PeerJ Computer Science

Hybrid-Module Transformer: enhancing speech emotion recognition with HuBERT, LSTM, and ResNet-50

Xindong Huang, Wuhui Lin, Maming Chen and Hua Shi

School of Opto-Electronic and Communication Engineering, Xiamen University of Technology, Xiamen City, China

ABSTRACT

Speech emotion recognition (SER) is a challenging task that involves identifying human emotions from speech. Traditional sequence models like recurrent neural network (RNN) and long short-term memory (LSTM) are limited by vanishing gradients and difficulty in capturing long-range dependencies. This article presents a novel model based on the Hybrid-Module-Transformer, which leverages the capabilities of Transformer modules to extract feature representations effectively, even with limited data. The model combines the strengths of Hidden-Unit BERT (HuBERT), LSTM, and Residual Network (ResNet-50) to achieve superior performance in speech emotion classification tasks. In the model, we utilized Mel-frequency cepstral coefficients (MFCC) and Spectrogram for feature extraction. Then, a HuBERT-LSTM framework is used to perform both speech-to-text recognition and emotion classification. We evaluate the model on two benchmark datasets: Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and Multimodal EmotionLines Dataset (MELD). On the RAVDESS dataset, the model achieves a maximum accuracy of 76% and precision of 78%, while on the more challenging MELD dataset, it attains an accuracy of 72.9% and precision of 72.3%. These results demonstrate the effectiveness and generalizability of our model in both controlled and real-world conversational scenarios, making it a competitive solution for robust speech emotion recognition.

Submitted 8 April 2025
Accepted 22 September 2025
Published 27 October 2025

Corresponding author
Wuhui Lin,
2422121014@stu.xmut.edu.cn

Academic editor
Othman Soufan

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj-cs.3292**

Subjects Algorithms and Analysis of Algorithms, Artificial Intelligence, Data Mining and Machine Learning, Natural Language and Speech, Neural Networks
Keywords Speech emotion recognition, Deep learning, HuBERT, NLP

INTRODUCTION

Speech emotion recognition (SER) is like a puzzle where we try to figure out how someone feels by listening to their voice. It is not easy because people show their feelings in many ways when they talk. Thus, researchers have proposed many methods to deal with this challenge.

Traditional methods involve hidden Markov models (HMMs), which are suited for time-series data and widely used in the fields, including speech recognition, natural language processing, and bioinformatics (*Mao et al., 2019*; *Feng & Narayanan, 2024*; *Dwivedi et al., 2022*; *Akbulut, Perros & Shahzad, 2020*). *Mao et al. (2019)* focus on the development and evaluation of HMM-based architectures for utterance-level speech emotion recognition. They propose three HMM-based architectures for speech emotion

recognition: Gaussian Mixture Model–Hidden Markov Model (GMM-HMM), Subspace Gaussian Mixture Model–Hidden Markov Model (SGMM-HMM), and Deep Neural Network–Hidden Markov Model (DNN-HMM). In the experimental evaluation, various HMM-based architectures were assessed on the CASIA *corpus*, achieving a maximum accuracy of 86.88%. However, HMMs rely heavily on hand-crafted features, which require expert knowledge and may not capture all the details of the data. Besides this, they cannot capture intricate patterns and relationships in the data.

Researchers have proposed deep neural networks (DNNs) as a means to learn complex patterns in data. These networks are effective for both feature extraction and classification (*Trinh Van et al., 2022*; *Wani et al., 2021*; *Fahad et al., 2021*). Compared to HMMs, DNNs are better at recognizing intricate patterns and relationships within the data, making them particularly effective for SER tasks. For example, in the work of *Fahad et al. (2021)* they combine the excitation source features and the Mel-frequency cepstral coefficients (MFCC) features to develop an emotion recognition system called DNN-HMM. The experimental results show that the epoch feature set is complementary to the MFCC feature set for emotion classification. The average emotion recognition rate of the proposed model using epoch features is 54.52% (*Fahad et al., 2021*). DNNs offer a robust and adaptable framework for modeling complex data, but they fail to process information from previous inputs, which is essential for time series prediction and language modeling.

Recurrent neural networks (RNNs) are more effective than traditional DNNs when dealing with sequential data because they can retain information over time. This capability enables RNNs to process inputs of varying lengths and to take into account the context and dependencies within a sequence. As a result, RNNs are well-suited for tasks such as language modeling, speech recognition, and time series prediction (*Yadav et al., 2022*; *Trinh Van et al., 2022*; *Kons et al., 2022*; *Jermsittiparsert et al., 2020*). For example, *Trinh Van et al. (2022)* present the results of speech emotion recognition with the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) *corpus*. In their case, three deep neural network models, convolutional neural network (CNN), convolutional recurrent neural network (CRNN), and gated recurrent unit (GRU), were used for emotion recognition. The results show that the proposed model gave the highest average recognition accuracy of 97.47% (*Trinh Van et al., 2022*). However, RNNs have difficulty learning long-range dependencies due to two main issues: the vanishing gradient problem, where gradients become too small to effectively update weights, and the exploding gradient problem, where gradients grow excessively large and destabilize the network.

Research shows that using self-supervised training on audio data enables full utilization of large volumes of unlabeled audio for learning latent features. This reduces the need for data labeling and associated costs (*Zhao & Zhang, 2022*; *Chang et al., 2021*; *Chiu et al., 2022*). Wav2Vec2, a transformer-based model trained with self-supervised methods, has offered several advantages over RNN-based models in SER tasks (*Yi et al., 2020*; *Jain et al., 2023*; *Shahgir, Sayeed & Zaman, 2022*). For example, *Yi et al. (2020)* applied the pre-trained wav2vec2.0 model to address the low-resource SER task. The model achieved over 20% relative improvements in six languages compared to previous work (*Yi et al., 2020*). Wav2Vec2 offers a combination of speed, efficiency, and accuracy that makes it a

preferred choice for modern SER systems. While Wav2Vec2 is a powerful model for SER, it does have limitations, such as the need for massive computational resources and reliance on the availability of unlabeled data. These limitations highlight the need for further research and improvements to the SER model.

In recent years, Transformers have become a trend in SER tasks (*Hazmoune & Bougamouza, 2024*; *Andayani et al., 2022*; *Wagner et al., 2023*; *Triantafyllopoulos et al., 2022*). The Transformer architecture, introduced by *Vaswani et al. (2017)*, represents a significant advancement in the field of sequence modeling, offering several advantages over traditional RNNs, DNNs, and HMMs. Unlike RNNs, which process data sequentially, Transformers are effective in capturing long-range dependencies. They are not constrained by the limitations of sequential processing, which allows for more efficient parallel computation and a significant reduction in training times. This advantage is beneficial for utilizing modern hardware capabilities, such as GPUs. *Roy et al. (2021)* propose the Routing Transformer, which utilizes content-based sparse attention inspired by non-negative matrix factorization. Unlike local attention models, the Routing Transformer does not rely on fixed attention patterns while maintaining a similar space-time complexity. Additionally, compared to previous approaches to content-based sparse attention, it does not require the computation of a full attention matrix. Instead, it selects sparsity patterns based on content similarity (*Roy et al., 2021*).

Recent advances have explored dynamic-scope Transformer architectures that allow models to adapt their attention span based on input characteristics (*Chen et al., 2023a*; *Wang et al., 2023*, *2024*; *Chen et al., 2023b*). Examples include the Deformable Speech Transformer (DST) (*Chen et al., 2023b*), which learns token-specific window sizes; and the Time-Frequency Transformer (*Wang et al., 2023*), which models temporal and spectral attention separately before fusing them. These models share a core philosophy: letting the data guide the scale and focus of attention, rather than relying on fixed windows. While they improve performance on benchmarks such as IEMOCAP and Multimodal EmotionLines Dataset (MELD), they also introduce higher model complexity. For instance, DST achieves up to 71.8% weighted accuracy (WA) and 73.6% unweighted accuracy (UA) on IEMOCAP, but its dynamic modules require careful regularization to avoid overfitting on smaller datasets, highlighting a trade-off between flexibility and efficiency.

Capsule networks (CapsNets) have also emerged as a promising direction in SER due to their ability to preserve hierarchical relationships between features. Unlike traditional CNNs that lose spatial relationships during pooling, capsule networks use dynamic routing to retain part-whole relationships, making them particularly effective for capturing complex emotional cues in speech. Recent studies have explored their potential in SER tasks (*Zhang et al., 2024*, *2025*). For example, *Zhang et al. (2024)* proposed a capsule-enhanced neural network (CENN) that integrates multi-head attention, residual blocks, and capsule layers. Their model achieved 72.88% accuracy on the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) dataset, showcasing the advantages of capsule-based architectures in capturing fine-grained emotion-related features (*Zhang et al., 2024*). However, CENN exhibits high computational complexity and

performs less effectively on smaller datasets like Surrey Audio-Visual Expressed Emotion Database (SAVEE), highlighting the trade-off between representational power and model efficiency.

Inspired by these works, this article presents an idea to enhance SER models through the integration of various techniques, including HuBERT, LSTM, and ResNet (*Yi et al., 2020*; *Jain et al., 2023*; *Hattori & Tamura, 2023*; *Li, 2021*). Traditional approaches often struggle with either extracting meaningful features from raw audio, modeling long-range temporal dependencies, or recognizing hierarchical spatial patterns. The proposed model uses HuBERT's self-supervised learning to extract richer and more discriminative audio representations without requiring large labeled datasets, leverages the sequential processing capabilities of LSTM to capture long-range dependencies, and incorporates convolutional layers inspired by ResNet-50 for effective pattern recognition. This hybrid design is motivated by the need to unify strengths across domains—self-supervised learning, temporal modeling, and spatial feature extraction—to form a more robust SER framework. This combination leads to a system that is more accurate and flexible for SER tasks. Additionally, it is capable of generalizing with limited labeled data, enhancing the overall performance and robustness of speech recognition tasks. The proposed method has been evaluated on two widely-used benchmark datasets for speech emotion recognition: Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and MELD. On RAVDESS, which features high-quality emotional speech samples in a controlled environment, the model achieves a maximum accuracy of 76% and a precision of 78%. On the more challenging MELD dataset, which contains multi-party conversations with rich emotional diversity, the model attains an accuracy of 72.9% and a precision of 72.3%. Compared to existing approaches, including AlexNet, ResNet-50, and ResNet-101, our Hybrid-Module-Transformer model demonstrates superior performance across both datasets.

The main contributions of our article are as follows:

- We have developed an integrated Transformer model, which is designed for SER tasks. The hybrid model leverages the strengths of HuBERT for robust feature extraction and LSTM for capturing the temporal dynamics of emotions, leading to enhanced accuracy and contextual understanding. This integrated approach allows for better generalization across different speakers and emotional states. Compared to Wav2Vec2, HuBERT focuses on learning high-level representations of unmasked inputs to accurately infer the targets of masked ones, leading to an improvement in low-resource scenarios where Wav2Vec2 may struggle.
- The HuBERT module serves as the feature extractor, leveraging its Transformer architecture to extract robust features from raw speech data. HuBERT is trained on a large amount of unlabeled audio data using a self-supervised learning approach, allowing it to learn complex patterns and representations of speech.
- The LSTM module effectively extracts features while managing temporal dynamics, offering considerable advantages for processing speech data. This approach not only enhances the precision of emotion recognition across a spectrum of emotional states but

Table 1 Comparative assessment of different speech emotion recognition models.

Model	Features	Advantages	Disadvantages
SGMM-HMM	MFCC + pitch + voicing (47-dim)	Compact; reduces overfitting	Weak in nonlinear representation
DNN-HMM	MFCC + epoch-based features (69-dim)	Effective speaker adaptation	Requires extensive training
GRU	Mel-spectrogram (153-dim)	Captures long-term dependencies	Gradient instability limits long-range learning
DST	Spectrogram + deformable attention	Learns flexible attention spans	Risk of overfitting on small datasets
CENN	MFCC + attention + capsule features	Strong spatial modeling; high accuracy	Performance drops on small datasets
Our model	MFCC + Spectrogram + HuBERT + LSTM + ResNet-50	Combines HuBERT, LSTM, and ResNet-50	High computational complexity and training time

also improves classification accuracy. By utilizing LSTM's capability to capture speech patterns and understand the sequential flow of emotions over time, this model delivers reliable and detailed results for SER tasks.

- ResNet-50 is traditionally used for image recognition, but we applied its convolutional layers for feature extraction in audio processing. This approach allows the model to learn complex hierarchical features from audio data, capturing both low-level and high-level characteristics. This is especially beneficial in speech classification tasks, where the frequency content of speech signals is important for identifying different languages or emotional states.
- The hybrid Transformer model demonstrates exceptional performance in SER tasks, surpassing traditional models such as LSTM and ResNets. Experiments conducted on the RAVDESS Emotional Speech Audio dataset achieved an accuracy rate of 78%.

To demonstrate the advancements brought by our proposed approach, we compare its performance with several representative SER models reported in the literature. The baseline models include traditional architectures such as SGMM-HMM and DNN-HMM, as well as deep learning-based models like gated recurrent unit (GRU). In addition, we include two recent and competitive models: the Deformable Speech Transformer (DST), which represents the adaptive-attention Transformer family, and the CENN, which incorporates capsule structures for improved spatial relationships. While these models have shown impressive results on the IEMOCAP and Berlin Emotional Speech Database (EMODB) datasets, our Hybrid-Module-Transformer model achieves competitive performance on both RAVDESS and MELD, demonstrating strong generalization across diverse SER scenarios. Table 1 presents a quantitative comparison of these models, highlighting their features, advantages, and disadvantages. This comparison further motivates the hybrid design of our model, which balances the strengths of HuBERT, LSTM, and ResNet-50 to enhance speech emotion recognition.

This article is organized as follows: We first introduce our proposed model in 'Methodology'. In 'Implementation and Experiments', we evaluate the proposed model using several evaluation metrics. After assessing the model's performance, we summarize our findings and discuss future work in 'Materials and Methods'.

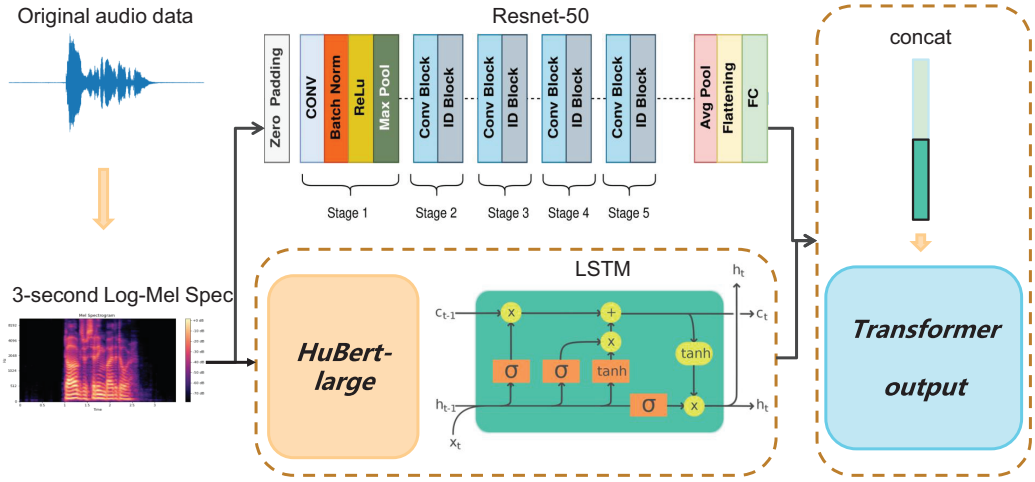

Figure 1 **The system architecture of the proposed model.** Full-size DOI: 10.7717/peerj-cs.3292/fig-1

METHODOLOGY

In this study, we propose a hybrid approach for feature extraction in SER by leveraging the complementary strengths of HuBERT-LSTM and ResNet-50. The process of speech emotion recognition in our model is shown in Fig. 1.

First, the original raw audio data is transformed into a 3-s Log-Mel Spectrograms. This transformation converts the audio signal into a format that is more manageable for our model. After the feature extraction, the Log-Mel Spectrograms is fed separately into the HuBERT-LSTM and ResNet-50 modules. The output from the ResNet-50 module consists of learned features that capture various aspects of the audio data. Next, the HuBERT model is employed to capture both the content and context of the audio. These features are then input into an LSTM network, which models the sequential dependencies in the audio data. Finally, all the feature sets are concatenated and input into a Transformer module for SER classification. By integrating these components, our model leverages the strengths of each module, enhancing overall system performance and leading to more accurate and robust emotion classification results.

ResNet-50 module

In order to better extract rich, hierarchical features from the Log-Mel Spectrograms, we have adopted ResNet-50 for audio processing. As shown in Fig. 2, the adapted ResNet-50 module consists of the following steps:

- **Audio to spectrograms:** Since ResNet-50 was originally designed for image recognition tasks, we convert raw audio signals into Log-Mel Spectrograms, which are 2D representations of the audio's frequency content over time. This transformation makes audio data compatible with ResNet-50.
- **Convolutional layers for feature extraction:** The model utilizes the initial layers of ResNet-50 to extract low-level features from the spectrograms. These layers are effective at capturing local patterns and textures within the spectrograms. Each convolutional

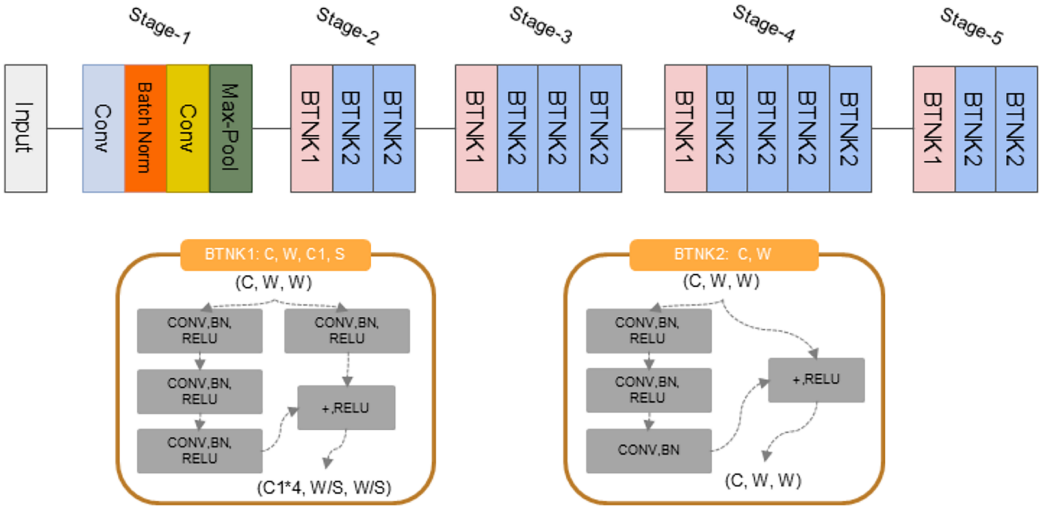

Figure 2 **The architecture of the ResNet-50 module.** Full-size DOI: 10.7717/peerj-cs.3292/fig-2

layer applies a set of filters to the input to extract features. The output of a convolution operation can be represented by the Eq. (1). Where $\boldsymbol{F}$ is the feature map, $\boldsymbol{I}$ is the input audio (Log-Mel Spectrograms), $\boldsymbol{K}$ is the kernel (filter), and * denotes the convolution operation.

$$F(x,y) = (I * K)(x,y) = \sum_m \sum_n I(m,n) \cdot K(x-m, y-n). \tag{1}$$

- **Batch normalization:** After convolutional layers, we apply batch normalization to normalize the activations. The normalized output is described by Eq. (2). In the equation, $\boldsymbol{z}$ represents the input to the batch normalization layer, μ is the mean of $\boldsymbol{z}$, σ^2 is the variance of $\boldsymbol{z}$, and ε is a small constant helps to prevent division by zero. After convolutions, Rectified Linear Unit (ReLU) activation functions are used to learn complex patterns in the audio data that indicate different emotions.

$$\hat{z} = \frac{z-\mu}{\sqrt{\sigma^2 + \varepsilon}}. \tag{2}$$

- **Max pooling:** A max pooling layer is employed to reduce the spatial dimensions of the feature maps, emphasizing the most significant elements. We use a 3×3 pooling window that moves across the feature map with a stride of 2. The window captures the highest value within its area, discarding the rest, and this process is repeated across the map. Max pooling reduces the network's parameter count and computational load, thereby enhancing efficiency.

HuBERT-LSTM module

The HuBERT-LSTM architecture for SER tasks combines the strengths of self-supervised learning and sequence modeling to recognize emotions from speech signals effectively. Detailed workflow and architecture of the HuBERT-LSTM are shown in Fig. 3.

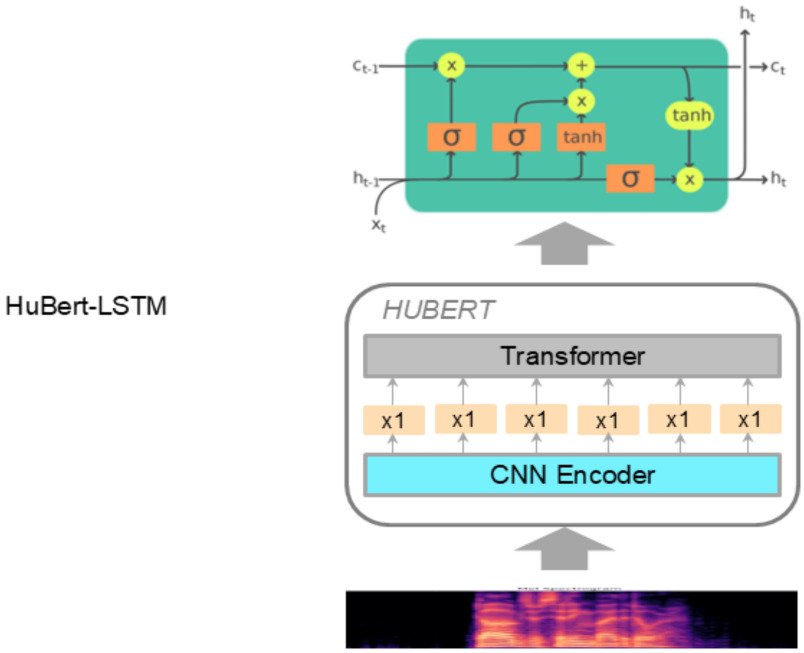

Figure 3 **The architecture of the HuBERT-LSTM module.**
Full-size DOI: 10.7717/peerj-cs.3292/fig-3

First, the raw audio data is transformed into Log-Mel Spectrograms and then fed into the HuBERT model. This model has been trained in a self-supervised manner, allowing it to learn robust speech representations without requiring labeled data. The HuBERT model outputs a sequence of embeddings that capture the acoustic features of the speech signal.

Next, these embeddings are processed by an LSTM network, which is effective at capturing long-term dependencies within sequence data due to its gating mechanisms. The workflow within an LSTM unit can be described by the following steps:

– ***Forget gate:*** The forget gate determines how much of the previous cell state should be retained. As shown in Eq. (3), σ represents the sigmoid function, W_f are the weights associated with the forget gate, h_{t-1} is the hidden state from the previous time step, x_t is the input at time step t, and b_f are the biases.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right). \tag{3}$$

– ***Input gate:*** The input gate controls how much of the new input will be added to the cell state, while the candidate cell state represents the new information to be stored. This is expressed in Eq. (4). Here, i_t denotes the input gate activation, $\tilde{c}$ is the candidate value for the cell state, and W_i, W_C, b_i, and b_C are the weights and biases for the input gate and candidate cell state, respectively.

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C). \end{aligned} \tag{4}$$

After the LSTM processes the sequence, the output is passed through a fully connected layer and a softmax layer to classify the input audio into different emotional categories. The softmax function is given by Eq. (5). Where z_k is the output of the fully connected layer for class κ, and $\boldsymbol{K}$ is the total number of emotion classes.

Finally, the outputs from HuBERT-LSTM and ResNet-50 are fed into a Transformer model for SER outputs. This combines the sequential information captured by LSTM with the spatial features extracted by ResNet-50. This combined architecture leverages HuBERT's ability to extract meaningful acoustic features and LSTM's capability to model the temporal dynamics of speech, providing a practical framework for recognizing emotions in speech signals.

$$P(y = k|x) = \frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)}. \tag{5}$$

IMPLEMENTATION AND EXPERIMENTS

In this experiment, we conducted a comparative analysis and ablation studies with traditional SER models, including AlexNet, ResNet50, ResNet101, LSTM, and LSTM-HuBERT. Our study utilized the RAVDESS Emotional Speech Audio Dataset, which is well-known for its comprehensive emotional annotations and diverse vocal expressions, serving as the foundation for our experiments.

The comparative experiments aimed to benchmark the performance of our proposed model against several commonly used SER models. We evaluated the models using standard metrics, including accuracy, precision, recall, and the F1-score. These metrics allowed us to perform a detailed assessment of each model's efficiency.

Additionally, we conducted an ablation study to assess the impact of various model components. By selectively disabling or modifying certain features, we were able to isolate the contributions of individual elements to the overall performance of the model. This process was instrumental in identifying which aspects of our model were most influential in achieving good performance in emotion recognition tasks.

Dataset

To evaluate the effectiveness and generalizability of the proposed model, we conduct experiments on two widely used emotional speech datasets: the RAVDESS Emotional Speech Audio Dataset and the MELD. These datasets differ in modality, speaker diversity, and conversational structure, providing complementary perspectives for benchmarking SER performance.

RAVDESS

The RAVDESS Emotional Speech Audio Dataset (*Livingstone & Russo, 2019*) consists of 1,440 audio files, generated from 60 trials per actor across 24 professional actors (12 female and 12 male). Each actor vocalizes two lexically matched statements in a neutral North American accent. The dataset covers eight emotion categories: calm, happy, sad, angry,

Table 2 Number of samples per emotion in the RAVDESS emotional speech audio dataset.

ID	Emotion	Size
1	Disgust	192
2	Happy	192
3	Fear	192
4	Sad	192
5	Angry	192
6	Neutral	96

fearful, surprise, disgust, and neutral. Except for the neutral class, all emotions are expressed at two levels of intensity (normal and strong). Each audio file is uniquely identified by a seven-part numerical filename (*e.g.*, 03-01-06-01-02-01-12.wav). Table 2 shows the distribution of audio samples across selected emotion categories used in our experiments.

MELD

The MELD dataset, or the Multimodal EmotionLines Dataset (*Poria et al., 2018*), is a comprehensive benchmark for emotion recognition in conversations. It is an extension of the EmotionLines dataset, enhanced with audio and visual modalities in addition to text. MELD comprises over 1,400 dialogues and 13,000 utterances from the popular TV series "Friends", capturing a wide range of emotional expressions. Each utterance is meticulously labeled with one of seven distinct emotions: anger, disgust, sadness, joy, neutral, surprise, and fear. Moreover, MELD also includes sentiment annotations (positive, negative, and neutral) for each utterance. The dataset's multi-party nature, with dialogues involving multiple speakers, adds to its complexity and makes it a challenging benchmark for evaluating emotion recognition models. The distribution of samples across emotion categories is shown in Table 3, highlighting the class imbalance that must be addressed in model development. This rich multimodal dataset provides a robust foundation for developing and testing advanced emotion recognition systems.

Experiment environment

The raw audio signals used in this study are sampled at a rate of 16 kHz. Each audio utterance is split into several segments. If a segment is less than 3 s long, we add zero padding to ensure it meets the required length. The final prediction for an audio utterance depends on all segments derived from that utterance.

In the experiments, both MFCCs and spectrogram features are used to capture diverse aspects of speech. MFCCs are 40-dimensional features derived from Mel frequencies, which are mainly designed to reflect the human auditory system's response to sound.

A series of Hamming windows is used to generate spectrograms. Each window segment lasts for 50 ms and advances with a hop length of 15 ms. These windowed segments, or frames, are then subjected to the Discrete Fourier Transform (DFT), which is applied with a length of 800. The first 200 DFT points from each frame are selected as

Table 3 Number of samples per emotion in the MELD dataset.

ID	Emotion	Sample count
1	Disgust	271
2	Joy	1,743
3	Fear	268
4	Sadness	683
5	Anger	1,109
6	Neutral	4,710
7	Surprise	1,205

input features, resulting in a spectrogram image of dimensions 300 × 300 for each audio segment.

The model is implemented using PyTorch. We use the AdamW Optimizer with a learning rate of 1e−4 and a weight decay of 5e−4. The training batch size is set to 8. To improve performance when the loss shows no improvement, we adjust the learning rate by a factor of *n*. Additionally, we have set the system to run for a maximum of 30 epochs. The complete processing steps and code can be found in the GitHub Repository.

During the feature extraction stage, two types of features are synchronously derived from each 3-s audio segment, corresponding to the dual-branch structure of the proposed model. First, a 300 × 300 Log-Mel spectrogram is generated using a window size of 50 ms, hop length of 15 ms, and FFT size of 800. The resulting single-channel image is duplicated across three channels and fed into the ResNet-50 network to capture localized time-frequency patterns. Second, a 40-dimensional MFCC feature sequence with 300 frames (aligned with the spectrogram frame count) is extracted and passed into a HuBERT-LSTM branch to model long-range temporal dependencies. The ResNet-50 branch outputs a 2,048-dimensional vector after global average pooling, while the HuBERT-LSTM branch produces a 768-dimensional hidden state. These two representations are concatenated into a 2,816-dimensional feature vector, which is then input into a four-layer Transformer encoder for high-level feature modeling and final emotion classification. This dual-branch structure effectively integrates both spatial and temporal characteristics of the audio signal, enhancing the model's ability to discriminate speech emotions.

Comparative study

To comprehensively assess the effectiveness of our model, we perform comparative experiments on two datasets: RAVDESS and MELD. The following subsections present the detailed results and analysis on each dataset respectively.

Experiments on RAVDESS

Table 4 provides a comparative analysis of the performance of various models in SER. According to the results, "Our model" demonstrates the highest Recall with 0.78, indicating they successfully recognize 78% of the actual emotional instances. This performance surpasses that of LSTM-HuBERT, ResNet-101, ResNet-50, and AlexNet,

Table 4 Evaluation results of various models in SER tasks.

Model	Precision	Recall	Accuracy	F1-score
Alexnet	0.55	0.54	0.54	0.55
ResNet-50	0.59	0.59	0.59	0.59
ResNet-101	0.64	0.67	0.65	0.65
LSTM-HuBERT	0.75	0.77	0.75	0.76
Our model	**0.78**	**0.78**	**0.76**	**0.78**

which have Recall scores of 0.77, 0.67, 0.59, and 0.54, respectively. This indicates that "Our model" is more effective at recognizing emotional states in audio data. In contrast, AlexNet is the least effective, detecting nearly half of the actual emotional instances.

When evaluating the accuracy, which reflects the overall correctness of a model's classifications, both LSTM-HuBERT and our model demonstrate high reliability with scores of 0.75 and 0.76, respectively. These results indicate the high accuracy of our model.

Precision also shows promising results. LSTM-HuBERT and our model lead with scores of 0.75 and 0.77, suggesting a lower rate of false positives compared to other models.

The F1-score, which balances precision and recall, further highlights the performance of LSTM-HuBERT and our model, which achieve the highest scores at 0.76 and 0.78, respectively. This indicates that our model provides a good trade-off between precision and recall, offering a more comprehensive measure of performance.

Overall, the proposed model performs competitively with other state-of-the-art models, as evidenced by its strong recall and F1-score. The high precision and accuracy further demonstrate the model's effectiveness in SER tasks. These results can be attributed to its integration of HuBERT-LSTM's ability to capture long-range dependencies in speech data with ResNet-50's strength in feature extraction, leading to a robust hybrid architecture that effectively recognizes emotional states.

Experiments on MELD

To assess the model's performance in more natural and conversational scenarios, we conduct additional experiments on the MELD dataset. Unlike RAVDESS, MELD contains multi-party dialogues and multimodal emotion expressions, posing greater challenges for recognition models. As shown in Table 5, our Hybrid-Module-Transformer model achieves superior results compared to several strong baselines, including AlexNet, ResNet-50, ResNet-101, and LSTM-HuBERT.

Our model achieved a precision of 0.723, a recall of 0.718, an accuracy of 0.729, and an F1-score of 0.720. In contrast, AlexNet recorded a precision of just 0.482, a recall of 0.469, an accuracy of 0.501, and an F1-score of 0.475. This clearly shows that our model surpasses AlexNet in all metrics, with improvements exceeding 0.20 in both precision and recall.

ResNet-50 fared better than AlexNet, achieving a precision of 0.615, a recall of 0.602, an accuracy of 0.628, and an F1-score of 0.608. However, our model still outshines it by over 0.10 in precision and recall.

Table 5 Comparison of different models on the MELD dataset.

Model	Precision	Recall	Accuracy	F1-score
AlexNet	0.482	0.469	0.501	0.475
ResNet-50	0.615	0.602	0.628	0.608
ResNet-101	0.630	0.621	0.645	0.625
LSTM-HuBERT	0.684	0.679	0.692	0.681
emotion2vec	–	–	0.5188 (WA)	0.487 (WF1)
1D-CNN with feature fusion	0.932	0.938	0.940	0.935
Our model	**0.723**	**0.718**	**0.729**	**0.720**

ResNet-101 further enhances these metrics, with a precision of 0.630, a recall of 0.621, an accuracy of 0.645, and an F1-score of 0.625. Even with these advancements, our model maintains a significant edge over ResNet-101, outperforming it by nearly 0.10 in both Precision and Recall.

LSTM-HuBERT achieved a precision of 0.684, a recall of 0.679, an accuracy of 0.692, and an F1-score of 0.681. While it shows commendable performance, our model distinctly surpasses it by approximately 0.04 in precision and recall.

Beyond these baselines, two recent works provide valuable points of comparison. The first is emotion2vec (*Ma et al., 2023*), which employs large-scale self-supervised pre-training with utterance-level and frame-level objectives. On MELD, emotion2vec reported a weighted accuracy (WA) of 51.88% and a weighted F1(WF1) score of 48.7%. Although these values are below our 72.9% accuracy, emotion2vec demonstrates impressive cross-lingual and cross-task generalization, highlighting the strength of universal pre-trained representations. Our model, in contrast, is specifically designed for speech emotion recognition and therefore achieves higher performance in this focused task.

The second is the CNN-based approach of *Waleed & Shaker (2025)*, which integrates MFCCs, Mel-spectrograms, and Chroma features through a 1D-CNN with feature fusion. Their method achieved an outstanding 94.0% accuracy on MELD, far surpassing our result. The difference lies mainly in methodology: their network directly exploits complementary handcrafted spectral features with an efficient CNN pipeline, while our approach relies on HuBERT for representation learning combined with LSTM and ResNet modules to capture temporal and hierarchical structure. Although our Transformer-based design provides a balanced framework with strong generalization, their feature-fusion strategy demonstrates that careful exploitation of complementary acoustic cues can significantly boost accuracy. This suggests a promising direction for future work, where our architecture may be further enhanced by incorporating similar multi-feature fusion strategies.

In summary, the Hybrid-Module-Transformer model demonstrates exceptional performance on the MELD dataset, outperforming multiple strong baselines and showing competitive accuracy against recent specialized models. While CNN-based fusion models

Table 6 Ablation study results for our model.

Model	Precision	Recall	Accuracy	F1-score
Full model	0.7727	0.7852	0.793	0.789
w/o transformer attention	0.7689	0.7714	0.7929	0.7820
w/o ResNet-50	0.7538	0.7518	0.7819	0.7665
w/o transformer and ResNet-50	0.7538	0.756	0.7741	0.7649

currently achieve higher accuracy, our results establish a strong balance between robustness, generalization, and architectural flexibility.

Ablation study

First, we examine the impact of different modules in our model. As shown in Table 6, the "Full Model" serves as the baseline, demonstrating strong performance with a precision of 0.7727, Recall of 0.7852, accuracy of 0.793, and an F1-score of 0.789. These results indicate a well-balanced trade-off between precision and recall.

The ablation study reveals the significance of each component for the model's performance. When the Transformer Attention mechanism is removed ("w/o Transformer Attention"), there is a slight decrease across all metrics, with precision dropping to 0.7689 and recall to 0.7714. This suggests that attention plays a crucial role in capturing relevant features for emotion recognition. When ResNet-50 is excluded ("w/o ResNet-50"), we observe a drop in precision (0.7538) and accuracy (0.7819), highlighting its importance in stabilizing the learning process and enhancing generalization. Additionally, omitting ResNet-50 leads to a marginal decrease in recall (0.7518), indicating its contribution to preventing overfitting.

Removing both the Transformer and ResNet-50 from the model results in a global decline in performance. Precision drops to 0.7538, recall to 0.756, and accuracy to 0.7741. Furthermore, the F1-score decreases to 0.7649. These results show how various components contribute to enhancing the model's overall performance.

In summary, the results show the role of each component in the model's architecture and their contribution to the performance of the Full Model. The relatively high performance of the Full Model compared to its ablated versions supports the design of the proposed architecture in achieving effective results in SER.

The impact of removing different modules from the proposed architecture on classification precision is evident in Fig. 4. For instance, when the 'sad' emotion classification is analyzed, there is a drop in accuracy from 0.85 in the full model (d) to 0.8 in the model without the Transformer (c), and further down to 0.7 in the model without ResNet-50 (b).

When both the Transformer and ResNet-50 are omitted (as shown in (a)), there is a substantial decrease in accuracy across multiple emotion classifications. The 'fear' emotion classification, in particular, drops from 0.88 in the full model (d) to 0.86 in the model without the Transformer and ResNet-50 (a).

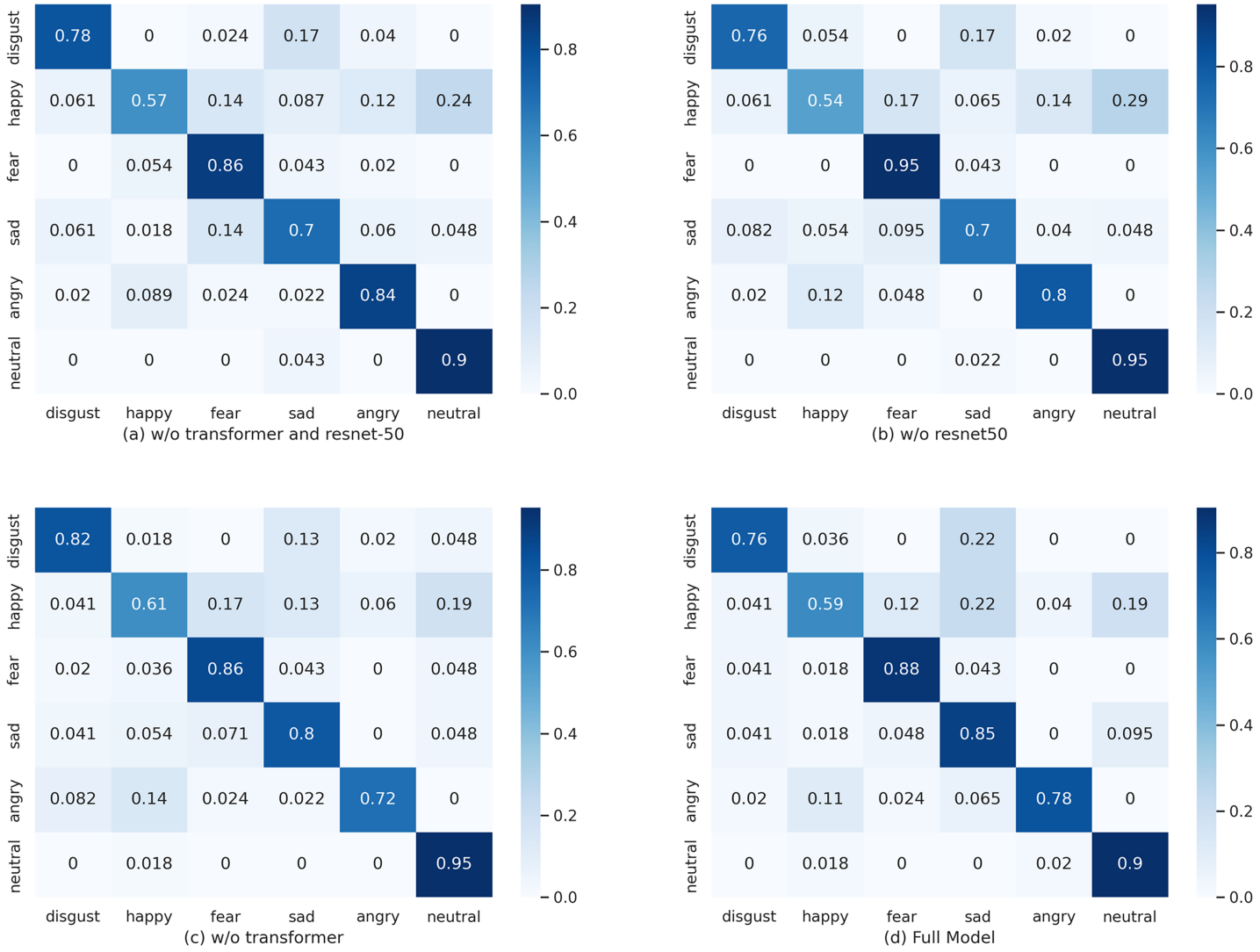

Figure 4 The normalized confusion matrix for the SER without and with the proposed modules. (A) w/o Transformer and ResNet-50: Diagonal (per-class) accuracies—disgust 0.780, happy 0.570, fear 0.860, sad 0.700, angry 0.840, neutral 0.900. Overall accuracy is lowest in this setting: although fear and angry reach relatively high values (0.860 and 0.840), sad is notably low (0.700) and the diagonal is generally lighter than in other configurations, indicating reduced discrimination across emotions. (B) w/o ResNet-50 (Transformer only): Diagonal accuracies—disgust 0.760, happy 0.540, fear 0.950, sad 0.700, angry 0.800, neutral 0.950. The Transformer strongly boosts fear and neutral (both 0.950); however, happy falls to 0.540 (the lowest across settings), suggesting the CNN/ResNet path is important for capturing timbral cues associated with happiness. (C) w/o Transformer (ResNet-50 only): Diagonal accuracies—disgust 0.820, happy 0.610, fear 0.860, sad 0.800, angry 0.720, neutral 0.950. Retaining ResNet-50 raises disgust (0.820) and keeps neutral high (0.950), but without attention the model struggles more with angry (0.720), indicating the Transformer's key role in modeling frame-to-frame dependencies for emotions such as anger. (D) Full model (Transformer + ResNet-50): Diagonal accuracies—disgust 0.760, happy 0.590, fear 0.880, sad 0.850, angry 0.780, neutral 0.900. The combined model yields the most balanced performance overall, with the highest sad accuracy (0.850) and strong fear accuracy (0.880). The diagonal is the darkest and off-diagonal errors are smallest, confirming the complementary benefits of integrating both Transformer and ResNet-50. Full-size DOI: 10.7717/peerj-cs.3292/fig-4

Removing the Transformer (c) also leads to decreased accuracy in classifying the 'sad' and 'angry' emotions, which fall from 0.85 and 0.78 in the full model (d) to 0.8 and 0.72, respectively. The 'fear' classification similarly decreases from 0.88 to 0.86.

The inclusion of both the Transformer and ResNet-50 (as in the full model (d)) results in an improvement in performance, with enhancements in the classification accuracy of 'fear' and 'sad' emotions. The 'sad' classification, in particular, benefits from the full model architecture, with a high accuracy of 0.85 compared to others.

In summary, the confusion matrices and their corresponding accuracy rates illustrate the significant contribution of each component to the model's performance in correctly classifying emotional states. The data underscores the importance of an integrated model architecture that includes both the Transformer and ResNet-50, as seen in the full model (d), to achieve optimal performance in emotion recognition tasks.

MATERIALS AND METHODS

System implementation and usage

The hybrid modular Transformer architecture combining HuBERT, LSTM, and ResNet-50 was implemented as an open-source project with the following implementation details and usage instructions:

Environment setup

All experiments were conducted in a containerized environment using the `pytorch/pytorch:1.13.1-cuda11.6-cudnn8-devel` Docker image to ensure reproducibility. This environment provided PyTorch 1.13.1 with CUDA 11.6 and cuDNN 8 support. Additional dependencies were managed through a requirements file.

Dataset acquisition

The RAVDESS and MELD datasets were employed for training and evaluating our proposed model. The RAVDESS dataset (https://zenodo.org/records/1188976) provides high-quality acted speech recordings with labeled emotional content, suitable for evaluating performance in controlled settings. In contrast, the MELD dataset (https://affective-meld.github.io/) offers multimodal, multi-speaker conversations sourced from television dialogues, making it ideal for assessing model robustness in more natural and dynamic scenarios. Both datasets are publicly available and widely used in speech emotion recognition research.

Feature extraction pipeline

A dedicated preprocessing pipeline was implemented to transform raw audio data into suitable input formats. The `feature_extract.py` script extracts intermediate features from speech samples and stores them in a structured format `(arrays.pkl): python feature_extract.py`

CONCLUSION

This article presents a Hybrid Module Transformer model for speech emotion recognition, demonstrating its effectiveness even with limited data. By combining HuBERT, LSTM, and ResNet-50, our model excels in both feature extraction and emotional classification.

Experiments conducted on the RAVDESS dataset achieved an impressive accuracy rate of 78%, highlighting the robustness of the model in low-data situations. Furthermore, the

model achieved an accuracy of 72.9% on the MELD dataset, demonstrating its robustness in more complex, real-world conversational scenarios involving multimodal emotional expressions. While our Hybrid Module Transformer model demonstrates promising SER results, a major limitation is its dependence on substantial computational resources, which could limit its applicability in low-resource environments. In future work, we plan to improve the model's accuracy by exploring additional features and refining its architecture. We also aim to expand its capabilities to recognize a broader range of emotions and adapt to new languages with minimal retraining.

CODE REPOSITORY URL

The experimental code has been uploaded to GitHub, and the URL of the code is: https://github.com/qpqpaall/ser-main.

THIRD-PARTY DATA

This study utilizes publicly available third-party datasets for training and evaluation. The datasets and their respective access links are listed below:

- RAVDESS: https://zenodo.org/records/1188976.
- MELD: https://github.com/declare-lab/MELD.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the National Key Research and Development Program of China (No. 2022YFF0708500); the Natural Science Foundation of Fujian Province grant number (Nos. 2022J011273, 2022J011275); and the National Natural Science Foundation of China (No. 62372392). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Key Research and Development Program of China: 2022YFF0708500.
Natural Science Foundation of Fujian Province: 2022J011273, 2022J011275.
National Natural Science Foundation of China: 62372392.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Xindong Huang conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

- Wuhui Lin conceived and designed the experiments, performed the experiments, prepared figures and/or tables, and approved the final draft.
- Maming Chen conceived and designed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, and approved the final draft.
- Hua Shi conceived and designed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The code is available in the Supplemental File and at GitHub and Zenodo: https://github.com/qpqpaall/ser-main.

Huang, X., lin, W., Chen, M., & Shi, H. (2025). Hybrid-module transformer: Enhancing speech emotion recognition with HuBERT, LSTM, and ResNet-50. Zenodo. https://doi.org/10.5281/zenodo.17190242.

The RAVDESS dataset is available at Zenodo: Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [Data set]. In PLoS ONE (1.0.0, Vol. 13, Number 5, p. e0196391). Zenodo. https://doi.org/10.5281/zenodo.1188976.

The MELD dataset is available at GitHub and ACL Anthology: https://github.com/declare-lab/MELD.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 527–536, Florence, Italy. Association for Computational Linguistics.

https://doi.org/10.18653/v1/P19-1050.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.3292#supplemental-information.

REFERENCES

Akbulut FP, Perros HG, Shahzad M. 2020. Bimodal affect recognition based on autoregressive hidden markov models from physiological signals. *Computer Methods and Programs in Biomedicine* **195(11)**:105571 DOI 10.1016/j.cmpb.2020.105571.

Andayani F, Theng LB, Tsun MT, Chua C. 2022. Hybrid LSTM-transformer model for emotion recognition from speech audio files. *IEEE Access* **10**:36018–36027 DOI 10.1109/access.2022.3163856.

Chang X, Maekaku T, Guo P, Shi J, Lu Y-J, Subramanian AS, Wang T, Yang S-W, Tsao Y, Lee H-Y, Watanabe S. 2021. An exploration of self-supervised pretrained representations for end-to-end speech recognition. In: *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Piscataway: IEEE, 228–235.

Chen W, Xing X, Xu X, Pang J, Du L. 2023b. DST: deformable speech transformer for emotion recognition. In: *ICASSP, 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE.

Chen S, Xing X, Zhang W, Chen W, Xu X. 2023a. DWFormer: dynamic window transformer for speech emotion recognition. In: *ICASSP, 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Chiu C-C, Qin J, Zhang Y, Yu J, Wu Y. 2022. Self-supervised learning with random-projection quantizer for speech recognition. In: *International Conference on Machine Learning*. PMLR, 3915–3924.

Dwivedi TU, Gupta S, Upadhyay SK, Shukla Y, Ahuja S. 2022. Automatic speech recognition system using hybrid hidden Markov model and human emotion recognition system. *SSRN Electronic Journal* DOI 10.2139/ssrn.4021329.

Fahad MS, Deepak A, Pradhan G, Yadav J. 2021. DNN-HMM-based speaker-adaptive emotion recognition using MFCC and epoch-based features. *Circuits, Systems, and Signal Processing* **40(1)**:466–489 DOI 10.1007/s00034-020-01486-8.

Feng T, Narayanan S. 2024. Foundation model assisted automatic speech emotion recognition: transcribing, annotating, and augmenting. In: *ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE, 12116–12120.

Hattori T, Tamura S. 2023. Speech recognition for minority languages using HuBERT and model adaptation. In: *ICPRAM*, 350–355.

Hazmoune S, Bougamouza F. 2024. Using transformers for multimodal emotion recognition: taxonomies and state of the art review. *Engineering Applications of Artificial Intelligence* **133(3)**:108339 DOI 10.1016/j.engappai.2024.108339.

Jain R, Barcovschi A, Yiwere MY, Bigioi D, Corcoran P, Cucu H. 2023. A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition. *IEEE Access* **11**:46938–46948 DOI 10.1109/access.2023.3275106.

Jermsittiparsert K, Abdurrahman A, Siriattakul P, Sundeeva LA, Hashim W, Rahim R, Maseleno A. 2020. Pattern recognition and features selection for speech emotion recognition model using deep learning. *International Journal of Speech Technology* **23(4)**:799–806 DOI 10.1007/s10772-020-09690-2.

Kons Z, Aronowitz H, Morais E, Damasceno M, Kuo H-K, Thomas S, Saon G. 2022. Extending RNN-T-based speech recognition systems with emotion and language classification. ArXiv DOI 10.48550/arXiv.2207.13965.

Li B. 2021. Hearing loss classification via AlexNet and extreme learning machine. *International Journal of Cognitive Computing in Engineering* **2**:144–153 DOI 10.1016/j.ijcce.2021.09.002.

Livingstone SR, Russo FA. 2019. RAVDESS emotional speech audio. Kaggle. *Available at https://www.kaggle.com/dsv/256618*.

Ma Z, Zheng Z, Ye J, Li J, Gao Z, Zhang S, Chen X. 2023. emotion2vec: self-supervised pre-training for speech emotion representation. ArXiv DOI 10.48550/arXiv.2312.15185.

Mao S, Tao D, Zhang G, Ching P, Lee T. 2019. Revisiting hidden markov models for speech emotion recognition. In: *ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE, 6715–6719.

Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. 2018. MELD: multimodal emotionlines dataset. Affective Computing Group, Carnegie Mellon University. *Available at https://affective-meld.github.io/*.

Roy A, Saffar M, Vaswani A, Grangier D. 2021. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics* **9(3)**:53–68 DOI 10.1162/tacl_a_00353.

Shahgir H, Sayeed KS, Zaman TA. 2022. Applying wav2vec2 for speech recognition on Bengali common voices dataset. ArXiv DOI 10.48550/arXiv.2209.06581.

Triantafyllopoulos A, Wagner J, Wierstorf H, Schmitt M, Reichel U, Eyben F, Burkhardt F, Schuller BW. 2022. Probing speech emotion recognition transformers for linguistic knowledge. ArXiv DOI 10.48550/arXiv.2204.00400.

Trinh Van L, Dao Thi Le T, Le Xuan T, Castelli E. 2022. Emotional speech recognition using deep neural networks. *Sensors* **22(4)**:1414 DOI 10.3390/s22041414.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. 2017. Attention is all you need. In: *Advances in Neural Information Processing Systems.* Vol. 30. Long Beach, CA, USA: Curran Associates, Inc.

Wagner J, Triantafyllopoulos A, Wierstorf H, Schmitt M, Burkhardt F, Eyben F, Schuller BW. 2023. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45(9)**:10745–10759 DOI 10.1109/tpami.2023.3263585.

Waleed GT, Shaker SH. 2025. Speech emotion recognition on MELD and RAVDESS datasets using CNN. *Information* **16(7)**:18 DOI 10.3390/info16070518.

Wang Y, Lu C, Lian H, Zhao Y, Schuller B, Zong Y, Zheng W. 2024. Speech swin-transformer: exploring a hierarchical transformer with shifted windows for speech emotion recognition. In: *ICASSP, 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE.

Wang Y, Lu C, Zong Y, Lian H, Zhao Y, Li S. 2023. Time-frequency transformer: a novel time frequency joint learning method for speech emotion recognition. In: *International Conference on Neural Information Processing (ICONIP)*.

Wani TM, Gunawan TS, Qadri SAA, Kartiwi M, Ambikairajah E. 2021. A comprehensive review of speech emotion recognition systems. *IEEE Access* **9**:47795–47814 DOI 10.1109/access.2021.3068045.

Yadav SP, Zaidi S, Mishra A, Yadav V. 2022. Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN). *Archives of Computational Methods in Engineering* **29(3)**:1753–1770 DOI 10.1007/s11831-021-09647-x.

Yi C, Wang J, Cheng N, Zhou S, Xu B. 2020. Applying wav2vec2. 0 to speech recognition in various low-resource languages. ArXiv DOI 10.48550/arXiv.2012.12121.

Zhang H, Huang H, Zhao P, Yu Z. 2025. Sparse temporal aware capsule network for robust speech emotion recognition. *Engineering Applications of Artificial Intelligence* **144(10)**:110060 DOI 10.1016/j.engappai.2025.110060.

Zhang H, Huang H, Zhao P, Zhu X, Yu Z. 2024. CENN: capsule-enhanced neural network with innovative metrics for robust speech emotion recognition. *Knowledge-Based Systems* **304**:112499 DOI 10.1016/j.knosys.2024.112499.

Zhao J, Zhang W-Q. 2022. Improving automatic speech recognition performance for low-resource languages with self-supervised models. *IEEE Journal of Selected Topics in Signal Processing* **16(6)**:1227–1241 DOI 10.1109/jstsp.2022.3184480.