2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

HMoE-SiMBA: Heterogeneous Mixture-of-Experts with SiMBA Attention for Robust Chinese Speech Emotion Recognition

Lu Wang[1][0009-0003-6738-0132] and Xinyue Duan[2][0009-0008-6297-3031]

[1] China University of Mining and Technology (Beijing), Haidian, Beijing 100083, China
[2] Shanghai University, Baoshan, Shanghai 201900, China
luw31258@gmail.com

Abstract. Speech Emotion Recognition (SER) for Mandarin Chinese is crucial for human-computer interaction, yet faces challenges in real-world applications due to unique tonal and prosodic features. Existing methods suffer from limitations in feature extraction, model generalization, and computational efficiency. To address these issues, we propose **HMoE-SiMBA**, a novel framework based on **H**eterogeneous **Mo**xture-of-**E**xperts and **SiMBA**(**Si**mplified **M**amba-**B**ased **A**rchitecture) [30] attention for addressing stability and generalization issues in Chinese SER. Our approach employs a multi-modal feature representation layer to comprehensively capture emotional cues, utilizes heterogeneous feature extractors with dynamic routing to enhance feature adaptability, and combines EinFFT and Mamba [23] for efficient sequence modeling. Experiments on the CASIA dataset demonstrate that HMoE-SiMBA achieves 92.2% accuracy, significantly outperforming existing methods with robust performance in complex acoustic environments.

Keywords: Chinese speech emotion recognition, State Space Models, Mixture-of-Experts.

1 Introduction

Mandarin Chinese Speech Emotion Recognition (SER) is crucial for advancing human-computer interaction (HCI) systems, with applications spanning intelligent customer service, mental health monitoring, personalized education, and affective computing [1]. Unlike many Western languages, Mandarin Chinese employs unique prosodic and tonal variations to encode emotional semantics, making accurate SER both essential and particularly challenging [2]. Despite progress, the practical application of Chinese SER systems is often limited by their robustness against real-world acoustic complexities, including environmental noise, dialectal variations, and the computational cost of processing long speech segments.

Recent advancements in Chinese SER have increasingly utilized deep learning (DL) architectures to address these challenges. Sequential capsule networks have been explored for modeling the spatio-temporal dynamics of speech spectral features [3].

Transformer-based encoders, pre-trained on extensive multilingual corpora, have demonstrated the ability to capture cross-linguistic emotional nuances, improving generalization [4]. Self-supervised learning (SSL) frameworks, such as HuBERT [5] and Wav2vec [6], have also emerged as powerful tools for learning robust feature representations directly from raw speech waveforms, even with limited labeled data. Furthermore, state-of-the-art methods incorporate multi-task learning [7] and contrastive predictive coding [8] to enhance model generalization and feature discriminability in SER. However, significant obstacles still hinder the development of truly robust and efficient Chinese SER systems. We identify three key limitations:

1.1 Feature Extraction Bottleneck

Traditional handcrafted features, like Mel-Frequency Cepstral Coefficients (MFCCs), and reliance on single-modality representations often fail to fully capture the multifaceted nature of emotional expression in Chinese speech. They may overlook critical emotional cues within harmonic structures, intricate pitch contours, and dynamic spectral variations. This incomplete feature representation limits model performance, especially for nuanced emotion classification.

1.2 Limited Model Generalization:

Conventional monolithic deep learning architectures, such as standalone Transformer networks, often exhibit suboptimal generalization when faced with noisy acoustic environments or variations in speech prosody (e.g., pitch-shifted speech). These models struggle to balance local temporal dynamics with broader global semantic context, reducing their adaptability and robustness across diverse real-world conditions.

1.3 Computational Inefficiencies

Standard attention mechanisms, particularly in Transformer architectures, have quadratic computational complexity ($O(N^2)$) relative to sequence length. This inefficiency makes them computationally prohibitive for long-duration speech sequences common in realistic scenarios. Training such models is also prone to gradient instability, further increasing computational demands and hindering scalability for real-time applications.

To overcome these limitations, we propose **HMoE-SiMBA**, a novel and robust framework for Chinese Speech Emotion Recognition, based on a **H**eterogeneous **M**ixture-**o**f-**E**xperts **(HMoE)** and the **SiMBA** (**Si**mplified **M**amba-**B**ased **A**rchitecture) attention mechanism. To address feature extraction limitations, HMoE-SiMBA uses a Speech Multi-modal Feature Representation Layer. This layer comprehensively represents speech signals by integrating spectral, energy, and harmonic information from MFCCs, Mel-spectrograms, and Chroma features. Data augmentation, including time stretching, pitch shifting, and additive white noise, enhances input feature diversity and robustness. To mitigate model generalization issues, HMoE-SiMBA is designed with a **Heterogeneous Feature Extractor Mixture-of-Experts (HMoE) Network**. This network includes a diverse set of specialized feature extractors, such as Transformer,

LSTM, and CNN modules, each designed to capture distinct aspects of speech from different perspectives. A Multi-Expert Routing Network dynamically weights expert outputs, enabling adaptive and optimized feature representation. Finally, to address computational inefficiencies, HMoE-SiMBA incorporates the **SiMBA Attention** mechanism. SiMBA utilizes an **EinFFT Module** for efficient frequency-domain channel selection, filtering key frequency information. Simultaneously, a **Mamba Block** provides lightweight sequence context modeling, achieving reduced computational complexity of $O(N\log N)$. This design effectively enhances emotion-relevant features through frequency-domain channel selection, offering a computationally efficient and robust solution for Chinese SER in complex environments, while maintaining global semantic understanding.

Our main contributions are:

- We propose **HMoE-SiMBA**, a framework for Chinese Speech Emotion Recognition using Heterogeneous Mixture-of-Experts (HMoE) that includes Transformer, LSTM and CNN modules to extract diverse features and improve robustness.
- We integrate the **SiMBA Attention** Mechanism[30], which synergistically combines EinFFT-based spectral gating and Mamba-driven [23] sequence modeling. This integration enhances emotional feature discrimination through efficient frequency domain filtering and sequence modeling, significantly improving computational efficiency compared to traditional attention mechanisms.
- We design a **Multi-Expert Routing Network** within the HMoE architecture that adaptively weights contributions from different feature experts based on input signal characteristics, demonstrating enhanced robustness in adverse acoustic conditions.

2 Related Work

2.1 Speech Emotion Recognition.

Traditional machine learning methods [2][9][10], such as Support Vector Machines (SVM), Random Forests (RF), and Hidden Markov Models (HMM), rely heavily on manually selected features like MFCCs, prosodic features (e.g., pitch and energy contours), and statistical models for speech emotion classification [1]. While somewhat effective, these approaches are limited by the quality and relevance of handcrafted features, which may fail to capture the complex, high-dimensional nature of speech signals.

Deep learning (DL) techniques [3][11][12][13], including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and Autoencoders, can automatically learn hierarchical feature representations directly from raw or minimally processed speech data. This capability to extract and model intricate temporal and spatial patterns has led to significant improvements. For example, Wu et al. [3] introduced sequential capsule networks that excel in capturing spatial and contextual information, enhancing SER performance.

Self-Supervised Representation Learning (SSRL) is an unsupervised learning paradigm designed to extract rich and meaningful representations directly from raw speech signals [7][14][15][16]. By leveraging the inherent structure in speech data, SSRL methods reduce the need for large labeled datasets and have been successfully applied to SER [7]. Popular SSRL frameworks, such as HuBERT [5] and Wav2vec [6], have shown significant potential in SER, providing robust pre-trained representations that improve downstream model performance, even with limited emotion datasets. For instance, Zhang et al. [4] used a Transformer-based encoder, pre-trained on extensive unlabeled audio from diverse datasets, to learn more general and robust acoustic representations. Li et al. [8] utilized contrastive predictive coding to learn salient representations from unlabeled datasets for SER.

2.2 State Space Models

State Space Models (SSMs) have demonstrated effectiveness in capturing the dynamic s and dependencies of sequences through state space transformations. The structured s tate-space sequence model (S4) [17][18][19] is specifically designed to handle long-ra nge dependencies with linear complexity. Following S4, models like S5 [20], H3 [21], and GSS [22] have been proposed.

Mamba distinguishes itself by incorporating a data-dependent SSM layer and a selection mechanism known as parallel scan (S6) [23]. Compared to Transformer-based models with quadratic attention complexity, Mamba excels at processing long sequences with linear complexity.

In computer vision, SSMs were initially applied to pixel-level image classification, while S4 was used for long-range temporal dependency modeling in movie clip classification.

Mamba's potential has motivated numerous studies, highlighting its superior perfor mance and GPU efficiency over Transformers in visual tasks such as object detection [24] and semantic segmentation [25].

3 Methodology

This section details our proposed HMoE-SiMBA framework for Mandarin Chinese speech emotion recognition. We begin with a formal problem definition, followed by a description of each methodological component: multi-modal feature representation, data augmentation, the Heterogeneous Mixture-of-Experts (HMoE) Network, the Multi-Expert Routing Network, the SiMBA attention mechanism, and model training and optimization.

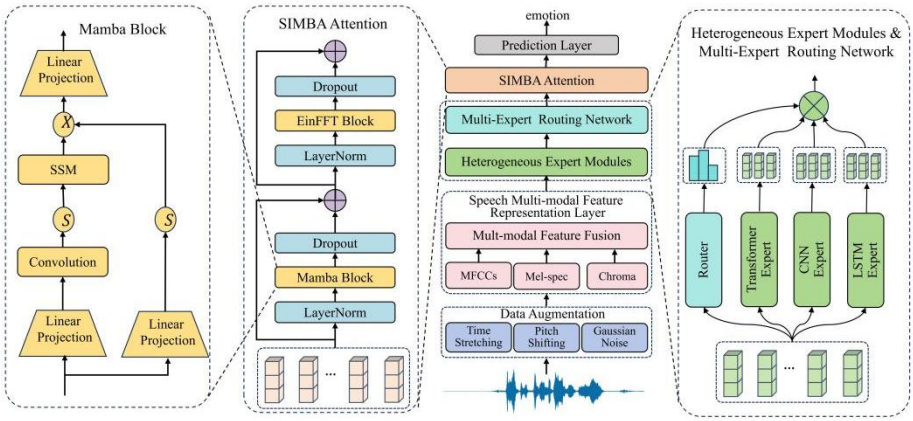

Fig. 1. HMoE-SiMBA Framework Architecture.

3.1 Problem Definition

Given a Mandarin Chinese speech utterance, our goal is to accurately classify its emotional state into a predefined emotion category. Let $\mathcal{X}$ be the space of Mandarin Chinese speech utterances, and $\mathcal{Y} = \{y_1, y_2, \ldots, y_C\}$ be the set of C emotion categories. We aim to learn a mapping function $f: \mathcal{X} \to \mathcal{Y}$ that predicts the emotion category $\hat{y} \in \mathcal{Y}$ for an input utterance $\mathbf{x} \in \mathcal{X}$.

For each utterance $\mathbf{x}$, we extract a set of multi-modal features. Let $\mathbf{F}(\mathbf{x})$ denote the multi-modal feature representation of utterance $\mathbf{x}$, where $\mathbf{F}(\mathbf{x}) = \text{Concat}\left(\mathbf{F}_{mfcc}(\mathbf{x}), \mathbf{F}_{mel}(\mathbf{x}), \mathbf{F}_{chroma}(\mathbf{x})\right)$. Here, $\mathbf{F}_{mfcc}(\mathbf{x})$, $\mathbf{F}_{mel}(\mathbf{x})$, and $\mathbf{F}_{chroma}(\mathbf{x})$ represent the MFCC, Mel-spectrogram, and chroma features extracted from $\mathbf{x}$, respectively, and $\text{Concat}(\cdot)$ denotes the concatenation operation. **HMoE-SiMBA** model, parameterized by $\boldsymbol{\Theta}$, takes these features as input and outputs a probability distribution over emotion categories, $p(y|\mathbf{F}(\mathbf{x}); \boldsymbol{\Theta})$. The predicted emotion $\hat{y}$ is determined by maximizing this probability:

$$\hat{y} = \text{argmax}_{y \in \mathcal{Y}} p(y|\mathbf{F}(\mathbf{x}); \boldsymbol{\Theta}) \qquad (1)$$

The challenge in Mandarin Chinese SER is capturing subtle emotional cues in speech, intricately linked to tonal variations, prosody, and language-specific contextual nuances. Robustness to acoustic variations like noise and dialectal diversity is also crucial for real-world applications. **HMoE-SiMBA** addresses these challenges through the integration of multi-modal feature representations, a heterogeneous mixture of expert feature extractors, and the computationally efficient SiMBA attention mechanism.

3.2 Speech Multi-modal Feature Representation Layer

To comprehensively represent emotional information in Mandarin Chinese speech, we extract multi-modal acoustic features: Mel-Frequency Cepstral Coefficients (MFCCs),

Mel-spectrograms, and chroma features. Each feature provides complementary perspectives on the speech signal and is selected for its established effectiveness in capturing different dimensions of emotional expression.

Feature Extraction.

Mel-Frequency Cepstral Coefficients (MFCCs) ($\mathbf{F}_{mfcc}$*).* MFCCs, a cornerstone of speech processing, effectively represent the spectral envelope, sensitive to phonetic content and emotional undertones. We extract 40-dimensional MFCCs per frame, with a 25 ms frame size and 10ms hop length. To capture temporal dynamics, we include first and second derivatives (delta and delta-delta coefficients), resulting in a $D_{mfcc} = 120$-dimensional MFCC feature vector $\mathbf{f}_{mfcc}(t) \in \mathbb{R}^{120}$ at time frame t. The MFCC feature sequence for utterance x is denoted as $\mathbf{F}_{mfcc}(\mathbf{x}) = \left[\mathbf{f}_{mfcc}(1),\ \mathbf{f}_{mfcc}(2),\ \dots,\ \mathbf{f}_{mfcc}(T)\right] \in \mathbb{R}^{T \times D_{mfcc}}$, where T is the number of frames.

Mel-spectrograms ($\mathbf{F}_{mel}$*):* Mel-spectrograms offer a perceptually relevant representation of the speech spectrum, emphasizing frequency bands critical for human auditory perception and emotion discrimination. We compute Mel-spectrograms using 128 Mel filter banks, with the same frame size and hop length as MFCCs. Each frame t yields a $D_{mel} = 128$-dimensional Mel-spectrogram feature vector $\mathbf{f}_{mel}(t) \in \mathbb{R}^{128}$. The Mel-spectrogram feature sequence is denoted as $\mathbf{F}_{mel}(\mathbf{x}) = [\mathbf{f}_{mel}(1),\ \mathbf{f}_{mel}(2),\ \dots,\ \mathbf{f}_{mel}(T)] \in \mathbb{R}^{T \times D_{mel}}$.

Chroma Features ($\mathbf{F}_{chroma}$*):*. Chroma features, or pitch class profiles, capture the harmonic content of speech, related to emotional expression and prosodic variations. They represent the intensity of the 12 pitch classes of the Western chromatic scale, robust to timbre and octave variations. We extract 12-dimensional chroma features per frame, resulting in $\mathbf{f}_{chroma}(t) \in \mathbb{R}^{12}$ at time t, and $\mathbf{F}_{chroma}(\mathbf{x}) = [\mathbf{f}_{chroma}(1),\ \mathbf{f}_{chroma}(2),\ \dots,\ \mathbf{f}_{chroma}(T)] \in \mathbb{R}^{T \times D_{chroma}}$, where $D_{chroma} = 12$.

Multi-modal Feature Fusion.

To integrate complementary emotional cues, we employ temporal averaging followed by channel concatenation. For each modality $m \in \{mfcc, mel, chroma\}$, we compute the temporal average of $\mathbf{F}_m(\mathbf{x})$ along the time dimension to obtain a fixed-length feature vector $\mathbf{v}_m(\mathbf{x}) \in \mathbb{R}^{D_m}$:

$$\mathbf{v}_m(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{f}_m(t) \qquad (2)$$

We then concatenate these averaged feature vectors across modalities to form the multi-modal input feature vector $\mathbf{F}_{in}(\mathbf{x}) \in \mathbb{R}^{D_{in}}$:

$$\mathbf{F}_{in}(\mathbf{x}) = \text{Concat}\left(\mathbf{v}_{mfcc}(\mathbf{x}), \mathbf{v}_{mel}(\mathbf{x}), \mathbf{v}_{chroma}(\mathbf{x})\right) \qquad (3)$$

where Concat$(\cdot)$ denotes vector concatenation.This fused feature vector $\mathbf{F}_{in}(\mathbf{x})$ serves as the input to our **Heterogeneous Feature Extractor Mixture-of-Experts (HMoE) Network**.

3.3 Data Augmentation Strategies

To enhance robustness and generalization, particularly in noisy real-world conditions, we apply data augmentation during training. These techniques increase the diversity o f the training data by introducing realistic perturbations to speech signals, thereby imp roving the model's ability to generalize to unseen data and noisy environments.

Time Stretching

Time stretching alters the speaking rate without changing the pitch, using the Phase V ocoder algorithm [26] with a stretching factor α_{ts} randomly sampled from a uniform d istribution $\mathcal{U}(0.8,1.2)$. This technique makes the model invariant to variations in spea king tempo observed in spontaneous speech. For an original speech signal $\mathbf{x}$, the time-stretched signal is denoted as $\mathbf{x}_{ts}$, with its duration scaled by α_{ts}.

Pitch Shifting

Pitch shifting [27] modifies the pitch of speech while preserving the speaking rate. We apply pitch shifting with a pitch shift factor α_{ps} randomly selected from the set $\{-2, -1, 0, +1, +2\}$ semitones. This augmentation improves the model's robustness to voc al pitch variations arising from different speaker characteristics and emotional states. The pitch-shifted signal is denoted as $\mathbf{x}_{ps}$.

Additive White Gaussian Noise

To simulate noisy acoustic environments, we add additive white Gaussian noise (AWGN) to speech signals. The noise component $\mathbf{n}$ is sampled from a normal distribution $\mathcal{N}(0,1)$. The noise level is controlled by a scaling factor β which is determined based on the desired signal-to-noise ratio (SNR), randomly sampled from $\mathcal{U}(10,25)$ dB. For a given SNR, β is calculated to achieve the target SNR. The noisy speech signal $\mathbf{x}_{noise}$ is generated as:

$$\mathbf{x}_{noise} = \mathbf{x} + \beta\mathbf{n} \qquad (4)$$

where β is the noise scaling factor corresponding to the sampled SNR.

These augmentations are applied stochastically during training epochs. For each training sample, we randomly apply one or more augmentations with a certain probability, effectively expanding the training dataset and enhancing the model's generalization and robustness.

3.4 Heterogeneous Feature Extractor Mixture-of-Experts Network

The Heterogeneous Feature Extractor Mixture-of-Experts (HMoE) Network is a core component of our framework. It is designed to adaptively extract and integrate diverse temporal dependencies and emotional nuances from the multi-modal input features. The HMoE network comprises a set of specialized expert modules and a Multi-Expert Routing Network.

Overall Architecture

The HMoE network consists of K expert feature extractors, $\{Expert_1, Expert_2, \ldots, Expert_K\}$, and a Multi-Expert Routing Network. In our implementation, these experts are heterogeneous deep learning modules, specifically an LSTM, a CNN, and a Transformer encoder. Each expert is designed to capture distinct aspects of the input speech features. The input $\mathbf{F}_{in}(\mathbf{x})$ is processed by each expert in parallel. The Multi-Expert Routing Network, denoted as G, dynamically aggregates the outputs of these experts based on the input features, enabling an adaptive and optimized feature representation.

Let $\mathbf{H}_k(\mathbf{x}) = Expert_k\big(\mathbf{F}_{in}(\mathbf{x})\big)$ be the output of the k-th expert. The routing network G takes $\mathbf{F}_{in}(\mathbf{x})$ as input and produces routing weights $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_K(\mathbf{x})]$, where $g_k(\mathbf{x})$ represents the weight assigned to the k-th expert. These weights are normalized such that $\sum_{k=1}^{K} g_k(\mathbf{x}) = 1$ and $g_k(\mathbf{x}) \geq 0$. The final output of the HMoE network, $\mathbf{H}_{HMoE}(\mathbf{x})$, is computed as a weighted sum of the expert outputs:

$$\mathbf{H}_{HMoE}(\mathbf{x}) = \sum_{k=1}^{K} g_k(\mathbf{x})\mathbf{H}_k(\mathbf{x}) = \sum_{k=1}^{K} g_k(\mathbf{x}) Expert_k\big(\mathbf{F}_{in}(\mathbf{x})\big) \qquad (5)$$

Heterogeneous Expert Modules.

The heterogeneity in our HMoE network stems from the use of different types of expert modules, each with unique strengths in capturing different patterns in sequential data:

Long Short-Term Memory Network (LSTM): Captures long-range temporal dependencies, modeling the evolution of emotional states over time.

Convolutional Neural Network (CNN): Extracts local patterns and short-duration emotional cues, robust to speech rate variations.

Transformer Encoder: Uses self-attention to capture both local and global dependencies, modeling complex interactions in speech.

By integrating these diverse expert modules, the HMoE network can comprehensively analyze speech signals from multiple perspectives, thereby enhancing the robustness and accuracy of emotional content representation.

3.5 Multi-Expert Routing Network

The Multi-Expert Routing Network, denoted as G, plays a crucial role in dynamically determining the contribution of each expert based on the input features. This allows for a selective and adaptive utilization of expert knowledge, tailored to the characteristics of the input speech.

Gating Network Design

The Multi-Expert Routing Network is implemented as a Multi-Layer Perceptron (MLP). It takes the multi-modal input feature vector $\mathbf{F}_{in}(\mathbf{x})$ as input and outputs a weight vector $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_K(\mathbf{x})]$. The MLP architecture consists of two

hidden layers, each followed by ReLU activation functions, and a final linear layer followed by a softmax activation function. The softmax activation ensures that the output weights are non-negative and sum to 1, effectively representing a probability distribution over the experts.

$$\mathbf{g}(\mathbf{x}) = \text{Softmax}\left(MLP\left(\mathbf{F}_{in}(\mathbf{x})\right)\right) \qquad (6)$$

The input $\mathbf{F}_{in}(\mathbf{x})$ provides a global representation of the utterance. The MLP learns to map these global features to an optimal set of expert weights, effectively deciding the relevance of each expert for processing the given input.

Expert Routing Mechanism

Expert routing is achieved through element-wise multiplication of the gating weights with the corresponding outputs from the LSTM, CNN, and Transformer experts, followed by summation, as described in Equation (1). This soft routing mechanism allows for end-to-end, gradient-based training of the entire **HMoE-SiMBA** framework. The gating network learns to assign higher weights to experts that are more specialized and effective in processing the current input, thus creating a dynamically adaptive ensemble of feature extractors.

3.6 SiMBA Attention Mechanism

To further refine the feature representation and enhance computational efficiency, we integrate the **SiMBA** mechanism. SiMBA is designed to perform frequency-domain channel selection and lightweight sequence context modeling, utilizing an efficient EinFFT module and Mamba blocks.

The SiMBA mechanism is structured as a sequential combination of an **EinFFT Module** and a **Mamba Block**. The input to SiMBA is the output from the HMoE network, $\mathbf{H}_{HMoE}(\mathbf{x})$.

EinFFT Module.

This module is responsible for performing efficient frequency-domain channel selection using Einstein summation. It first transforms the input $\mathbf{H}_{HMoE}(\mathbf{x})$ from the time domain to the frequency domain using Efficient FFT (EinFFT):

$$\mathbf{H}_{freq}(\mathbf{x}) = \text{EinFFT}\left(\mathbf{H}_{HMoE}(\mathbf{x})\right) \qquad (7)$$

Subsequently, it applies Einstein summation for channel selection. Let $\mathbf{W}_{select} \in \mathbb{R}^{D_{freq} \times D_{selected}}$ be a learnable weight matrix, where D_{freq} is the dimension in the frequency domain and $D_{selected}$ is the number of selected channels. The channel selection process is defined as:

$$\mathbf{H}_{selected}(\mathbf{x}) = \text{einsum}\left('btd, dfc -> btc', \mathbf{H}_{freq}(\mathbf{x}), \mathbf{W}_{select}\right) \qquad (8)$$

where einsum denotes Einstein summation. This operation effectively weights different frequency components to select the most salient channels for emotion recognition, focusing on informative spectral features while filtering out redundant or noisy components. Note that the dimension of the time frame is moved to the second position in the output tensor.

Mamba Block.
Following the EinFFT module, a Mamba block, denoted as $Mamba_{SIMBA}$, is employed to model temporal dependencies within the selected frequency channels. This step further refines the contextual representation while maintaining computational efficiency. The output of the SiMBA mechanism, $\mathbf{H}_{SIMBA}(\mathbf{x})$, is given by:

$$\mathbf{H}_{SIMBA}(\mathbf{x}) = Mamba_{SIMBA}\left(\mathbf{H}_{selected}(\mathbf{x})\right) \qquad (9)$$

The SiMBA attention mechanism provides a computationally efficient approach for frequency-domain feature selection and lightweight sequence modeling. It reduces the computational burden compared to traditional attention mechanisms while enhancing the discrimination of emotion-relevant features.

3.7 Model Training and Optimization

The entire **HMoE-SiMBA** model is trained end-to-end using the categorical cross-entropy loss function, which is standard for multi-class classification tasks. Given a training dataset $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$, consisting of N samples $(\mathbf{x}^{(i)}, y^{(i)})$, the loss function is defined as:

$$\mathcal{L}(\boldsymbol{\Theta}) = -\frac{1}{N}\sum_{i=1}^{N} \log p\left(y^{(i)}|\mathbf{F}(\mathbf{x}^{(i)}); \boldsymbol{\Theta}\right) \qquad (10)$$

where $\mathbf{F}(\mathbf{x}^{(i)})$ is the multi-modal feature vector extracted from the i-th utterance $\mathbf{x}^{(i)}$, and $p\left(y^{(i)}|\mathbf{F}(\mathbf{x}^{(i)}); \boldsymbol{\Theta}\right)$ is the predicted probability of the true emotion label $y^{(i)}$. The model parameters $\boldsymbol{\Theta}$ encompass all learnable parameters within the HMoE network, the SiMBA mechanism, and the final classification layer. We optimize $\boldsymbol{\Theta}$ by minimizing $\mathcal{L}(\boldsymbol{\Theta})$ using backpropagation and the Adam optimizer [28], employing learning rate scheduling and weight decay techniques to ensure stable and efficient training.

4 EXPERIMENTS

This section details the experimental framework to evaluate our proposed **HMoE-SiMBA** model using the CASIA dataset, a public real-world dataset. These experiments aim to answer the following research questions (RQ):

RQ1: Does HMoE-SiMBA outperform state-of-the-art (SOTA) models in Mandarin Chinese speech emotion recognition?

RQ2: Are individual modules within HMoE-SiMBA effective and beneficial to overall performance?

RQ3: How sensitive is HMoE-SiMBA to key hyperparameter variations?

RQ4: What is the impact of data augmentation strategies used in HMoE-SiMBA?

4.1 Experimental Setup

Datasets.
We evaluate HMoE-SiMBA on the CASIA dataset, a recognized benchmark for Mandarin Chinese SER. The CASIA dataset contains recordings from **4** speakers, totaling

1200 utterances, annotated with six emotion categories: *Anger*, *Sadness*, *Happiness*, *Neutral*, *Fear*, and *Surprise*. The dataset exhibits a natural emotion distribution, reflecting real-world expressions. Dataset statistics are in Table 1.

Benchmark Methods.
To rigorously assess HMoE-SiMBA, we compare its performance against benchmark methods in two groups: (1) **Traditional Machine Learning Models**: **SVM**, **RF**, and **LR**, representing conventional SER approaches; and (2) **Deep Learning-based Models**: **LSTM**, **Transformer**, **DCNN+LSTM**, **TLFMRF(5-fold)**, **CNN+Bi-GRU**, and **CPAC**. These deep learning baselines represent SOTA architectures, including Transformer, CNN, and LSTM-based models, providing a comprehensive comparison with HMoE-SiMBA.

Evaluation Metrics and Protocol.
We use standard evaluation metrics: **Accuracy**, **Precision**, **Recall**, and **F1-score**. Accuracy measures overall emotion classification correctness. Precision and Recall assess the model's ability to correctly identify each emotion category. F1-score, the harmonic mean of Precision and Recall, provides a balanced performance measure. We use stratified 5-fold cross-validation for robust and unbiased evaluation. The dataset is divided into five folds, preserving class distribution. In each fold, we train on four folds and evaluate on the remaining fold. Reported results are average performance across folds, estimating generalization capability.

Implementation Details.
HMoE-SiMBA is implemented in PyTorch and trained on NVIDIA RTX 4090 GPUs. We use Librosa [29] for feature extraction: Mel-Frequency Cepstral Coefficients ($\mathbf{F}_{mfcc}$), Mel-spectrograms ($\mathbf{F}_{mel}$), and Chroma features ($\mathbf{F}_{chroma}$). The Heterogeneous Feature Extractor Mixture-of-Experts (HMoE) Network has three experts ($K = 3$): CNN, LSTM, and Transformer, each using Mamba blocks with differentiated state-space parameters. The Multi-Expert Routing Network is a two-layer MLP with ReLU activations. The SiMBA Attention Mechanism includes an EinFFT Module and a Mamba Block for frequency-domain channel selection and sequence modeling. We use the AdamW optimizer with an initial learning rate of $1e-4$, batch size of 32, and weight decay of $1e-5$. A learning rate scheduler with a decay factor of 0.9 is applied every 10 epochs. Training is conducted for 100 epochs, and the model with the best validation accuracy is selected for testing.

Table 1. CASIA Dataset Statistics.

Statistic	Value
Number of Speakers	4
Total Utterances	1200
Emotion Categories	Anger, Sadness, Happiness, Neutral, Fear, Surprise
Average Utterance Length	32.5 seconds
Sampling Rate	16 kHz

4.2 Overall Performance (RQ1)

To address RQ1, we compare HMoE-SiMBA's overall performance with benchmark methods on the CASIA dataset. Table 2 shows class-wise accuracy, overall accuracy, precision, recall, and F1-score for HMoE-SiMBA and baselines.

Table 2. Overall Performance Comparison on the CASIA Dataset.

Model	Class-wise Accuracy (%)						Overall Performance (%)			
	Anger	Sadness	Happiness	Neutral	Fear	Surprise	Accuracy	Precision	Recall	F1-score
SVM	43.6	48.7	40.4	52.5	41.3	32.6	44.2	47.2	43.4	45.2
RF	62.1	39.1	50.0	63.3	36.6	43.1	50.0	50.7	48.8	49.8
LR	60.2	52.1	63.3	77.3	58.8	67.9	62.9	65.5	62.9	64.3
LSTM	73.2	76.1	74.5	75.8	72.9	76.7	75.2	74.8	75.0	74.9
Transformer	76.8	79.0	77.3	78.5	76.2	79.5	78.7	78.5	78.8	78.6
DCNN+LSTM	81.7	84.2	82.5	83.8	80.9	85.3	84.1	83.9	84.2	84.0
TLFMRF (5-fold)	82.8	85.3	83.6	84.9	81.7	86.4	85.8	85.6	85.9	85.7
CNN+Bi-GRU	87.3	89.2	87.5	88.8	85.7	90.4	89.2	89.0	89.3	89.1
CPAC	88.4	90.3	88.5	89.8	86.7	91.6	90.7	90.5	90.8	90.6
HMoE-SiMBA	**90.4**	**92.3**	**90.5**	**92.8**	**88.7**	**93.4**	**92.2**	**92.0**	**92.3**	**92.1**

Table 2 shows HMoE-SiMBA consistently outperforms all baselines across metrics and emotion categories. HMoE-SiMBA achieves 92.2% overall accuracy, exceeding the best baseline (CPAC) by 1.5 percentage points. Class-wise accuracy also demonstrates HMoE-SiMBA's superior performance in recognizing each emotion. This is attributed to the synergistic integration of the Heterogeneous Feature Extractor Mixture-of-Experts (HMoE) Network and SiMBA Attention Mechanism. Unlike traditional methods and monolithic deep learning architectures, HMoE-SiMBA uses a diverse expert ensemble to capture heterogeneous emotional cues from multi-modal features. The

Multi-Expert Routing Network dynamically aggregates expert outputs, enabling adaptive feature representation. Furthermore, the SiMBA Attention Mechanism efficiently filters frequency-domain information and models sequence context with reduced computational complexity, enhancing HMoE-SiMBA's performance and efficiency.

4.3 Ablation Study (RQ2)

To evaluate each module's contribution and address RQ2, we conduct an ablation study on the CASIA dataset. Table 3 shows HMoE-SiMBA performance with key components removed.

Table 3. Ablation Study on the CASIA Dataset.

Model Variant	Accuracy (%)
HMoE-SiMBA (Full Model)	92.2
w/o SiMBA	89.5
w/o Multi-Expert Routing Network	90.1
w/o CNN Expert	91.3
w/o LSTM Expert	91.5
w/o Transformer Expert	91.0
w/o $\mathbf{F}_{mfcc}$	88.7
w/o $\mathbf{F}_{mel}$	89.1
w/o $\mathbf{F}_{chroma}$	91.8

Table 3 shows that removing any key module from HMoE-SiMBA decreases performance, highlighting each component's effectiveness. Removing SiMBA Attention (w/o SiMBA) significantly drops performance by 2.7%, demonstrating SiMBA's critical role in enhancing feature discrimination and efficiency. Removing the Multi-Expert Routing Network (w/o Multi-Expert Routing Network) also noticeably decreases accuracy (2.1%), highlighting dynamic expert weighting's importance for adaptive feature representation. Removing individual experts (w/o CNN Expert, w/o LSTM Expert, w/o Transformer Expert) causes marginal degradation, suggesting each expert provides unique and complementary information. Removing each multi-modal feature ($\mathbf{F}_{mfcc}$, $\mathbf{F}_{mel}$, $\mathbf{F}_{chroma}$) also reduces accuracy, with $\mathbf{F}_{mfcc}$ and $\mathbf{F}_{mel}$ removal causing larger drops than $\mathbf{F}_{chroma}$, indicating spectral features' relative importance in Mandarin Chinese SER. These results validate the effectiveness and necessity of each module within HMoE-SiMBA, confirming each component's positive contribution to overall performance.

4.4 Sensitivity Analysis (RQ3)

To investigate HMoE-SiMBA's sensitivity to key hyperparameters and address RQ3, we analyze the state dimension (D_{state}) of Mamba blocks in the HMoE Network and SiMBA Attention Mechanism. We vary D_{state} to assess its impact.

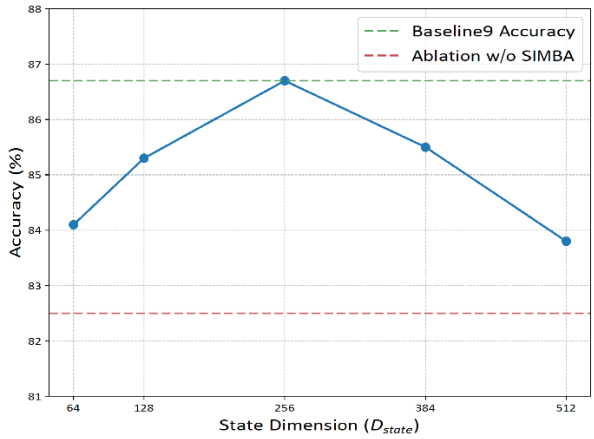

Fig. 2. Sensitivity to Mamba Block State Dimension (D_{state}).

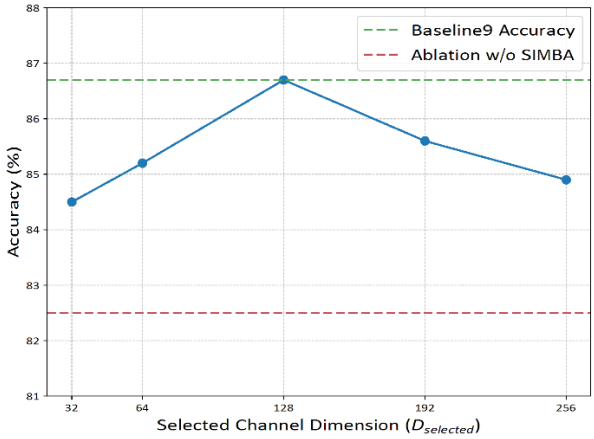

Fig. 3. Sensitivity to Selected Channel Dimension (D_{selected}).

Fig. 2 shows the impact of varying Mamba block state dimension (D_{state}) from 64 to 512. Performance improves as D_{state} increases from 64 to 256, peaking at D_{state}. Beyond $D_{state} = 256$, performance slightly declines, suggesting overfitting or diminishing returns. Smaller D_{state} may limit modeling complex temporal dependencies, while excessively large D_{state} may lead to overfitting and higher computational cost without significant gains. Optimal D_{state} tuning is crucial for balancing model capacity and generalization.

Fig. 3 shows HMoE-SiMBA's sensitivity to selected channel dimension $D_{selected}$ in SiMBA, varying from 32 to 256. Performance increases with $D_{selected}$ up to 128, indicating that selecting sufficient frequency channels is beneficial. However, beyond 128, performance plateaus and slightly decreases, suggesting excessively high-dimensional frequency channels may include redundant or noisy information, hindering focus on salient emotional cues. These analyses provide insights for hyperparameter tuning, guiding optimal parameter configuration for maximizing HMoE-SiMBA performance.

4.5 Impact of Data Augmentation (RQ4)

To evaluate data augmentation effectiveness and address **RQ4**, we compare HMoE-SiMBA performance with different augmentation configurations: (1) **No Augmentation**, (2) **Time Stretching Only**, (3) **Pitch Shifting Only**, (4) **Additive Noise Only**, and (5) **All Augmentations Combined**. Fig. 4 shows HMoE-SiMBA accuracy under these settings.

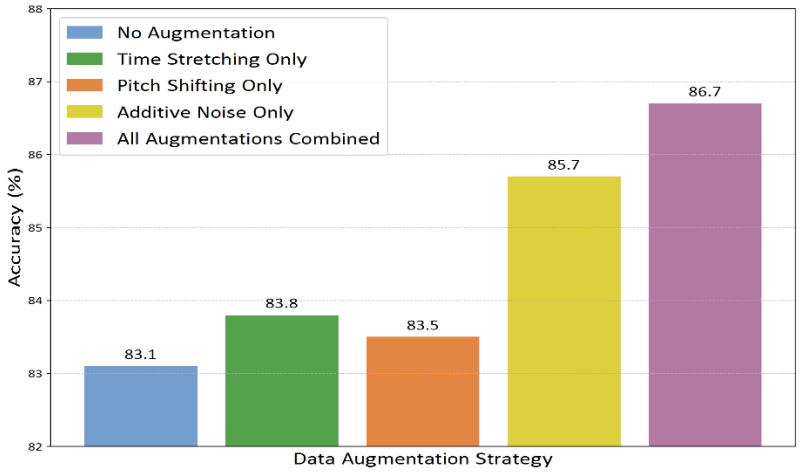

Fig. 4. Impact of Data Augmentation Strategies.

Fig. 4. Impact of Data Augmentation Strategies. shows data augmentation consistently improves HMoE-SiMBA performance compared to No Augmentation. Time Stretching Only provides modest improvement, while Pitch Shifting Only and Additive Noise Only yield more substantial gains, enhancing robustness to speaking rate, pitch, and noise variations. Combining all three augmentations (Time Stretching, Pitch Shifting, and Additive Noise) achieves the highest accuracy, outperforming other configurations. This demonstrates the synergistic effect of augmentations in improving generalization and robustness, particularly in diverse acoustic conditions of real-world SER tasks. Combined augmentation effectively increases training data diversity, allowing the model to learn more robust and invariant feature representations, improving emotion recognition accuracy.

4.6 Conclusion

This paper addressed critical challenges in Mandarin Chinese Speech Emotion Recognition (SER): feature representation, model generalization, and computational efficiency. We introduced **HMoE-SiMBA**, a novel framework integrating a **Heterogeneous Mixture-of-Experts (HMoE)** network with the **SiMBA**(**Si**mplified **M**amba-**B**ased **A**rchitecture) attention mechanism. Experiments on real-world Chinese speech emotion datasets demonstrate that HMoE-SiMBA achieves state-of-the-art accuracy (92.2%) and superior robustness compared to Transformer-based and SSRL baselines. The SiMBA mechanism reduces computational complexity while enhancing feature discriminability. Future work will focus on further improving robustness in noisy environments and exploring real-time deployment and applications across diverse languages and emotional contexts.

References

1. Singh, Youddha Beer, and Shivani Goel. "A systematic literature review of speech emotion recognition approaches." Neurocomputing 492 (2022): 245-263.Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016)
2. Lin, Yi-Lin, and Gang Wei. "Speech emotion recognition based on HMM and SVM." 2005 international conference on machine learning and cybernetics. Vol. 8. IEEE, 2005.
3. Chen, Jingyuan, et al. "Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention." Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval. 2017.
4. Atrey, Pradeep K., et al. "Multimodal fusion for multimedia analysis: a survey." Multimedia systems 16 (2010): 345-379.
5. Chang, Heng-Jui, Shu-wen Yang, and Hung-yi Lee. "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.
6. Chen, Li-Wei, and Alexander Rudnicky. "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.
7. Tao, Zhulin, et al. "Self-supervised learning for multimedia recommendation." IEEE Transactions on Multimedia 25 (2022): 5107-5116.
8. Nefian, Ara V., et al. "Dynamic Bayesian networks for audio-visual speech recognition." EURASIP Journal on Advances in Signal Processing 2002 (2002): 1-15.
9. Wang, Kunxia, et al. "Speech emotion recognition using Fourier parameters." IEEE Transactions on affective computing 6.1 (2015): 69-75.
10. Lanjewar, Rahul B., Swarup Mathurkar, and Nilesh Patel. "Implementation and comparison of speech emotion recognition system using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques." Procedia computer science 49 (2015): 50-57.
11. Khalil, Ruhul Amin, et al. "Speech emotion recognition using deep learning techniques: A review." IEEE access 7 (2019): 117327-117345.
12. Jahangir, Rashid, et al. "Deep learning approaches for speech emotion recognition: state of the art and research challenges." Multimedia Tools and Applications 80.16 (2021): 23745-23812.

13. Thakur, Anuja, and Sanjeev Dhull. "Speech emotion recognition: A review." International Conference on Advanced Communication and Computational Technology. Singapore: Springer Nature Singapore, 2019.
14. Mohamed, Abdelrahman, et al. "Self-supervised speech representation learning: A review." IEEE Journal of Selected Topics in Signal Processing 16.6 (2022): 1179-1210.
15. Luo, Hui, and Jiqing Han. "Cross-corpus speech emotion recognition using semi-supervised transfer non-negative matrix factorization with adaptation regularization." INTERSPEECH. 2019.
16. Lin, Wei-wei, Man-Wai Mak, and Jen-Tzung Chien. "Multisource i-vectors domain adaptation using maximum mean discrepancy based autoencoders." IEEE/ACM Transactions on Audio, Speech, and Language Processing 26.12 (2018): 2412-2422.
17. Gu, Albert, Karan Goel, and Christopher Ré. "Efficiently modeling long sequences with structured state spaces." arXiv preprint arXiv:2111.00396 (2021).
18. Gu, Albert, et al. "On the parameterization and initialization of diagonal state space models." Advances in Neural Information Processing Systems 35 (2022): 35971-35983.
19. Tang, Haoran, et al. "Muse: Mamba is efficient multi-scale learner for text-video retrieval." arXiv preprint arXiv:2408.10575 (2024).
20. Smith, Jimmy TH, Andrew Warrington, and Scott W. Linderman. "Simplified state space layers for sequence modeling." *arXiv preprint arXiv:2208.04933* (2022).
21. Fu, Daniel Y., et al. "Hungry hungry hippos: Towards language modeling with state space models." *arXiv preprint arXiv:2212.14052* (2022).
22. Mehta, Harsh, et al. "Long range language modeling via gated state spaces." *arXiv preprint arXiv:2206.13947* (2022).
23. Gu, Albert, and Tri Dao. "Mamba: Linear-time sequence modeling with selective state spaces." *arXiv preprint arXiv:2312.00752* (2023).
24. Zou, Zhengxia, et al. "Object detection in 20 years: A survey." *Proceedings of the IEEE* 111.3 (2023): 257-276.
25. Ma, Jun, Feifei Li, and Bo Wang. "U-mamba: Enhancing long-range dependency for biomedical image segmentation." *arXiv preprint arXiv:2401.04722* (2024).
26. Laroche, Jean, and Mark Dolson. "Improved phase vocoder time-scale modification of audio." *IEEE Transactions on Speech and Audio processing* 7.3 (1999): 323-332.
27. Lent, Keith. "An efficient method for pitch shifting digitally sampled sounds." *Computer Music Journal* 13.4 (1989): 65-71.
28. Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).
29. McFee, Brian, et al. "librosa: Audio and music signal analysis in python." *SciPy* 2015 (2015): 18-24.
30. Patro B N, Agneeswaran V S. "SiMBA: Simplified mamba-based architecture for vision and multivariate time series." *arXiv preprint arXiv:2403.15360* (2024).