foundation for parallel dual-branch feature extraction and subsequent decision-level fusion.

### C. wav2vec 2.0 Branch

The wav2vec 2.0 branch is designed to learn context-dependent emotional representations directly from raw audio waveforms. Unlike conventional methods that rely on handcrafted acoustic descriptors or shallow convolutional features, wav2vec 2.0 leverages self-supervised learning techniques applied to extensive unlabeled speech corpora. Through this paradigm, the model acquires high-level semantic and prosodic patterns of speech without explicit emotion annotations, enabling strong transferability in low-resource SER scenarios.

*1) Model principle*: The wav2vec 2.0 framework consists of two major parts: a convolutional feature encoder and a context network based on the Transformer model. The convolutional encoder converts a continuous speech waveform $\mathbf{x} = [x_1, x_2, \dots, x_T]$ into a low-dimensional latent representation $Z$ through multiple one-dimensional convolutional layers:

$$Z = f_{enc}(\mathbf{x}) \in \mathbb{R}^{L \times d_z} \tag{1}$$

where, $L$ signifies the quantity of frames after temporal downsampling, while $d_z$ indicates the dimensionality of the convolutional features. This stage primarily captures local acoustic structures and short-term energy variations in the waveform.

The Transformer encoder further models global dependencies within the feature sequence:

$$H = f_{tr}(Z) = [h_1, \dots, h_L],\ h_i \in \mathbb{R}^{768} \tag{2}$$

The Transformer model learns long-range contextual relationships among speech frames by employing the multi-head self-attention mechanism, effectively encoding intonation, rhythm, and prosodic fluctuations that are strongly correlated with emotional expression.

In its pretraining stage, wav2vec 2.0 learns rich speech context representations by reconstructing masked speech frames through contrastive loss on large-scale unlabeled corpora. In the present study, we implement partial fine-tuning of the pretrained wav2vec 2.0 model on the downstream emotion recognition task, thereby activating its latent emotion-discriminative capability while preserving the general acoustic knowledge learned during pretraining.

*2) Partial fine-tuning strategy*: Since emotional speech datasets are typically small in scale, fully fine-tuning all parameters of the wav2vec 2.0 model can often result in overfitting. To address this issue, a systematic comparison was conducted across twelve fine-tuning configurations, ranging from unfreezing only the topmost layer to unfreezing all twelve Transformer layers. The experimental results revealed that unfreezing the last three Transformer encoder layers together with the feature projection layer achieved the best validation performance and the most stable convergence.

Accordingly, this study adopts a "three-layer unfreezing" partial fine-tuning strategy, in which only the high-level semantic parameters are updated, while the lower-level acoustic encoder remains frozen. Let the model parameters be denoted as $\Theta = \{\Theta_f, \Theta_t, \Theta_p\}$, where $\Theta_t^{(K-3:K)}$ represents the trainable parameters of the top three Transformer layers, and $\Theta_p$ denotes the feature projection parameters. The training objective is defined as follows:

$$\min_{\Theta_t^{(K-3:K)}, \Theta_p} \mathcal{L}(y, p) \tag{3}$$

where, $\mathcal{L}$ represents the cross-entropy loss function, $y$ is the true emotion label, and $p$ signifies the predicted probability distribution over the emotional states.

*3) Utterance-level representation and classification structure*: The Transformer encoder outputs a sequence of frame-level hidden states, which are aggregated via temporal mean pooling to form an utterance-level embedding:

$$\bar{h} = \frac{1}{L}\sum_{i=1}^{L} h_i \tag{4}$$

where, $\bar{h} \in \mathbb{R}^{768}$ denotes the global representation of the entire speech segment.

During the classification stage, a two-layer multilayer perceptron (MLP) is employed to project and nonlinearly transform $\bar{h}$:

$$a_1 = \sigma(W_1\bar{h} + b_1) \tag{5}$$

$$a_2 = \sigma(W_2 a_1 + b_2) \tag{6}$$

$$P_{w2v} = \mathrm{softmax}(W_3 a_2 + b_3) \tag{7}$$

where, $\sigma(\cdot)$ denotes the ReLU activation function. A Dropout layer with a ratio of 0.3 is applied after each fully connected layer to alleviate overfitting. The dimensional transitions are: $\bar{h}(768) \to a_1(256) \to a_2(256) \to P_{w2v}(8)$.

This hierarchical mapping progressively compresses the semantic space while enhancing the discriminability of emotional representations. By integrating nonlinear transformations and moderate regularization, the classifier effectively bridges the contextual embedding from wav2vec 2.0 to the eight-category emotion prediction task.

*4) Functional role and advantages of the branch*: The wav2vec 2.0 branch is capable of directly capturing emotion-related prosodic variations, energy fluctuations, and contextual dependencies from raw waveforms without relying on handcrafted acoustic features. Compared with the spectrogram branch, its advantages can be summarized as follows:

- Pretrained structural awareness: Benefiting from large-scale self-supervised pretraining, the model possesses a strong perceptual understanding of speech structure, enabling rapid adaptation to downstream emotion recognition tasks.
- Efficient partial fine-tuning: By employing the partial fine-tuning strategy, the number of trainable parameters is effectively reduced, thereby enhancing the model's

robustness and generalization when operating with limited data.

- Complementary abstract representation: The generated utterance-level embeddings are highly semantic and abstract, complementing the explicit time–frequency features extracted by the spectrogram branch.

Finally, the emotion probability vector $P_{w2v}$ produced by this branch is integrated with the spectrogram branch output at the decision-fusion stage, providing high-level semantic support for the subsequent multimodal emotion recognition process.

### D. Spectrogram Branch

The spectrogram branch is designed to extract explicit acoustic representations from the time–frequency domain, thereby complementing the semantic and contextual modeling capability of the wav2vec 2.0 branch. This branch builds upon our previous work [9], with structural refinements and fusion optimization introduced in the current study. As illustrated in Fig. 1 (highlighted in light orange), it encompasses the complete processing pipeline—from data preprocessing and Mel-spectrogram extraction to convolution–attention hybrid modeling.

*1)* Input representation: The input to this branch is a log-Mel spectrogram generated from 48 kHz audio, comprising 128 Mel filters and 278-time frames. The resulting input tensor has a shape of 1×128×278, corresponding to the channel, frequency, and time dimensions, respectively. This representation inherits the spectrogram construction and additive white Gaussian noise (AWGN) augmentation strategy proposed in [9], while further improvements are introduced in feature normalization and temporal alignment, ensuring that the spectrogram input is strictly synchronized with the waveform input used in the wav2vec 2.0 branch.

*2)* Network architecture: The spectrogram branch adopts a hybrid Convolution–Transformer–Cross-Attention framework for hierarchical acoustic modeling.

*a) Convolutional module*: The convolutional front-end consists of four residual CNN blocks. Each block comprises two convolutional layers, followed by Batch Normalization and ReLU activation functions. These blocks are interconnected through skip connections, which serve to improve gradient propagation and facilitate the reuse of features. This module captures short-term energy variations and formant structures while progressively downsampling the feature map. The feature dimensions evolve as follows: (1×128×278) → (16×64×139) → (32×32×69) → (64×16×34) → (64×8×17).

*b) Transformer encoder*: A four-layer Transformer encoder is utilized to effectively model long-range temporal dependencies. The convolutional output is first downsampled using a 2D pooling window of [2, 4], resulting in a 64 × 69 feature map. Before entering the Transformer, the dimensions are rearranged to treat the time frames as the primary sequential dimension, fitting the encoder's input format. After encoding, the output tensor (shape = 69×64) is restored to batch-first form. Each Transformer layer has a hidden size of 64 and 4 attention heads, enabling it to model global rhythm and intensity variations of emotional speech along the temporal dimension.

*c) Cross-attention module*: To achieve complementary modeling between convolutional and sequential features, a Cross-Attention module is employed. Here, the convolutional features function as queries (Q), while the Transformer outputs act as keys (K) and values (V). Through attention-weighted interaction, the two feature spaces are fused to produce a 64-dimensional integrated representation that simultaneously encodes time–frequency detail and contextual dependency. The fused vector is then concatenated with the flattened convolutional feature, forming an 8768-dimensional joint representation, which is fed into the final classifier.

*3)* Classification and feature evolution: The classification head operates on a joint representation obtained by concatenating the flattened CNN feature with a 64-dimensional vector produced by the Cross-Attention module. Concretely, the spectrogram input 1×128×278 is processed by the convolutional front-end into a feature map 64×8×17, which is flattened to 8704 dimensions. In parallel, the raw spectrogram is downsampled by a 2-D pooling [2, 4] to a 64×69 map and passed through a 4-layer Transformer (d=64, 4 heads), yielding a sequence of shape 69×64. Using Cross-Attention (query = CNN flatten 8704 → proj., key/value = Transformer output 69×64), the model produces a 64-dimensional fused vector that captures contextual dynamics. The two parts are then concatenated to form an 8768-dimensional feature (8768=8704⊕64), which a single fully connected layer maps to the 8 emotion categories.

*4)* Functional role: Within the overall system, the spectrogram branch is responsible for explicit acoustic modeling, focusing on energy distribution, formant resonance, and prosodic patterns that characterize emotional tone at the physical level. When integrated with the semantic representations learned by the wav2vec 2.0 branch, the two pathways form a complementary feature hierarchy—from time–frequency to semantic space—providing a more stable and discriminative foundation for final emotion classification.

### E. Decision-Level Fusion

To fully exploit the complementary strengths of the two branches at the semantic and acoustic levels, a logistic regression–based fusion mechanism is introduced after the classification layer. This approach maintains the independence of each branch while achieving optimal emotional decision integration through probability-space learning.

Let, the wav2vec 2.0 branch and the spectrogram branch respectively output the predicted probability vectors over $C$ emotion classes as: $P_{w2v} = [p_{w2v}^{(1)}, p_{w2v}^{(2)}, \ldots, p_{w2v}^{(C)}]$, $P_{spec} = [p_{spec}^{(1)}, p_{spec}^{(2)}, \ldots, p_{spec}^{(C)}]$, where $C$=8 denotes the number of emotion categories, and each element represents the confidence score assigned by the corresponding branch.

The two vectors are concatenated in probability space to form the fusion input feature:

$$f = [P_{w2v}; P_{spec}] \in \mathbb{R}^{2C} \quad (8)$$

where, "[;]" denotes vector concatenation. This combined vector encodes the independent judgments of the semantic and acoustic branches for each emotion category.

The logistic regression model performs a linear mapping followed by softmax normalization to generate the final fused prediction:

$$P_{fuse} = softmax(W_{LR} f + b_{LR}) \quad (9)$$

here, $W_{LR} \in \mathbb{R}^{C \times 2C}$ and $b_{LR}$ denote the trainable parameters and bias term of the logistic regression layer. This formulation allows the model to learn optimal linear combinations of branch-level probabilities across emotion classes.

During the training process, the cross-entropy loss is employed to minimize the divergence between the predicted and true distributions:

$$\mathcal{L}_{\text{fuse}} = -\sum_{i=1}^{C} y_i \log\left(p_{\text{fuse}}^{(i)}\right) \quad (10)$$

where, $y = [y_1, \ldots, y_C]$ denotes the one-hot ground-truth label, while $p_{\text{fuse}}^{(i)}$ represents the predicted probability for class $i$. By maximizing the likelihood of the correct class, the logistic regression layer learns the optimal fusion parameters $[W_{LR}, b_{LR}]$.

Unlike fixed-weight or gating-based strategies, the proposed fusion mechanism possesses learnable and adaptive parameters, enabling it to dynamically adjust the relative contributions of both branches across different emotion types.

For instance, when the wav2vec 2.0 branch excels at semantically driven categories (e.g., happy, sad), and the spectrogram branch better captures acoustically dominant emotions (e.g., angry, calm), the logistic regression model automatically rebalances their influences through parameter optimization.

The resulting fused probability vector $P_{fuse}$ serves as the final system output, providing a unified basis for performance evaluation and result analysis. Experimental results demonstrate that this fusion strategy, while maintaining computational efficiency, yields notable improvements in overall accuracy and robustness.

## IV. Experiments

### A. Dataset and Data Augmentation Strategy

All experiments were conducted on the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset. This corpus comprises recordings from 24 professional actors, evenly divided by gender (12 male and 12 female), and covers eight emotion categories: happy, fearful, surprised, sad, neutral, angry, disgust, and calm. Each utterance is expressed at two intensity levels, namely normal and strong. This study utilized exclusively the speech subset, consisting of 1,440 audio samples with an average duration of approximately 3 to 4 seconds per clip.

To ensure balanced emotion distribution, a stratified random split was employed, dividing the data into training, validation, and test subsets with an approximate proportional distribution of 8:1:1. During training, additive white Gaussian noise (AWGN) augmentation was applied only to the training set to enhance model robustness. For each training utterance, three augmented versions were generated with randomly selected signal-to-noise ratios (SNRs) of 10, 20, and 30 dB. The validation and test datasets remained clean and unaltered, preserving the original audio quality for fair evaluation.

The detailed data split statistics are summarized in Table I.

TABLE I Data Partition and Sample Statistics

| Subset | Number of Samples | Percentage (%) | Description |
|---|---|---|---|
| Training set | 1147 | 80 | Used for model training with AWGN-based augmentation (3 SNR levels: 10, 20, 30 dB) |
| Validation set | 143 | 10 | Used for adjusting hyperparameters and choosing the best model |
| Test set | 150 | 10 | Used for the ultimate performance assessment under clean conditions |
| Total | 1440 | 100 | Speech subset of RAVDESS (8 emotion classes × 24 speakers) |

Note: The training set was augmented threefold using additive white Gaussian noise (AWGN) at various SNR levels, while the validation and test datasets remained unaltered to ensure fair evaluation.

### B. Experimental Setup

*1) Experimental environment*: All experiments were conducted on a single Linux workstation. The system configuration is summarized as follows:

- Operating System: Ubuntu 22.04
- GPU: NVIDIA RTX 5880 Ada (48 GB VRAM)
- Python Version: 3.10
- PyTorch Version: 2.1.0
- CUDA Version: 12.4

This hardware and software setup provides sufficient computational capacity for parallel training and fusion experiments involving the dual-branch architecture.

*2) Training configuration*: Model training was performed in three stages:

- Training the wav2vec 2.0 branch
- Training the CNN–Transformer spectrogram branch
- Optimizing the logistic regression fusion layer

The main hyperparameter configurations for each stage are summarized in Table II.

TABLE II TRAINING PARAMETERS FOR DIFFERENT STAGES

| Parameter | wav2vec 2.0 Branch | Spectrogram Branch | Logistic Regression Fusion Layer |
|---|---|---|---|
| Optimizer | AdamW | SGD | Adam |
| Initial Learning Rate | $1\times10^{-4}$ | $1\times10^{-2}$ | $5\times10^{-4}$ |
| Batch Size | 32 | 32 | 32 |
| Epochs | 120 | 500 | 30 |
| Loss Function | CrossEntropyLoss | CrossEntropyLoss | CrossEntropyLoss |
| Learning Rate Scheduler | CosineAnnealing | StepLR (step=10, $\gamma$=0.7) | Fixed learning rate |
| Early-Stopping Criterion | Stop if validation accuracy does not improve for 10 epochs | Stop if validation loss does not decrease for 25 consecutive epochs | Stop if validation Macro-F1 remains stable for 5 epochs |

All models employed CrossEntropyLoss as the classification objective. During training, gradient clipping (clip = 1.0) was applied to prevent gradient explosion and ensure stable optimization. The input utterances were standardized to a fixed duration of 3 seconds, with sampling rates of 16 kHz for the wav2vec 2.0 branch and 48 kHz for the spectrogram branch, respectively.

### C. Comparative Experiments

To verify the effectiveness of the proposed parallel dual-branch fusion architecture, this section presents comparative experiments against multiple baseline models and fusion strategies. Both quantitative and qualitative analyses are employed to assess the performance improvements and the interpretability of the results.

*1) Comparative methods and evaluation criteria*: To objectively evaluate the performance advantages of the proposed model, several representative Speech Emotion Recognition (SER) methods from recent years (2023–2025) were selected as comparative baselines.

The selection of baseline models follows these principles:

- Recency and relevance: methods published within the last three years to reflect the current research trends;
- Completeness of evaluation: models reporting comprehensive metrics (Precision, Accuracy, Recall, and F1-score) for fair comparison;
- Speaker-dependent setting: consistent with this study's evaluation protocol;
- Audio-only modality: models that perform SER using acoustic features only, without incorporating visual or textual information;
- Dataset consistency: all methods are evaluated on the RAVDESS dataset to ensure comparability under identical data conditions.

The comparative methods are summarized as follows:

*a) MFCC-fusion (2023) [4]*: This method integrates MFCC, ΔMFCC, and $\Delta^2$MFCC features via a PCA-based contribution fusion strategy and employs a BiLSTM–CNN hybrid network for emotion recognition.

*b) AFEA-Net (2025) [14]*: AFEA-Net is an audio-based SER framework that fuses low-level Fbank features and high-level WavLM embeddings through an excitation-and-aggregation mechanism under a multi-task learning framework, effectively enhancing emotion-relevant feature alignment.

*c) CNN-n-GRU (2025) [13]*: CNN-n-GRU is an end-to-end model that directly learns emotional representations from raw speech waveforms by integrating convolutional layers to capture local features and gated recurrent units (GRUs) to model temporal dependencies.

*d) MFGCN (2025) [1]*: MFGCN introduces a multimodal fusion graph convolutional network that captures semantic–emotional dependencies among acoustic features. For fair comparison, only the audio-stream branch (MFGCN-a) is considered here, which utilizes WavLM acoustic embeddings and enhances emotion classification through multi-perspective fusion.

*e) IGRFXG (2025) [26]*: IGRFXG is an ensemble-based feature selection framework integrating Random Forest, XGBoost ranking, and Information Gain mechanisms to select the most informative acoustic descriptors. The selected feature kernel is used with SVM and MLP classifiers, achieving strong results on audio-only RAVDESS data.

*f) RELUEM (2025) [25]*: This model combines deep reinforcement learning with a convolution–recurrent hybrid structure, enabling adaptive speech emotion recognition through learned decision policies. It dynamically adjusts classification strategies to capture emotional transitions, achieving high stability and accuracy on the RAVDESS dataset.

To comprehensively evaluate the performance of all models on the SER task, four widely used classification metrics were adopted: Accuracy, Precision, Recall, and F1-score. These metrics jointly reflect the ability to recognize patterns, generalization stability, and class-wise balance across emotional categories.

*2) Experimental results*: Table III summarizes the performance comparison between the proposed dual-branch fusion model and several representative SER methods using the RAVDESS dataset. The findings highlight the proposed framework's ability to enhance both recognition accuracy and feature robustness across emotional categories.

TABLE III PERFORMANCE COMPARISON OF VARIOUS SER METHODS ON THE RAVDESS DATASET (%)

| Method | Year | Accuracy (%) | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| MFCC-Fusion [4] | 2023 | 81.5 | 82.5 | 85.5 | - |
| AFEA-Net [14] | 2025 | 80.3 | 80.8 | 80.6 | 80.4 |
| CNN-n-GRU [13] | 2025 | 86.6 | 87.1 | 86.6 | 86.7 |
| MFGCN [1] | 2025 | 85.7 | 85.7 | 85.1 | 85.4 |
| IGRFXG [26] | 2025 | 79.3 | - | - | - |
| RELUEM [25] | 2025 | 89.5 | 87 | 85 | 86 |
| Ours | 2025 | **92.7** | **92.4** | **92.5** | **92.2** |

From Table III, it can be observed that the most recent state-of-the-art SER methods have achieved recognition accuracies exceeding 80%, with deep feature learning models showing a clear advantage over traditional acoustic-feature-based approaches. Specifically, the conventional MFCC-Fusion (2023) model relies solely on handcrafted low-level acoustic descriptors and achieves an accuracy of 81.5%. In contrast, deep learning–based approaches such as AFEA-Net (2025), CNN-n-GRU (2025), and MFGCN (2025) leverage multi-task learning, temporal feature extraction, and graph-based fusion, reaching accuracies of 80.3%, 86.6%, and 85.7%, respectively. The ensemble feature-selection model IGRFXG (2025) attains 79.3%, while the reinforcement-learning-driven RELUEM (2025) demonstrates outstanding performance with an accuracy of 89.5%, highlighting the potential of dynamic policy optimization in emotion classification.

In comparison, the proposed dual-branch fusion model attains the highest overall performance on all evaluation metrics, achieving 92.7% accuracy, 92.4% precision, 92.5% recall, and 92.2% F1-score. Compared with the strongest baseline, RELUEM, the proposed method yields an accuracy gain of approximately 3.2 percentage points.

This performance improvement primarily stems from the complementary nature of the dual-branch architecture:

- wav2vec 2.0 branch: captures high-level semantic and prosodic cues, providing contextual information critical for emotional differentiation.
- Spectrogram branch: models energy distribution and spectral dynamics, offering structured acoustic representations for emotion discrimination.
- Logistic-regression fusion layer: adaptively re-weights the two feature spaces in the probability domain, balancing semantic and acoustic contributions for optimal decision fusion.

These findings collectively highlight the effectiveness and robustness of the proposed dual-branch fusion framework in speech emotion recognition. To further investigate class-wise performance, the confusion matrix of the fused model applied to the RAVDESS dataset is shown in Fig. 2.

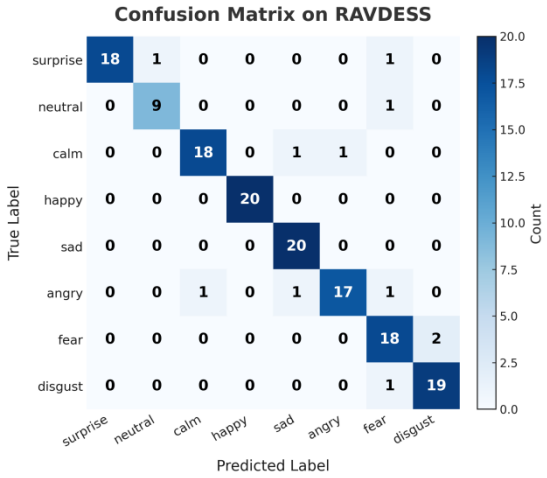



Fig. 2. Confusion matrix of the proposed model applied to the RAVDESS dataset.

Each cell indicates the count of samples corresponding to the true (row) and predicted (column) emotion columns, while the diagonal entries represent correctly classified samples. Overall, the model exhibits balanced performance across all eight emotion categories, with the diagonal cells showing notably higher counts than off-diagonal ones—demonstrating strong discriminative capability and stable classification behavior across emotional states.

As shown in Fig. 2, the model achieves perfect recognition for happy and sad (20/20 correct each). For surprise, 18 out of 20 samples are correctly identified, with 1 misclassified as neutral and 1 as fear. The disgust class records 19/20 correct predictions, with 1 sample misclassified as fear. Fear achieves 18/20 correct, with 2 samples misclassified as disgust. Angry attains 17/20 correct, with 1 sample confused with calm and 1 with sad. Mild confusions occur among low-arousal emotions: neutral has 9/10 correct with 1 misclassified as disgust, while calm has 18/20 correct with 1 misclassified as sad and 1 as angry.

Overall, the model performs best on emotions with pronounced prosodic dynamics (e.g., happy, sad, surprise), whereas residual errors arise between acoustically similar or adjacent categories (e.g., neutral–calm and angry–fear–disgust). This pattern confirms that the proposed dual-branch fusion effectively captures complementary semantic and prosodic cues, delivering stable and interpretable SER performance on this split.

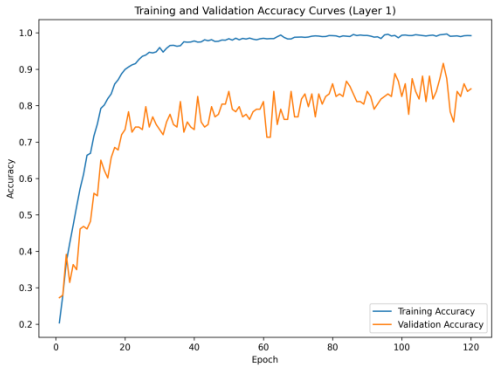
### D. Layer Unfreezing Experiments of wav2vec 2.0

To investigate the contribution of different representation layers within wav2vec 2.0 to the speech emotion recognition task, a systematic layer unfreezing experiment was conducted across its 12 Transformer encoder blocks. In this analysis, the range of trainable parameters was progressively expanded to explore the relationship between feature hierarchy and task adaptability. All experiments were performed under identical training data, optimizer, and hyperparameter settings, with only the number of unfrozen layers varied.

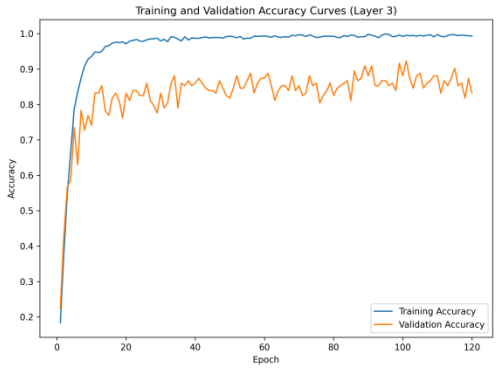
For brevity and representativeness, four typical configurations—unfreezing 1, 3, 5, and 9 layers—are presented and analyzed. Table IV presents a summary of the performance results obtained from the RAVDESS dataset.

TABLE IV PERFORMANCE COMPARISON OF WAV2VEC 2.0 UNDER DIFFERENT UNFREEZING DEPTHS (RAVDESS)

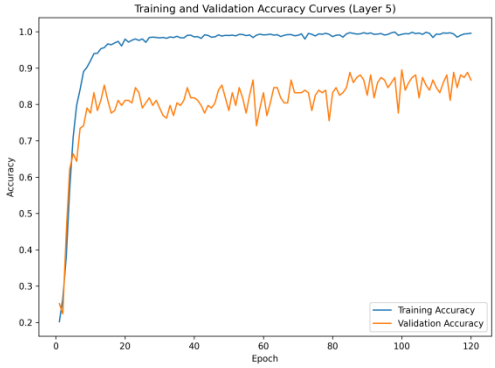
| Unfrozen Layers | Accuracy (%) | Macro Precision (%) | Macro Recall (%) | Macro F1 (%) | Performance Description |
|---|---|---|---|---|---|
| 1 layer | 82.67 | 83.63 | 80.63 | 81.22 | Dominated by shallow acoustic cues; limited semantic perception |
| 3 layers | **88.67** | **88.22** | **88.75** | **88.19** | Best overall performance; balanced semantic–acoustic representation |
| 5 layers | 88 | 87.22 | 86.88 | 86.81 | Slight overfitting observed as deeper layers are updated |
| 9 layers | 17.33 | 5.0 | 16.25 | 7.42 | Model collapse due to the disruption of pretrained representations |



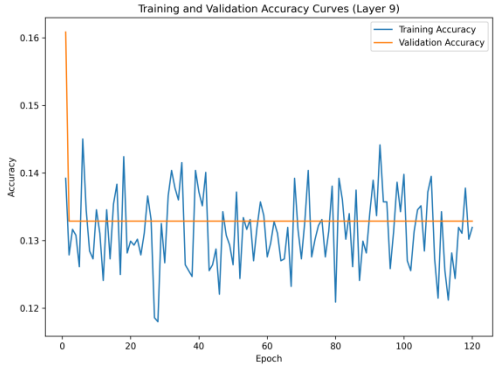

(a) Unfreezing 1 layer.




(b) Unfreezing 3 layers




(c) Unfreezing 5 layers.




(d) Unfreezing 9 layers

Fig. 3. Training and validation accuracy curves of wav2vec 2.0 fine-tuning.

As shown in Table IV, model performance exhibits a distinct "rise–then–fall" trend as the number of unfrozen layers increases. When only the first layer is unfrozen, the model mainly relies on low-level acoustic patterns, achieving 82.67% accuracy, which reflects limited emotional abstraction capability. Performance peaks when three layers are unfrozen, reaching 88.67% accuracy and 88.19% Macro-F1, indicating that the integration of low-level acoustic and mid-level semantic representations yields the most effective balance.

Further unfreezing up to five layers leads to a slight decline, suggesting feature drift and a growing risk of overfitting. When more than nine layers are unfrozen, the model collapses (accuracy drops to 17.33%), as the pretrained representations become severely distorted, undermining semantic stability and generalization ability.

The corresponding training and validation curves under different unfreezing configurations are illustrated in Fig. 3, further confirming the degradation pattern observed with excessive parameter unfreezing.

In summary, across the full range of 1–12 layers unfreezing experiments, the configuration with the last three layers unfrozen (Layer-3 setting) achieved the best overall performance, highlighting the crucial role of mid-level semantic representations in wav2vec 2.0 for speech emotion recognition. These findings indicate that appropriately controlling the unfreezing depth not only maximizes the utilization of pretrained knowledge but also enhances both the transferability and stability of the model in downstream emotional tasks.

### *E. Fusion Strategy Comparison Experiments*

To evaluate the performance differences among various fusion mechanisms within the dual-branch architecture, five groups of experiments were designed while keeping the backbone structure identical. The first two groups represent single-branch models, used to assess the independent modeling capability of the acoustic and semantic branches. The remaining three groups implement different fusion strategies, aiming to analyze the effectiveness of probability-space fusion. All experiments were conducted on the RAVDESS dataset, and the performance was evaluated using Accuracy, Precision, Recall, and Macro-F1 as the primary metrics. The detailed results are presented in Table V.

TABLE V PERFORMANCE COMPARISON OF DIFFERENT FUSION STRATEGIES ON THE RAVDESS DATASET

| Fusion Strategy | Accuracy (%) | Macro Precision (%) | Macro Recall (%) | Macro F1 (%) | Description |
|---|---|---|---|---|---|
| Spectrogram-only model | 77 | 76 | 74 | 74 | Uses only acoustic spectrogram features; lacks semantic modeling capability |
| wav2vec 2.0-only model (3-layer unfreeze) | 88.67 | 88.22 | 88.75 | 88.19 | Utilizes semantic representations; strong emotional discriminability |
| Weighted fusion | 90.67 | 90.4 | 90.62 | 90.32 | Linear weighted combination of branch outputs; improves overall performance |
| Gated fusion | 88 | 87.97 | 86.87 | 87.08 | Employs dynamic weight allocation; slightly unstable under small-sample conditions |
| Logistic regression fusion | 92.67 | 92.36 | 92.5 | 92.24 | Adaptive probability-space fusion; achieves the best overall performance |

As shown in Table V, the choice of fusion strategy significantly impacts model performance.

Under single-branch conditions, the spectrogram-only model achieves only 77.0% accuracy, indicating that low-level acoustic features alone are insufficient to represent complex emotional states. In contrast, the wav2vec 2.0-only model achieves 88.7% accuracy, demonstrating that semantic-level representations derived from pretrained models offer stronger discriminative power for emotion classification.

Among the fusion strategies, the weighted fusion approach linearly combines the outputs of both branches, improving performance to 90.7%, which confirms the complementarity between semantic and acoustic features. The gated fusion mechanism adaptively adjusts feature weights based on the input but exhibits training instability under limited data, leading to a slight drop in accuracy (88.0%). In comparison, the proposed logistic regression fusion achieves adaptive weighting in probability space with a smaller parameter scale and more stable convergence. It attains the best results across all metrics (Accuracy = 92.7%, Macro-F1 = 92.2%), demonstrating the superiority and robustness of the proposed fusion mechanism.

These findings indicate that probability-space fusion effectively integrates the discriminative strengths of both feature domains, achieving simultaneous improvements in accuracy and stability without increasing model complexity. To provide a more intuitive comparison, Fig. 4 illustrates the performance differences among various fusion strategies.

As shown in Fig. 4, the proposed logistic regression fusion method significantly outperforms other strategies in both Accuracy and Macro-F1, exhibiting stronger feature integration and discriminative capability. This further validates the effectiveness and efficiency of the proposed adaptive fusion framework developed for speech emotion recognition.

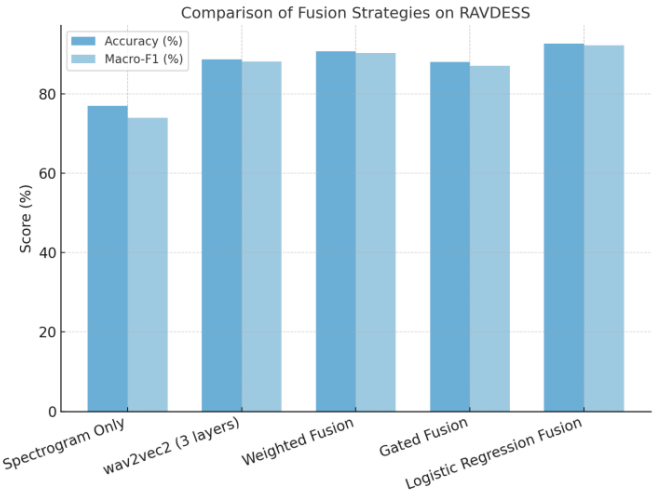



Fig. 4. Comparison of different fusion strategies on the RAVDESS dataset.

### *F. Discussion*

Section IV(C) to Section IV(E) have systematically verified the efficacy of the proposed parallel dual-branch fusion architecture through comprehensive experimental results. To attain a deeper understanding of the underlying mechanisms driving its performance improvements, this section provides a detailed analysis from three complementary perspectives—

architectural design, optimization strategy, and feature fusion mechanism—and further discusses the empirical findings from the layer unfreezing and fusion strategy experiments.

*1) Model mechanism and sources of performance improvement*: The superior performance of the proposed model primarily stems from the synergistic effects of three key factors: feature complementarity, branch-wise decoupled optimization, and probability-space fusion.

*a) Complementarity of feature representations*: The wav2vec 2.0 branch, pretrained on large-scale unlabeled speech corpora, captures high-level semantic and prosodic representations—such as rhythm, intonation variation, and emotional dynamics—thereby enhancing the model's global awareness of emotional cues. Meanwhile, the spectrogram branch, based on a CNN–Transformer hybrid structure, models acoustic energy distributions and harmonic patterns, exhibiting higher sensitivity to fine-grained spectral variations. By jointly focusing on the semantic–prosodic and acoustic–spectral domains, the two branches form a multi-level complementary representation. This enables the fusion layer to learn more discriminative emotion features from multi-perspective information sources.

*b) Branch decoupling and optimization stability*: A stage-wise training strategy is employed—each branch is optimized independently before fusion. This decoupled design effectively mitigates the gradient interference and feature competition commonly observed in end-to-end joint training, allowing each branch to converge independently under its optimal learning rate and scheduling policy. As a result, the model generates more stable and diverse emotional representations, leading to enhanced training stability and generalization performance compared with single-stage joint optimization.

*c) Effectiveness of probability-space fusion*: During fusion, logistic regression is applied to learn the optimal weighting of branch-level Softmax probability vectors. Compared with complex gating or attention-based mechanisms, this lightweight design achieves adaptive fusion with significantly fewer parameters. It accelerates convergence, improves stability, and achieves the highest recognition accuracy (92.7%) and Macro-F1 (92.2%) on the RAVDESS dataset. As illustrated in Fig. 4 and Table V, probability-space fusion achieves an optimal balance between performance and robustness.

*2) Insights from the layer unfreezing experiments*: The wav2vec 2.0 layer unfreezing experiments (Section IV D) further reveal the relationship between model performance and the depth of pretrained feature adaptation. Results indicate a clear "rise–then–fall" trend as the number of unfrozen layers increases, with the three-layer unfreezing configuration (Layer-3) achieving the best performance (Accuracy = 88.67%, Macro-F1 = 88.19%).

This finding underscores that mid-level semantic representations in wav2vec 2.0 possess the highest transferability for emotion recognition tasks. Shallow layers mainly encode low-level acoustic information with limited semantic abstraction, while deeper layers tend to introduce task-specific bias, causing overfitting and feature drift.

Therefore, moderate unfreezing depth effectively balances pretrained stability and task adaptability, yielding more discriminative semantic embeddings for downstream fusion.

*3) Emotion category discriminability and limitations*: As observed from the confusion matrix (Fig. 2), the model achieves near-perfect recognition for high-energy emotions such as happy, sad, and surprise, with diagonal entries approaching 100%. In contrast, mild confusion occurs among low-energy or acoustically similar emotions, particularly neutral–calm and angry–fear–disgust. This suggests that the model exhibits strong discriminative capability for emotions with distinct prosodic variations, yet still faces challenges in distinguishing acoustically overlapping or data-sparse categories.

Two major factors contribute to this phenomenon:

- Several high-arousal negative emotions share overlapping spectral characteristics at the acoustic level, making them inherently harder to separate.
- The current model still faces limitations in feature distribution learning under class-imbalance conditions.

Future work could address these challenges by introducing adversarial learning or emotion-aware reweighting mechanisms, which may enhance recognition performance for ambiguous or boundary emotions.

## V. Conclusion

This study addresses the long-standing challenge in SER—that is, the difficulty of simultaneously capturing time–frequency structures and semantic dependencies using a single feature representation. To this end, a parallel dual-branch fusion architecture is proposed, consisting of a wav2vec 2.0 branch and a CNN–Transformer spectrogram branch, which respectively extract semantic-level and acoustic-level features. A logistic regression fusion layer is further introduced to achieve adaptive weighting in probability space, effectively integrating contextual and time–frequency information. Experimental results on the RAVDESS dataset demonstrate that the proposed model achieves an accuracy of 92.7% and a Macro-F1 score of 92.2%, outperforming the best existing baseline by approximately 3.2 percentage points. The layer unfreezing experiment validates the effectiveness of unfreezing the last three layers of wav2vec 2.0, while the fusion strategy comparison confirms the superiority of probability-space fusion in both performance and stability.

Result analysis shows that the model excels at recognizing high-energy and distinct emotions such as happy, sad, and surprise, with diagonal recognition rates approaching 100% in the confusion matrix. However, slight confusion remains among acoustically similar or data-sparse categories such as neutral–calm and angry–fear. In summary, the proposed framework effectively balances performance and robustness through feature complementarity, branch decoupling, and lightweight fusion, fully validating the effectiveness of parallel dual-branch modeling in SER tasks.

Future work will proceed in three directions:

*1)* Validating the model's transferability on multilingual and cross-domain corpora to enhance generalization;

*2)* Extending the framework to multimodal emotion perception by integrating speech, facial expressions, and textual cues; and

*3)* Pursuing model lightweighting and real-time inference optimization, as well as exploring emotion dynamics modeling and interpretability mechanisms, to promote practical deployment in education, healthcare, and human–computer interaction scenarios.

## ACKNOWLEDGMENT

This research work was supported in part by the Medical Special Cultivation Project of Anhui University of Science and Technology (No. YZ2023H2C011), the National Natural Science Foundation of China (Grant No. 62476005).

## REFERENCES

[1] X. Qi, Y. Wen, P. Zhang, et al., "MFGCN: Multimodal Fusion Graph Convolutional Network for Speech Emotion Recognition," Neurocomputing, vol. 611, p. 128646, 2025.

[2] C. Luna-Jiménez, D. Griol, Z. Callejas, et al., "Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning," Sensors, vol. 21, no. 21, p. 7665, 2021.

[3] S. P. Mishra, P. Warule, and S. Deb, "Speech Emotion Recognition Using MFCC-Based Entropy Feature," Signal, Image and Video Processing, vol. 18, pp. 153–161, 2024.

[4] X. Huang, Q. Du, H. Long, et al., "Speech Emotion Recognition Algorithm Based on MFCC Feature Fusion," Journal of Shaanxi University of Technology, vol. 39, no. 4, pp. 17–25, 2023.

[5] J. Zhao, X. Mao and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks", Biomed. Signal Process. Control, vol. 47, pp. 312-323, 2019.

[6] S. Y. Lin, H. L. Chang, J. J. Hwang, et al., "Automatic audio-based screening system for Alzheimer's disease detection," Proc. SMC, pp. 2770-2775, 2022.

[7] A. Satt, S. Rozenberg, and R. Hoory, "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms," Proc. INTERSPEECH, pp. 1089–1093, 2017.

[8] F. Liu and L. Wang, "Research on Speech Emotion Recognition Based on Res-Transformer Model," Internet of Things Technology, vol. 6, pp. 36–42, 2023.

[9] Z. Wei, C. Ge, C. Su, et al., "A Deep Learning Model for Speech Emotion Recognition," Int. J. Adv. Comput. Sci. Appl. (IJACSA), vol. 16, no. 5, pp. 316–323, 2025.

[10] A. Baevski, H. Zhou, A. Mohamed, et al., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Proc. NeurIPS, 2020.

[11] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, et al., "HuBERT: Self-Supervised Speech Representation Learning," Proc. NeurIPS, 2021.

[12] G. Trigeorgis, F. Ringeval, R. Brueckner, et al., " Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," Proc. ICASSP, pp. 5200-5204, 2016.

[13] N. Nfissi, R. Saidi, A. El Hannani, et al., "From unaltered raw waveform to emotion: Synergizing convolutional and gated recurrent networks for holistic speech emotion analysis," Applied Intelligence, vol. 55, no. 737, 2025.

[14] X. Qi, Q. Song, G. Chen, et al., "Acoustic Feature Excitation-and-Aggregation Network Based on Multi-Task Learning for Speech Emotion Recognition" Electronics, vol. 14, no. 5, p. 844, 2025.

[15] C. Luna-Jiménez, R. Kleinlein, D. Griol, et al., "A Proposal for Multimodal Emotion Recognition Using Aural Transformers," Appl. Sci., vol. 12, no. 1, p. 327, 2022.

[16] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," PLOS ONE, vol. 13, no. 5, p. e0196391, 2018.

[17] Z. Gu, B. Wangdui, and J. Qi, "Tibetan Speech Emotion Recognition Based on Multi-Feature Fusion," Modern Electronic Technology, vol. 46, no. 21, pp. 129–133, 2023.

[18] P. Warule, S. P. Mishra, S. Deb, et al., "Sinusoidal modelbased diagnosis of the common cold from the speech signal," Biomedical Signal Processing and Control, vol. 83, p. 104653, 2023.

[19] S. Bhattacharya, S. Borah, B. K. Mishra, et al., "Emotion Detection from Multilingual Audio," Multimedia Tools Appl., vol. 81, pp. 41309–41338, 2022.

[20] A. M. Badshah, J. Ahmad, N. Rahim, et al., "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," Proc. PlatCon, pp. 1–5, 2017.

[21] S. Zhang, S. Zhang, T. Huang, et al., " Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching," IEEE Access, vol.20, no.6, pp. 1576-1590, 2017.

[22] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings" Proc. INTERSPEECH, pp. 3400-3404, 2021.

[23] Mustaqeem and S. Kwon, "AAD-Net: Advanced End-to-End System for Emotion Detection," Knowl.-Based Syst., vol. 270, p. 110525, 2023.

[24] Ł. Smietanka and T. Maka, "Enhancing Embedded Space with Low-Level Features," Appl. Sci., vol. 15, p. 2598, 2025.

[25] L. Zhang, J. Wang, Y. Zhao, et al., "RELUEM: Reinforcing Emotional Features Using Deep Reinforcement Learning for SER," Cognitive Computation, vol. 17, no. 110, 2025.

[26] A. Tripathi and P. Rani, "Multilingual Speech Emotion Recognition Using IGRFXG – Ensemble Feature Selection Approach," Appl. Acoust., vol. 240, p. 110905, 2025.