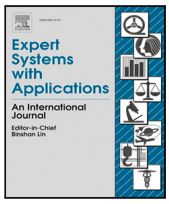
Expert Systems With Applications 273 (2025) 126855

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



MTLSER: Multi-task learning enhanced speech emotion recognition with pre-trained acoustic model

Zengzhao Chen [a,b], Chuan Liu [a], Zhifeng Wang [a,*], Chuanxu Zhao [a], Mengting Lin [a], Qiuyu Zheng [a]

[a] Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan, 430079, China
[b] National Intelligent Society Governance Experiment Base (Education), Wuhan, 430079, China

ARTICLE INFO

Keywords:
Multi-task learning
Speech emotion recognition
Speaker identification
Automatic speech recognition
Speech representation learning

ABSTRACT

This study proposes a novel Speech Emotion Recognition (SER) approach employing a Multi-Task Learning framework (MTLSER), designed to boost recognition accuracy by training multiple related tasks simultaneously and sharing information via a joint loss function. This framework integrates SER as the primary task, with Automatic Speech Recognition (ASR) and speaker identification serving as auxiliary tasks. Feature extraction is conducted using the pre-trained wav2vec2.0 model, which acts as a shared layer within our multi-task learning (MTL) framework. Extracted features are then processed in parallel by the three tasks. The contributions of auxiliary tasks are adjusted through hyperparameters, and their loss functions are amalgamated into a singular joint loss function for effective backpropagation. This optimization refines the model's internal parameters. Our method's efficacy is tested during the inference stage, where the model concurrently outputs the emotion, textual content, and speaker identity from the input audio. We conducted ablation studies and a sensitivity analysis on the hyperparameters to determine the optimal settings for emotion recognition. The performance of our proposed MTLSER model is evaluated using the public IEMOCAP dataset. Results from extensive testing show a significant improvement over traditional methods, achieving a Weighted Accuracy (WA) of 82.63% and an Unweighted Accuracy (UA) of 82.19%. These findings affirm the effectiveness and robustness of our approach. Our code is publicly available at https://github.com/CCNU-nercel-lc/MTL-SER.

1. Introduction

Speech is one of the most direct and effective ways for people to convey information, containing both linguistic and paralinguistic information (Wani, Gunawan, Qadri, Kartiwi, & Ambikairajah, 2021; Zeng et al., 2024). The primary goal of a SER system is to accurately identify the emotional state embedded within a speaker's speech signal by analyzing and understanding it. Studies show that it is difficult to reliably analyze and predict human emotions from linguistic information (Ambady & Rosenthal, 1992). However, the paralinguistic information in speech provides a wealth of acoustic features that encode the speaker's emotional state, such as prosodic and spectral features (Latif et al., 2023; Wang et al., 2022). Due to the inherent limitations of handcrafted features, various deep learning techniques are also proposed and used for automatic feature extraction from speech signals (Jahangir, Teh, Hanif, & Mujtaba, 2021; Wang, Zhan, Zhang, Ouyang, & Guo, 2023; Zeng et al., 2024). A typical ASR system can be considered a collection of methods for separating, extracting, and classifying speech signals to detect the emotions embedded within them (Atmaja & Sasou, 2022). In recent years, SER technology has found widespread applications in fields such as intelligent healthcare, vehicle driving, and call centers (Jahangir et al., 2021).

SER mainly consists of two steps: extracting speech features and selecting and classifying these features (Liu, Liu, Wang, Guo, & Dang, 2020). Traditional machine learning methods usually extract paralinguistic features of speech, including fundamental frequency, intensity, linear prediction coefficients (LPC), and Mel frequency cepstral coefficients (MFCC) (Alu, Zoltan, & Stoica, 2017). Based on these features, classic models such as the Gaussian Mixture Model (GMM) (Cheng & Duan, 2012/08), the Hidden Markov Model (HMM) (Schuller, Rigoll, & Lang, 2003), the Support Vector Machine (SVM) (Jain et al., 2020), the Random Forest (RF) (Noroozi, Sapiński, Kamińska, & Anbarjafari, 2017), and the K-Nearest Neighbors (KNN) (Umamaheswari & Akila, 2019) are used for classification. These features and classification methods all have their applicable scenarios and advantages. For

* Corresponding author.
E-mail addresses: zzchen@ccnu.edu.cn (Z. Chen), liuchuan@mails.ccnu.edu.cn (C. Liu), zfwang@ccnu.edu.cn (Z. Wang), cxzhao@mails.ccnu.edu.cn (C. Zhao), linmengting@mails.ccnu.edu.cn (M. Lin), qiuyu@mails.ccnu.edu.cn (Q. Zheng).

https://doi.org/10.1016/j.eswa.2025.126855
Received 5 May 2024; Received in revised form 23 December 2024; Accepted 10 February 2025
Available online 18 February 2025

example, Zhou, Sun, Zhang, and Yan (2009) use a GMM-based supervector SVM to fuse spectral features with prosodic features to obtain the final decision. Although researchers design many clever feature extraction and classification models, there are still some limitations in the effectiveness of SER.

With the development of deep learning (Zeng, Feng, Zhu, & Wang, 2023), more researchers use deep features for SER. Compared with traditional machine learning methods, deep learning methods eliminate the step of extracting complex acoustic features, achieve end-to-end SER, and extract more emotional information from speech signals. Deep Neural Networks (DNN) (Li et al., 2013) can capture the potential non-linear relationships between data. Many neural network models are proven effective in SER tasks, such as Convolutional Neural Networks (CNN) (Huang, Dong, Mao, & Zhan, 2014), Deep Belief Networks (DBN) (Le & Provost, 2013), Long Short-Term Memory networks (LSTM) (Xie et al., 2019), and Recurrent Neural Networks (RNN) (Mirsamadi, Barsoum, & Zhang, 2017; Zeng et al., 2024). For instance, Zhu, Chen, Zhao, Zhou, and Zhang (2017) combine DBN with SVM to achieve excellent performance in the Chinese speech database. However, the features contained in speech signals are complex and vary. Single-task SER ignores potential information in speech affecting emotions, such as the identity of the speaker. Such information generalizes and improves the performance of the SER system (Zhang, Wu and Schuller, 2019).

Adopt MTL as a learning paradigm to enhance the generalization performance of all tasks by leveraging useful information embedded in multiple related tasks. MTL can be integrated with various other learning paradigms to boost the effectiveness of learning tasks, such as semi-supervised learning, active learning, and more (Zhang & Yang, 2022). In multi-task SER, the SER as the main task and its related auxiliary tasks are solved simultaneously (Latif et al., 2022). During the process of training multiple related tasks simultaneously, information exchange between different tasks can be achieved through parameter sharing, thereby improving their own learning effects. Therefore, MTL is widely used in various natural language processing related problems (Atmaja, Sasou and Akagi, 2022; Yunxiang & Kexin, 2023). Some research has pointed out that the SER method based on MTL is more effective than the single-task method (Kim et al., 2017). Similarly, cooperative multi-agent reinforcement learning methods outperform single-agent counterparts (Wen, Fu, Dai, & Zhou, 2021). However, the emotion information extracted from speech by simple neural networks is very limited, and the model pre-trained with a large amount of speech data can obtain higher-quality speech features from the original speech signal, which provides a way to further improve the performance of the SER system (Atmaja, Zanjabila and Sasou, 2022). For example, Sharma (2022) proposed a multi-language and SER system based on the pre-trained model wav2vec2.0, and their validation results on the corpus showed the effectiveness of the model. In addition, in the multi-task SER system, the selection of auxiliary tasks is crucial. Auxiliary tasks related to the main task will be given priority. Li, Zhao, and Kawahara (2019) used gender recognition as an auxiliary task, and Cai, Yuan, Zheng, Huang, and Church (2021) selected ASR as an auxiliary task based on the use of the pre-trained model wav2vec2.0. Their methods have all improved the performance of the SER task. However, the auxiliary tasks used in previous research have certain limitations, and the emotion information contained in the extracted speech signals is not comprehensive enough.

As outlined above, the performance of traditional machine learning methods in Speech Emotion Recognition (SER) tasks is limited by feature selection and model design. Single-task deep learning approaches often fail to capture latent emotional information embedded in speech signals, whereas multitask learning (MTL) methods are generally more effective. In an MTL framework, the selection of auxiliary tasks plays a pivotal role. Therefore, our motivation is to better extract emotional information from speech, address the limitations of traditional feature-based and single-task deep learning methods, and improve the accuracy of SER. In this study, we propose a multitask learning-based framework for speech emotion recognition. Leveraging a pre-trained acoustic model to extract deep features, we incorporate two auxiliary tasks that are closely related to emotion recognition. This approach enables the extraction of richer emotional information from speech signals and significantly enhances the accuracy of the SER system.

The main contributions of this paper are summarized as follows:

- **Proposing a Novel MTL Framework for SER:** This paper introduces an innovative MTL framework specifically designed for SER. Unlike traditional single-task deep learning methods, the proposed framework positions SER as the primary task, augmented by ASR and speaker recognition(SR) as auxiliary tasks. By harnessing the inherent synergy among these interrelated tasks, the framework effectively exploits shared and complementary information in speech data, leading to significant improvements in emotion recognition accuracy.
- **End-to-End Feature Extraction Using Pre-Trained Acoustic Models:** We present a new methodology that utilizes pre-trained acoustic models for end-to-end deep feature extraction from speech signals. This eliminates the reliance on manually engineered and often complex feature extraction processes. Trained on extensive corpora of unlabeled speech data, these pre-trained models autonomously extract robust and versatile features that are resilient across diverse speakers and environmental conditions. These features enable superior performance in the target tasks by capturing the fundamental structural characteristics of speech signals.
- **Comprehensive Analysis of Auxiliary Tasks and Model Sensitivity:** Through ablation studies, we systematically quantify the contribution of auxiliary tasks to SER performance, demonstrating the substantial benefits of integrating ASR and SR. Furthermore, sensitivity analysis was conducted to evaluate the impact of various hyperparameters, resulting in the identification of an optimal configuration that maximizes the model's recognition accuracy.

The rest of this paper is structured as follows: Section 2 briefly reviews the current work related to SER. In Section 3, we introduce the concepts that appear in this paper and define the related problems of this research. Section 4 introduces the SER method based on MTL. Section 5 introduces the dataset and evaluation metrics, provides the details of the experiments and analysis of the results, and conducts ablation experiments for the two auxiliary tasks respectively. Finally, Section 6 summarizes the entire paper.

2. Related work

SER is usually regarded as a classification task, where the input speech is recognized and labeled with emotional tags such as happiness, sadness, anger, and neutrality (Cai et al., 2021). One of the daunting tasks of SER is to identify and extract the most suitable information for calculating, recognizing, and distinguishing emotions from speech (Madanian et al., 2023). This section introduces the related work of SER, mainly divided into three parts: SER using traditional methods, single-task SER based on deep learning, and multi-task SER based on deep learning. Finally, based on the experience of predecessors, we propose the multi-task emotion recognition method we use.

2.1. SER based on statistical machine learning

Human speech contains linguistic features and paralinguistic features. Linguistic features refer to the basic elements that constitute language, including vocabulary, grammar, etc. Paralinguistic features refer to features that occur with speech or affect speech, such as prosodic features, stress, spectral features.

SER techniques based on statistical machine learning often use manually extracted acoustic features and apply classification algorithms based on these extracted acoustic features. For example, Kwon, Chan, Hao, and Lee (2003) chose pitch, log energy, formants, Mel band energy, and MFCCs as basic features, and added the velocity/acceleration of the fundamental frequency and MFCCs to form a feature stream. They used Quadratic Discriminant Analysis (QDA), SVM, Linear Discriminant Analysis (LDA), and HMM to analyze the extracted features, and found that pitch and energy are the most important factors in SER. Koolagudi, Murthy, and Bhaskar (2018) studied how to choose classifiers based on the characteristics of the dataset, explored different combinations of spectral and prosodic features related to emotions, and conducted experiments using techniques such as K-means clustering and vector quantization. Le and Provost (2013) proposed and evaluated a hybrid classifier based on HMM and DBN. This network can simulate complex and nonlinear high-level relationships between low-level features, providing insights into important similarities and differences between speech and emotions. Jha, Kavya, Christopher, and Arunachalam (2022) proposed a SER framework that can work in real-time environments, using non-lexical or paralinguistic properties of speech to train supervised machine learning models for emotion recognition. The experimental analysis used a combination of prosodic and spectral features, and classification used algorithms such as Gaussian Naive Bayes, Random Forests, k-Nearest Neighbors, Support Vector Machines, and Multilayer Perceptrons (MLP). The analysis showed that SVM and MLP performed best. Recently, Costantini, Parada-Cabaleiro, Casali, and Cesarini (2022) explored the feasibility and characteristics of cross-language, cross-gender SER. They applied three ML classifiers (SVM, Naïve Bayes, and MLP) to acoustic features. The results on the Emoffilm database showed that MLP is the most effective classifier, cross-gender tasks are more difficult than tasks involving two languages, and the differences in emotions expressed by male and female subjects are greater than the differences between different languages. However, these SER methods based on statistical machine learning involve complex manual feature extraction and the selection of classification algorithms. The emotional information in speech is difficult to fully utilize, and the recognition effect is limited.

2.2. Single-task SER based on deep learning

With the rapid development of deep learning, there are many studies in the field of SER using deep learning methods. These studies mostly focus on the model structure to train a single SER task.

Fahad, Deepak, Pradhan, and Yadav (2021) proposed an SER system based on DNN-HMM. They used the Maximum Likelihood Linear Regression technique in the feature space for speaker adaptation during the training and testing stages to adapt to the changes in acoustic features of different speakers. At the same time, they used MFCC and epoch-based features. Their results highlighted the importance of speaker adaptation to the SER system and emphasized the complementarity of MFCC and epoch-based features in using speech for SER. Li, Bell, and Lai (2022) studied how to use language features based on speech transcription to improve SER performance. They proposed to integrate the output of ASR into the joint training of SER, examined various ASR outputs and fusion methods, and the experimental results showed that in the joint ASR-SER training, using a hierarchical co-attention fusion method combined with ASR's hidden output and text output can maximize the performance of SER. Attention mechanisms can also be employed to enhance the architecture of CNN, improving computational efficiency (Geng, Li, Han, & Zhang, 2022). With the development of attention mechanisms, more and more researchers are devoting themselves to this area of research. For example, Lieskovská, Jakubec, Jarina, and Chmulík (2021) studied the impact of various attention mechanisms on SER performance. Ioannides, Owen, Fletcher, Rozgic, and Wang (2023) proposed a method to aggregate information from the outputs of various transformer layers of a general speech encoder (such as WavLM, HuBERT) for downstream task SER, reducing the model's prediction dependence on language content. Zou, Si, Chen, Rajan, and Chng (2022) proposed an end-to-end SER system. They used multi-level acoustic information and a newly designed co-attention module to extract and utilize deeper audio information. Recently, the method of creating downstream SER tasks based on pre-trained models and then fine-tuning the models is very popular. Chen, Xing, Chen and Xu (2023) proposed a research direction to optimize pre-trained models for specific tasks to generate more compact and effective specific task pre-trained models. In addition, integrating multi-modal features for SER is also an effective way. Maji, Swain, Guha, and Routray (2023) proposed a new cross-modal MSER interaction method based on speech and text.

However, the above single-task deep learning methods ignore the interrelatedness and complementarity between multiple tasks, and there is still room for improvement in the emotional information extracted from speech signals.

2.3. Multi-task learning

MTL (Caruana, 1997) is an important component of machine learning and a type of transfer learning. MTL improves the overall recognition performance, generalization, and robustness by simultaneously processing multiple related tasks, mining the relationships between related tasks, and sharing information and knowledge between different tasks.

There have been many researchers who have done significant work in the direction of multi-task SER, such as Li et al. (2019), who built an MTL framework that simultaneously trains emotion classification and gender classification, with the gender task as an auxiliary task. They directly extract features from spectrograms to replace traditional manual features and use a self-attention mechanism to focus attention on the prominent stages of emotion in speech. Cai et al. (2021) proposed a multi-task framework that simultaneously trains ASR and emotion classification, uses the pre-trained wav2vec2.0 model for speech feature extraction, and implements an end-to-end SER system. Additionally, they validated the positive impact of ASR tasks on emotion classification in ablation experiments and discussed the values of hyperparameters for auxiliary tasks. The results show that different hyperparameters have different impacts on ASR tasks, and there is an optimal choice of hyperparameters. wook Lee (2023) proposed a method to effectively use feature space by using MTL to map the same speech to different clusters to extract different features, and then study fusion methods according to the relevance between different mapping features to obtain more complementary information. The combination of multi-modal and multi-task is also a research direction for many researchers. Ghosh, Tyagi, Ramaneswaran, Srivastava, and Manocha (2023) proposed a SER system based on the multi-modal MTL method MMER, which uses multiple modal inputs of audio, text, enhanced audio, and enhanced text. They use a multi-modal network based on early fusion and cross-modal self-attention between text and acoustic modalities and employ three auxiliary tasks: SER based on supervised contrastive learning, ASR, and SER based on enhanced contrastive learning. Khare, Parthasarathy, and Sundaram (2020) built a multi-modal architecture for machine learning tasks using audio, video, and text information, used Transformers to extract features from multi-modal inputs, and performed SER, ASR, and SR tasks using MTL on embedded features.

Summary: Traditional SER methods mainly focus on the study of manually extracted acoustic features of speech, but there are still limitations in extracting emotional features from them. Single-task deep learning methods have not fully utilized the feature information extracted by deep networks, and there is still room for improvement in recognition results. The tasks used in current multi-task SER methods are not the same, and the recognition results also vary greatly. In order to improve the performance and effect of SER, this paper proposes a deep learning method based on an MTL framework. The three tasks trained are SER, ASR, and SR.

Table 1
Main notations and descriptions.

Notation	Description
X, x	Audio features
A, A^*, a	Text sequences
N	Length of audio features
M	Length of text sequences
$Audio_{raw}$	Input raw speech audio
FC	Fully connected layer
PL	Pooling layer
P_σ	Pre-trained model, where σ represents model parameters
d	Last hidden layer dimension of the pre-trained model
v	Shared features
v^*	Pooled shared features
e	Emotional prediction logits vector
s	Speaker prediction logits vector
y'	Text prediction logits vector

3. Preliminaries

In this section, we provide the definitions of the tasks involved in multi-task SER and introduce the related concepts. At the same time, we give a brief introduction to the MTL framework. Finally, combining the related concepts mentioned earlier, we propose the SER model framework based on MTL used in this paper. The primary mathematical notations and their descriptions in this paper are presented in Table 1.

3.1. Definition and concepts

Definition 1 (*ASR Task*)**.** We use the ASR task as an auxiliary task for multi-task SER. The purpose of this task is to convert the input speech signal into text for output. Suppose $X = [x_1, x_2, \ldots, x_{N-1}, x_N]$ is the audio feature extracted from the input speech, where N is the length of the audio features. Let $A = [a_1, a_2, \ldots, a_{M-1}, a_M]$ denote the corresponding text sequence, where M is the length of the text sequence. The conditional probability of the text sequence A given the audio features X is represented as $P(A|X)$. The goal of the ASR task is to find the text sequence A^* that maximizes $P(A|X)$ among all possible text sequences. Mathematically, this can be expressed as:

$$A^* = \underset{A}{\operatorname{argmax}} \left(P(A|X)\right) \tag{1}$$

Here, this text sequence A^* is the most likely text corresponding to the input audio features X.

Definition 2 (*SR Task*)**.** The SR task is to convert the given speech signal feature $X = [x_1, x_2, \ldots, x_{N-1}, x_N]$ into the corresponding speaker label s^*. The mathematical representation is as follows:

$$s^* = \underset{s'}{\operatorname{argmax}} \{f(K_1, X, W), f(K_2, X, W), \ldots, f(K_S, X, W)\} \tag{2}$$

Here, $f(\cdot)$ is the network model for predicting speaker labels, $\{K_s|s' = 1, 2, \ldots, S\}$ represents the speaker features of the training speech, S is the number of speaker labels in the training set, and, W represents the backend parameters of the model. We also use this task as another auxiliary task for multi-task SER.

Definition 3 (*SER Task*)**.** In multi-task SER, the SER task will be trained as our main task. The purpose is to recognize the input speech signal feature $X = [x_1, x_2, \ldots, x_{N-1}, x_N]$ as a predefined emotion label e^* for output. The mathematical representation of this task is as follows:

$$e^* = \underset{e'}{\operatorname{argmax}} \{g(Y_1, X, W'), g(Y_2, X, W'), \ldots, g(Y_E, X, W')\} \tag{3}$$

Here, $g(\cdot)$ refers to the neural network model used for predicting speech emotion, $\{Y_e|e' = 1, 2, \ldots, E\}$ represents the emotional features present in the training set, E denotes the number of emotion labels, and W' corresponds to the model parameters.

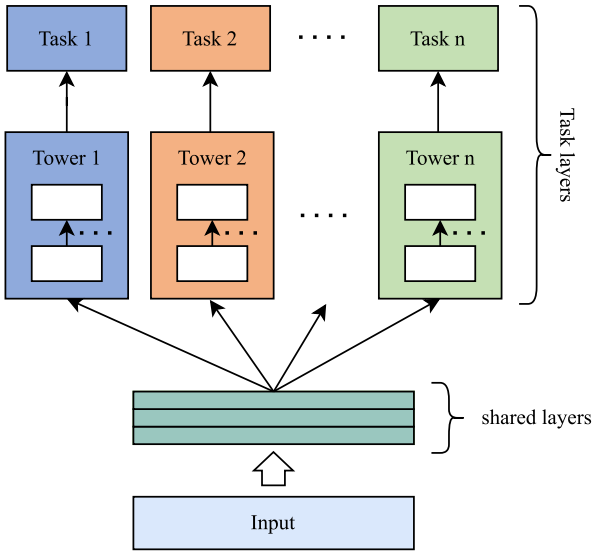


Fig. 1. The MTL framework.

Definition 4 (*MTL Framework*)**.** Unlike single-task training frameworks, MTL uses a shared backbone model to simultaneously optimize multiple objectives in different tasks. The advantage of this is that multiple tasks can improve each other's performance by sharing information and complementing each other. Experience has shown that training multiple related tasks simultaneously will yield better performance. MTL learns multiple related tasks simultaneously by sharing features, making the training results of these tasks better than training them separately. Essentially, it is a form of transfer learning. The MTL framework is mainly divided into a feature sharing layer and a task layer. The features extracted from the input are no longer used independently by a single task, but are shared by multiple tasks. Different tasks will process the features differently to train their specific tasks. This kind of MTL can avoid overfitting problems and improve the generalization ability of the model. In the MTL framework, there are two strategies for the design of the shared layer: hard parameter sharing and soft parameter sharing. We use the former as the design idea of the shared layer. The MTL framework is shown in Fig. 1.

3.2. Problem formulation

Definition 5 (*Problem of SER Based on MTL*)**.** The information contained in speech signals is complex and diverse. Single-task training models have difficulty fully utilizing the extracted features. We hypothesize that there is a certain correlation between ASR, SR, and SER tasks. To improve the utilization rate of features, we introduce an MTL framework to perform SER task. We train three tasks separately: SER, ASR, and SR. These tasks have a certain degree of relevance and complement each other in terms of feature utilization. At the same time, through the setting of hyperparameters, we distinguish SER as the main task and the other two tasks as auxiliary tasks, which makes the generalization of SER better. After the original audio is input, we extract shared features from it. The SER task inputs the extracted features into the SER network for emotion classification and outputs the predicted emotion labels. The ASR task performs speech recognition and outputs the predicted text content. The SR task predicts the identity of the speaker corresponding to the input audio and outputs its speaker label. The multi-task SER framework is shown in Fig. 2. We combine the loss functions of the

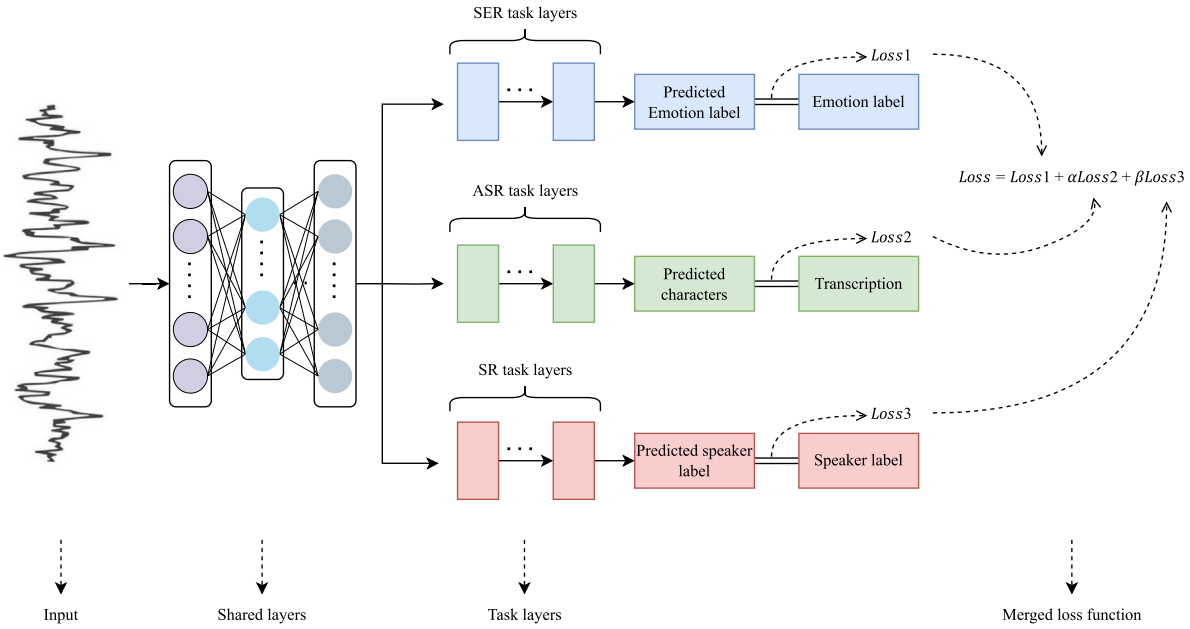


Fig. 2. SER framework based on MTL.

three tasks into one loss function. During the training process, this loss function will serve as their common loss function for backpropagation. The hyperparameters α and β represent the weights of the auxiliary tasks.

In this section, we provide the relevant definitions and concepts of SER based on MTL, which lays the foundation for understanding the subsequent multi-task SER framework.

4. Methods

In order to make full use of the information in speech features and improve the accuracy of SER, this paper proposes an end-to-end SER model based on MTL(MTLSER). MTLSER is divided into two parts: shared feature extraction and MTL. The shared feature extraction part uses the pre-trained model wav2vec2.0, and the extracted features are separately entered into three downstream tasks with certain relevance for training. The three tasks are SER, ASR, and SR. The joint loss of multiple tasks will be backpropagated to fine-tune the model parameters. The specific model framework is shown in Fig. 3.

4.1. Shared layer of MTL framework : pre-trained acoustic model

One of the main differences between the MTL framework and single-task learning is the shared layer. By sharing lower-level parameters, the risk of overfitting is reduced, making it suitable for handling tasks with strong relevance. Pre-trained acoustic models are very powerful. Common ones include Wav2vec, HuBert, WavLM, to name a few. These models learn from a large amount of unlabeled speech data through self-supervised learning, and can convert raw speech signals into high-quality feature representations. These representations can be used for various speech-related downstream tasks. In the model we proposed, we use the pre-trained wav2vec2.0 as the shared layer. The parameters of this model are the shared parameters, which are shared by multiple downstream tasks that follow. Fig. 3 shows the structure of this model in the shared layer part.

From an overall perspective, wav2vec2.0 consists of a feature encoder, quantization module, and context network (Baevski, Zhou, Mohamed, & Auli, 2020). The role of the feature extractor is to reduce the dimensionality of audio data, converting the original waveform into a series of feature vectors. These feature vectors correspond to different time segments of the audio, usually one feature vector every 20 ms, and they are called hidden speech representations. The feature extraction module includes a 7-layer convolutional neural network, each layer has 512 channels. During the pretraining process of wav2vec2.0, these continuous hidden variables pass through the quantization module, which maps them to a set of discrete quantized vectors q, thereby reducing the complexity of the model. At the same time, wav2vec2.0 masks these hidden representations, with the aim of generating spans in the hidden representations for contrastive learning. This task helps the model learn the contextual information of speech. Then, these masked hidden representations are input into the encoder of the Transformer network to capture long-distance dependencies in the sequence, thereby better representing speech features. Finally, the output of the Transformer encoder is compared with the quantized vector q for contrastive learning, to obtain the output features closest to q.

In summary, in the MTL framework we proposed, the pre-trained acoustic model can be used as a shared layer to extract features from speech data. We chose the pre-trained wav2vec2.0 model as the shared layer. The shared parameters are the weight parameters of various modules in the model structure, such as CNN, Transformer. During the training process, the shared parameters are updated by fine-tuning.

4.2. Learning of shared feature representations

One of the keys to the MTL framework is to find shared feature representations. By sharing feature representations, the model can share relevant information between different tasks, thereby improving the generalization ability and effectiveness of the model. This paper uses the pre-trained wav2vec2.0 to extract shared features. In our model, wav2vec2.0 takes the original speech as input and extracts abstract audio representation features from it. The subsequent training tasks will use the shared features as input for training.

Let the pre-trained model be represented as $P_\sigma(\cdot)$, where σ is the parameter related to this model, d is the dimension of the last hidden layer of this pre-trained model, and the input audio is $Audio_{raw}(Length = L)$, L is the length of the input $Audio_{raw}$, then the shared feature obtained is:

$$v = P_\sigma(Audio_{raw}) \in R^{L \times d} \tag{4}$$

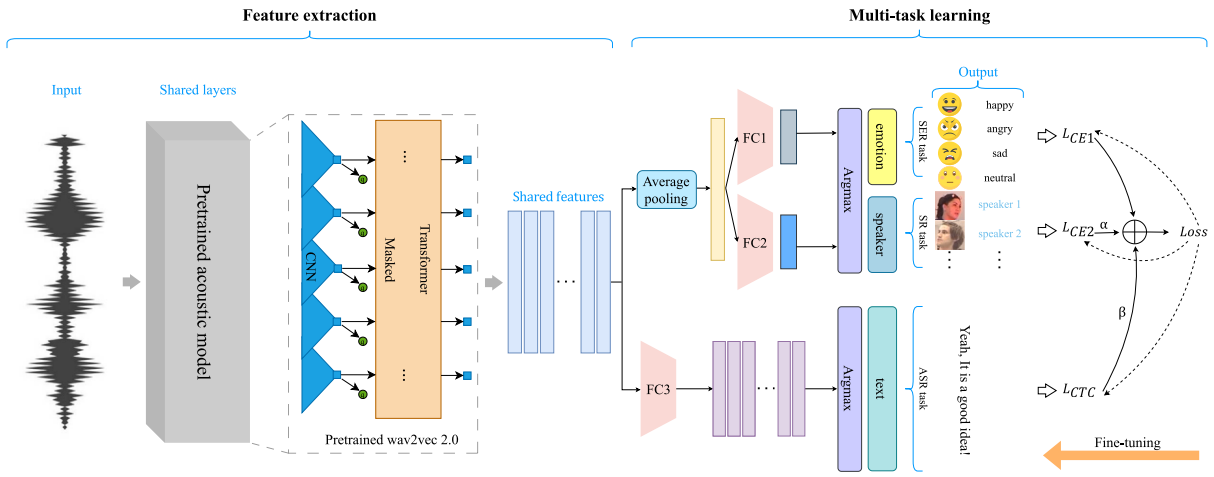


Fig. 3. The proposed SER framework based on MTL.

After the shared feature extraction is completed, we input the feature v into three different downstream tasks for training. The specific process of the three tasks will be introduced below.

4.3. SER task

In this section, we will introduce the main task in multi-task SER. This task obtains the emotional information in the speech signal after the shared features are processed at different levels, and obtains the predicted emotional labels.

Specifically, in this part, after we obtain the feature vector v extracted by the pre-trained model, we first perform average pooling on it. The pooling layer can downsample the input data, reduce the amount of data, and retain important feature information. The average pooling layer here can reduce the dimension of acoustic features, thereby improving the processing speed of downstream tasks. We input the shared feature v into an average pooling layer PL, which performs average pooling on the sample length, converting the vector sequence $v \in R^d$ into a one-dimensional vector $v^* \in R^d$. The mathematical representation is as follows:

$$v^* = PL(v) \tag{5}$$

The feature v^* output after the average pooling layer integrates the sound signal into a unified feature representation, retaining important information. Then it is input into a fully connected layer FC_1. The fully connected layer maps the feature space to the emotional label sample space for emotional label classification. If there are E classification emotion labels in total, FC_1 will map v^* to the logical vector $e \in R^E$. Its mathematical expression is as follows:

$$e = FC_1(v^*) = FC_1(\sum_{i=0}^{L-1} v_i) \tag{6}$$

Finally, the logical vector e is calculated through the argmax layer to obtain the predicted value of the emotional label. In this task, we designed the model structure of the downstream task, using the average pooling layer and the fully connected layer to process the shared features, in order to fully mine the emotional information contained in the shared features.

4.4. SR task

SR is an auxiliary task in the MTL framework. Its purpose is to identify the speaker identity label represented by the input speech. Similar to SER, SR is a classification task (Zheng et al., 2023; Zheng, Chen, Wang, Liu, & Lin, 2024). We use a structure similar to SER in the processing of downstream tasks.

For the input shared feature v, the SR task and the SER task share an average pooling layer. This pooling layer performs dimensionality reduction on the shared features, retaining important feature information while also being more conducive to subsequent processing. After the shared feature v passes through the pooling layer PL, we get a single vector $v^* \in R^d$. Then it is input into another fully connected layer FC_2. This fully connected layer maps the feature space to the sample space of speaker labels for subsequent speaker identity classification. Assuming there are S speaker categories, we will get the logical vector $s \in R^S$ of the speaker. The formula is as follows:

$$s = FC_2(v^*) = FC_2(\sum_{i=0}^{L-1} v_i) \tag{7}$$

Like the processing of SER, we process the obtained logical vector s through the argmax layer to get the predicted speaker label. In the SR task, we extract the speaker identity information from the input audio. The speaker identity has a certain relevance to the SER task, which can supplement the emotional information that the aforementioned SER task failed to extract.

4.5. ASR task

ASR, as another auxiliary task in the MTL framework, aims to convert the input audio into English text. By extracting the text information, it enhances the accuracy of the main task of emotion recognition. The vocabulary used in this task has a total of $U = 32$ characters, including 26 English letters and 6 punctuation marks.

Wav2vec2.0 has an outstanding performance in the field of ASR. Its advantage is that it can build a very robust speech recognition system by training a small amount of data, and it also has excellent performance in the training of other downstream tasks. After obtaining the shared feature v, we input it into a fully connected layer FC_3. This fully connected layer maps the shared feature space to the character space of the vocabulary, that is, the feature $v \in R^d$ is mapped to the logical vector $y \in R^{L \times U}$, and finally, the character prediction represented by the logical vector is obtained. The mathematical representation is as follows:

$$y = FC_3(v) \tag{8}$$

Then, the logical prediction value of the character is converted into the probability distribution of the character y' through the softmax layer, thereby obtaining the final character prediction.

$$y' = softmax(y) \tag{9}$$

Algorithm 1: Speech Emotion Recognition Model Based on Multitask Learning

Input: $\mathcal{RA}$: Raw speech Audio(length=L)

Output: Emotion prediction labels;speaker prediction labels;speech prediction texts

Using the pre-trained wav2vec2.0 model P_σ(The last hidden layer dimension is d) for extracting shared feature

$$v \longleftarrow P_\sigma(\mathcal{RA}) \in R^{L\times d}$$

Input shared feature v into average pooling layer PL and fully connected layer FC_3

$$v^* \longleftarrow PL(v)$$

$$y \longleftarrow FC_3(v)$$

Speech emotion recognition: Input v^* into FC_1 to obtain predicted emotion labels

$$e \longleftarrow FC_1(v^*) \longleftarrow FC_1(\sum_{i=0}^{L-1} v_i)$$

Speaker recognition: Input v^* into FC_2 to obtain predicted speaker labels

$$s \longleftarrow FC_2(v^*) \longleftarrow FC_2(\sum_{i=0}^{L-1} v_i)$$

Automatic speech recognition: Input y into softmax layer to obtain predicted text

$$y' \longleftarrow softmax(y)$$

ASR recognizes the input speech signal as text, mining the emotional information related to the text contained in the speech. This task, together with the SR task, serves as an auxiliary task for SER, extracting more emotional information from the speech from different dimensions, thereby improving the accuracy of SER.

Algorithm 1 shows the algorithm of the SER model based on MTL that we proposed. It decomposes the algorithm process and explains the algorithm steps in detail.

4.6. MTL framework for SER

After determining the training tasks, setting the loss function of the MTL model is a crucial step. It includes the selection and synthesis of different task loss functions, as well as the setting of hyperparameters. In the model proposed in this paper, the training process is supervised. In the final stage of ASR, we use the softmax operator to convert its logical vector into a corresponding probability vector, thereby obtaining a text sequence, and calculate the CTC loss with the true transcription. At the end of SER and SR, we use argmax to get the prediction label, and calculate the cross-entropy loss with the true label. Finally, we combine these three loss functions into a joint loss function. This loss function will affect these three tasks simultaneously during the training process. Next, we will discuss the loss functions of these three tasks in detail:

For the SER task, the model calculates the cross entropy of the predicted probability vector and the true emotion label, and uses this cross entropy loss L_{CE1} as the loss function.

$$L_{CE1} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{K} e_{jk} log(p_{jk}) \tag{10}$$

Here, N is the number of samples, K is the number of emotion categories, e_{jk} is the true label of the kth emotion category of the jth sample, and p_{jk} is the model's predicted probability for the kth emotion category of the jth sample.

For the ASR task, when the length of the input and output signals is different or no alignment information is provided, Connectionist Temporal Classification (CTC) technology can map the input signal to the output signal. Since the length of the input signal and the output signal in this task is generally different, we calculate the CTC loss of the predicted character sequence and the true sequence transcription, which can effectively backpropagate the gradient. The CTC loss is:

$$L_{CTC} = -\frac{1}{N}\sum_{j=1}^{N} ln(p(\pi|Z_j)) \tag{11}$$

Here, N is the number of input samples, and, $p(\pi|Z_j)$ is the probability of the output corresponding to the true label π for the input sample Z_j.

For the SR task, the model calculates the cross entropy of the predicted probability vector and the true speaker label, and uses this cross entropy loss L_{CE2} as the loss function.

$$L_{CE2} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{m=1}^{M} s_{jm} log(p_{jm}) \tag{12}$$

Here, M is the number of speaker categories, s_{jm} is the true label of the mth speaker category of the jth sample, and p_{jm} is the model's predicted probability for the mth emotion category of the jth sample.

The model combines the above three loss functions into a total loss function. For this, we introduce two hyperparameters α and β, which control the weights of the CTC loss and the cross-entropy loss of SR, respectively. We find the optimal choices of α and β through grid search. Finally, the objective function to be optimized by the model is as follows:

$$Loss = L_{CE1} + \alpha L_{CTC} + \beta L_{CE2} \tag{13}$$

Algorithm 2 shows the optimization and training steps of the model proposed in this paper.

4.7. Theoretical time complexity analysis

In this section, we analyze the time complexity of the proposed algorithm. Our model integrates the wav2vec2.0 model with an MTL approach for SER, ASR, and SR tasks. The overall time complexity is primarily determined by the wav2vec2.0 model, the fully connected layers, and the loss calculations. The detailed analysis is as follows:

The wav2vec2.0 model comprises multiple convolutional layers and an attention mechanism. The time complexity of the convolutional layers is $O(N_{layer} \cdot L \cdot K_{size}^2 \cdot C)$, where N_{layer} is the number of layers, L is the input sequence length, K_{size} is the kernel size, and C is the number of channels. The time complexity of the attention mechanism

Algorithm 2: Algorithm for model optimization and training

Input: Speech Audio

Initialization: Parameters of the pre-trained wav2vec2.0(σ), Sample number $sn = N$, Loss of SER is L_{CE1}, Loss of ASR is L_{CTC} and weights α, Loss of SR is L_{CE2} and weights β, epoch $(E) = 100$

for $e \leftarrow 1$ **to** E **do**

  for $batch \leftarrow 1$ **to** $N/batchsize$ **do**

    Calculate losses:

$$L_{CE1} \leftarrow -\frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{K} e_{jk}\log(p_{jk})$$

$$L_{CE2} \leftarrow -\frac{1}{N}\sum_{j=1}^{N}\sum_{m=1}^{M} s_{jm}\log(p_{jm})$$

$$L_{CTC} = -\frac{1}{N}\sum_{j=1}^{N}\ln(p(\pi|Z_j))$$

$$Loss = L_{CE1} + \alpha L_{CTC} + \beta L_{CE2}$$

    Backpropagation and update model parameters

  end

end

return the model after updating the parameters

is $O(N_{layer} \cdot L^2 \cdot d)$, where d is the hidden layer dimension. Therefore, the overall time complexity of the wav2vec2.0 model is:

$$O(N_{layer} \cdot L \cdot K_{size}^2 \cdot C + N_{layer} \cdot L^2 \cdot d) \tag{14}$$

The model includes three fully connected layers, which are used for the language model, emotion classification, and speaker classification. Their time complexities are $O(L \cdot d \cdot V)$, $O(L \cdot d \cdot C_{ser})$ and $O(L \cdot d \cdot C_{sr})$ respectively, where V is the vocabulary size, C_{ser} is the number of emotion classes, and C_{sr} is the number of speaker classes.

The model calculates three types of loss: CTC loss, cross-entropy loss for emotion classification, and cross-entropy loss for speaker classification. The time complexity of CTC loss is $O(L \cdot U \cdot d)$, where L is the input sequence length and U is the target sequence length. The time complexities of the cross-entropy losses for emotion classification and SR are $O(L \cdot C_{ser})$ and $O(L \cdot C_{sr})$ respectively.

Combining the complexities of all parts, we derive the overall time complexity of the proposed algorithm. After simplification and merging, it can be expressed as:

$$O(N_{layer} \cdot L(L \cdot d + K_{size}^2 \cdot C) + L \cdot d \cdot (V + C_{ser} + C_{sr} + U)) \tag{15}$$

The time complexity is primarily influenced by L_{layer}, L and d. Therefore, optimizing and controlling these parameters in practical applications will directly impact the efficiency of the model.

5. Experimental results and analysis

The main purpose of this study is to explore the performance of our proposed SER model based on an MTL framework. Specifically, our research attempts to answer the following research questions:

RQ1: What are the differences between the SER method based on MTL proposed in this study and previous methods?

RQ2: How do different tasks affect the recognition of speech emotion?

RQ3: What is the impact of different hyperparameters on SER?

5.1. Dataset introduction

IEMOCAP (Busso et al., 2008) is a widely used dataset in the field of speech, which records about 12 h of speech emotions and text information from five dialogues and 10 actors. This study also conducts experiments on this dataset. We selected 5531 data points from it, with a sampling rate of 16 000 Hz for each sample and an audio duration ranging from 1s to 40 s. Emotion and text information were annotated for each, with emotions divided into four categories: Happy, Angry, Neutral, and Sad. The number of speech samples for different emotions in this dataset is shown in Table 2.

Table 2
The number of speech with different emotions in IEMOCAP.

Emotion	Count	Label
Neutral	1708	0
Happy	1636	1
Anger	1103	2
Sad	1084	3
Total	5531	{0,1,2,3}

5.2. Evaluation metrics

To effectively evaluate the performance of our proposed model, we employed speaker-independent 10-fold cross-validation to demonstrate the method's efficacy. The evaluation metrics employed included WA, UA, F1 score, and Standard Deviation (SD). Higher WA, UA, and F1 scores, along with lower SD, indicate better model performance. The training and validation sets comprised 10% and 90% of the IEMOCAP dataset, respectively. We also utilized Floating Point Operations Per Second (FLOPs) and the p-value obtained from a two-tailed Welch's t-test to assess the model's performance.

FLOPs measure the computational complexity of the model by counting the number of floating-point operations required for a forward pass. This metric is essential for understanding the model's efficiency in terms of both computational power and speed. Lower FLOPs signify a more efficient model, which is crucial for real-time applications and deployment on devices with limited computational resources. In addition, we provide the inference time for a single sample of models.

The statistical significance of the performance differences between our proposed model and baseline models was evaluated using a two-tailed Welch's t-test. This test is particularly suitable for comparing two groups with unequal variances and sample sizes. The p-value from the Welch's t-test indicates the probability that the observed differences are due to random chance. A lower p-value (typically less than 0.05) suggests that the performance differences are statistically significant.

Table 3
Important parameters in the experiment.

Parameter	Value
Sample frequency	16 kHz
Training epochs	100
Optimizer	Adam
Learning rate	10^{-5}
Batch size	8
Pre-trained acoustic model	Wav2vec2.0
α	0.1
β	0.8

5.3. Experimental setup

The experiments were conducted on hardware featuring an Intel(R) Xeon(R) Gold 6126 CPU, a Tesla V100 SXM2 GPU, and 32 GB of RAM. The proposed model was implemented, trained, and tuned using the PyTorch framework. Table 3 presents additional specific parameter details.

The two most critical hyperparameters for our model are α and β. Detailed descriptions of hyperparameter tuning can be found in Section 5.8. In the following experiments, we used the best-performing set of hyperparameters, specifically $\alpha = 0.1$ and $\beta = 0.8$. Additionally, the pre-trained model used in our experiments was wav2vec2.0. In Section 5.6.3, we conducted comparative experiments to explain the rationale for using this model.

5.4. Baseline

In this section, we present several baseline models for SER, demonstrating the superiority of our model in SER tasks. The specific comparison results are shown in Table 4.

SVM (Kurpukdee, Kasuriya, Chunwijitra, Wutiwiwatchai, & Lamsrichan, 2017): This method uses different SVM kernels (linear, RBF, and polynomial) and utterance-based features, the accuracy of 12 models was compared. The results showed that the SVM model using a polynomial kernel and utterance-based Fbank features achieved the highest average accuracy.

STRL-SER (Chen, Lin, Wang, Zheng and Liu, 2023): This method establishes deep emotion learning modules at the frame level and utterance level, respectively, obtaining the spatio-temporal feature representation at the frame level and the global feature representation at the utterance level. It also uses a multi-head attention mechanism to fuse the features.

AMSNet (Chen, Li et al., 2023): This method uses a multi-scale SER network, AMSNet, to fuse frame-level features and utterance-level features. It proposes a fusion method based on connection attention mechanism, increasing the complementarity between different features.

Wav2vec2.0+MTL (Cai et al., 2021): This method establishes a multi-task SER framework based on the pre-trained wav2vec2.0, using SER as the main task and ASR as the auxiliary task to obtain more emotional information from the audio.

TIM-Net (Ye et al., 2023): The TIM-Net used in this paper is a time emotion modeling method. This method can capture remote temporal dependencies through bidirectional temporal modeling and dynamically fuse multi-scale information to better adapt to changes in time scale.

MFCC+Spectrogram+Wav2vec2.0 (Zou et al., 2022): This method uses features extracted by MFCC, spectrogram, and wav2Vec2.0, and uses a common attention mechanism to extract complementary acoustic information from the original audio for SER.

Self-attention+MTL (Li et al., 2019): This method first extracts features from the speech spectrogram, then uses a self-attention mechanism to further obtain emotional features. Finally, it enhances SER performance by incorporating gender classification as an auxiliary task through MTL.

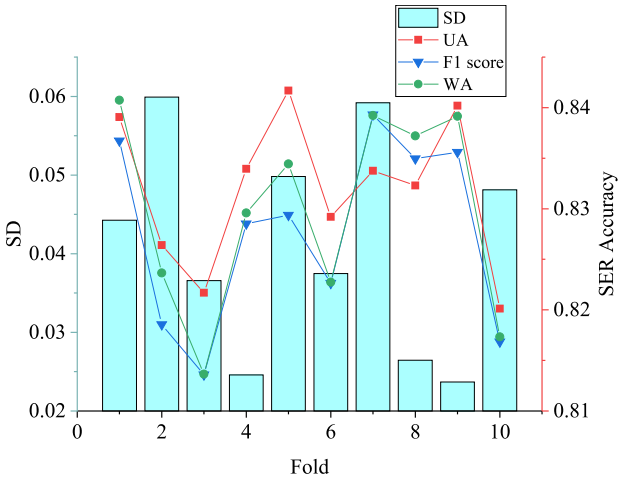


Fig. 4. 10-fold cross validation evaluation results.

MMER (Ghosh et al., 2023): This method utilizes multiple modalities, including audio, text, augmented audio, and augmented text. It also employs three auxiliary tasks: emotion recognition based on supervised contrastive learning, speech recognition, and emotion recognition based on augmented contrastive learning.

By comparing with these state-of-the-art methods, the superiority of our proposed model is demonstrated.

5.5. Comparison with baseline methods

5.5.1. Comparison and analysis of experimental results

We conducted experiments with our proposed model on the public dataset IEMOCAP and achieved excellent performance. Figs. 4–6 present the results of our model's performance under 10-fold cross-validation. To demonstrate the superiority of our model, we provide the prediction results for each fold. Fig. 4 displays the confusion matrices for each fold, while Fig. 6 shows the overall results averaged across all folds. From the confusion matrices, as well as the WA and UA metrics, it is evident that our model achieves highly accurate emotion classification on the IEMOCAP dataset. Table 4 shows the comparison results of WA and UA between our model and the baseline methods.

The previous methods for SER have been diverse, each focusing on feature extraction and processing to achieve optimal performance. However, these methods have inherent limitations. Chen, Li et al. (2023), Chen, Lin et al. (2023), and Ye et al. (2023) have made significant contributions to speech feature extraction and fusion, enhancing the complementarity between different features. Nevertheless, this approach has constraints in extracting deeper emotional information from speech signals. Our method leverages pre-trained models for feature extraction, enabling the extraction of more emotion-related information. Zou et al. (2022) extracted complementary acoustic information using Wav2Vec2.0 and a common attention mechanism. However, the emotional features extracted were not fully utilized in downstream tasks.

In recent approaches utilizing MTL for SER, Cai et al. (2021) initially employed a pre-trained model to extract deep speech features, followed by further processing within an MTL framework. However, using only an ASR auxiliary task failed to fully exploit the emotional information embedded in the features. MMER, despite leveraging multimodal features from both speech and text, increased the preprocessing requirements for speech signals. The auxiliary tasks predominantly utilized text and speech, lacking complementary information from other relevant tasks. Li et al. (2019) extracted features from speech spectrograms, employing a self-attention mechanism to focus on prominent emotional periods within speech utterances. They then shared useful information

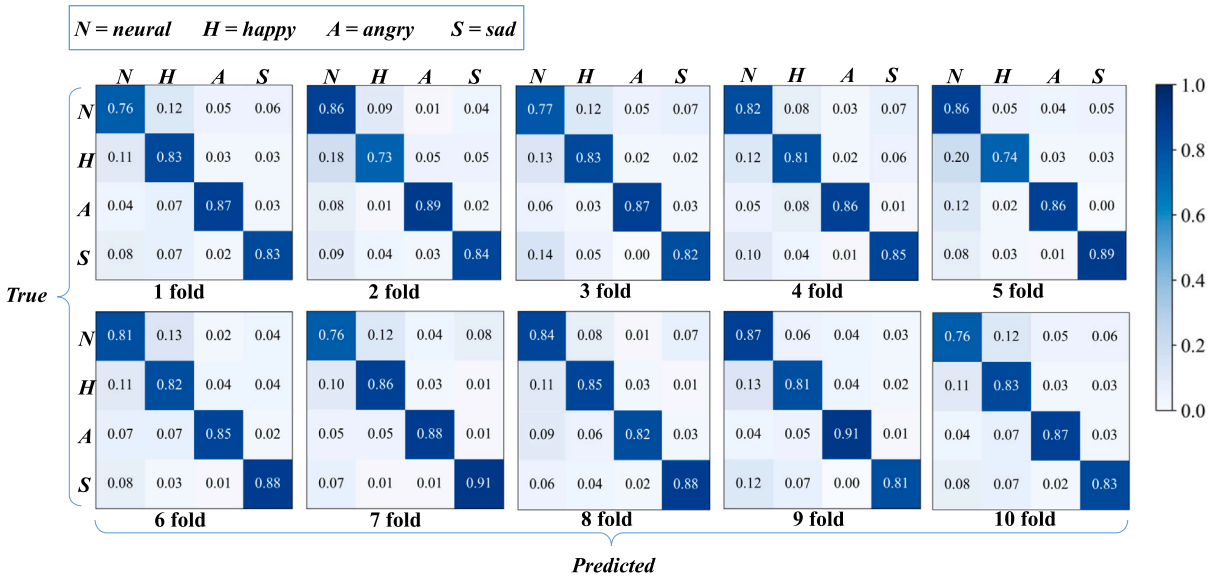


Fig. 5. 10-fold confusion matrix.

Table 4
Experimental results of comparison with the baseline method.

Method	UA (%)	WA (%)	FLOPs(G)	Inference time (s)	*p*-value
SVM (Kurpukdee et al., 2017)	–	58.40 ± 0.40	0.02	0.02	4.52E−13
STRL-SER (Chen, Lin et al., 2023)	79.32 ± 0.50	81.60 ± 0.85	12.00	0.34	1.78E−03
AMSNet (Chen, Li et al., 2023)	70.51 ± 0.65	69.22 ± 0.70	10.75	0.27	7.27E−11
Wav2vec2.0+MTL (Cai et al., 2021)	–	78.15 ± 0.60	17.31	0.55	4.59E−07
TIM-Net (Ye et al., 2023)	72.50 ± 0.80	71.65 ± 0.78	71.57	15.50	3.89E−10
MFCC+Spectrogram+wav2vec2.0 (Zou et al., 2022)	72.95 ± 0.70	71.64 ± 0.65	45.09	0.35	3.86E−10
Self-attention+MTL (Li et al., 2019)	80.60 ± 0.60	80.82 ± 0.55	26.10	0.96	1.46E−04
MMER (Ghosh et al., 2023)	76.20 ± 0.65	78.90 ± 0.70	138.77	1.60	1.70E−06
Proposed method	**82.19 ± 0.75**	**82.63 ± 0.77**	20.80	0.59	–

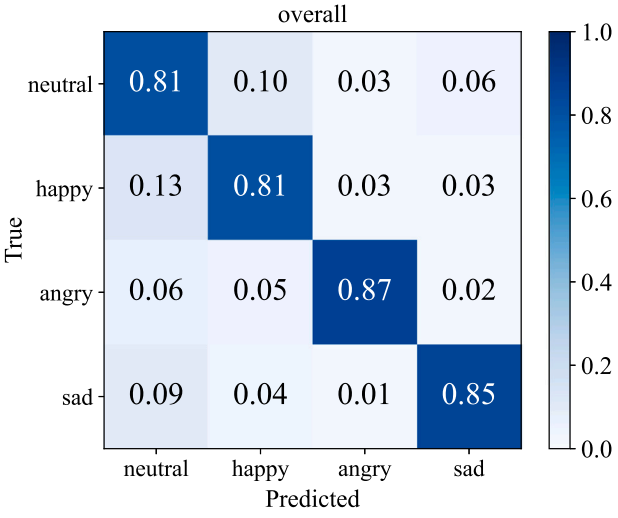


Fig. 6. Overall confusion matrix.

between gender classification and emotion classification tasks through MTL. Nevertheless, the features extracted in this manner carried fewer deep representations and did not fully capitalize on the advantages of MTL. Furthermore, using only gender recognition as an auxiliary task for SER had its limitations in enhancing SER performance. Our proposed MTLSER model addresses the shortcomings of baseline methods. Firstly, we utilized the pre-trained wav2vec2.0 model for feature extraction, capturing deep speech features and obtaining initial emotional representations rich in emotional information. Secondly, we selected three related tasks for MTL in the downstream task setup, allowing these tasks to influence each other during training, thereby extracting more complementary emotional information from speech signals. Performance-wise, our proposed MTLSER model exhibited outstanding results, outperforming the best baseline method by Li et al. (2019) by 1.59% in UA and Chen, Lin et al. (2023) by 1.03% in WA. Additionally, a two-tailed Welch's t-test with a significance threshold of $\alpha = 0.05$ is typically used to determine the statistical significance of the model's superiority (Nasiri & Ebadzadeh, 2023). This test evaluates whether there is a significant difference between two independent samples by calculating their means and variances. We applied the two-tailed Welch's t-test to the 10-fold cross-validation results of our proposed model and all baseline methods. As shown in Table 4, The *p*-values obtained from the experiments were all below 0.05, leading to the rejection of the null hypothesis. This indicates that our experimental results were not due to random chance and are statistically significant. However, this test has certain limitations: even if the *p*-value indicates statistical significance, it does not reflect the magnitude or practical importance of the difference. Therefore, effect size should be considered to provide a more comprehensive interpretation.

5.5.2. Comparison of computational complexity and inference time

We use FLOPs (Floating Point Operations) and inference time as metrics to measure the computational complexity and inference efficiency of our model. There is a certain correlation between FLOPs and inference time. Due to the inclusion of the pre-trained acoustic model, the complexity of our proposed model has increased. Consequently, both FLOPs and inference time are higher than those of traditional machine learning methods such as SVM and some deep

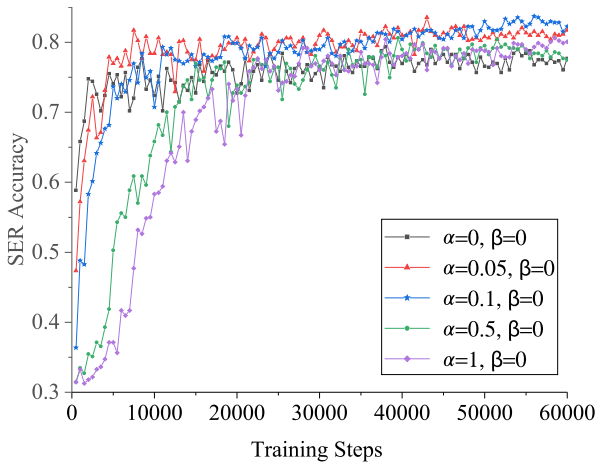


Fig. 7. acc of SER when $\beta = 0$, $\alpha = [0, 0.05, 0.1, 0.5, 1]$.

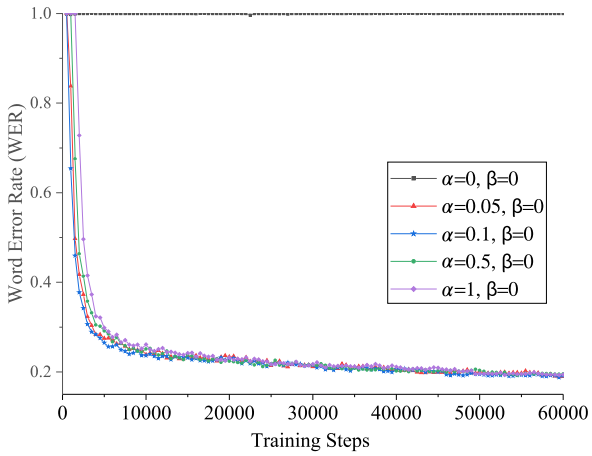


Fig. 8. wer of ASR when $\beta = 0$, $\alpha = [0, 0.05, 0.1, 0.5, 1]$.

learning methods like AMSNet and STRL-SER. However, in terms of FLOPs, our model remains significantly lower than the most complex models such as Multimodal+MTL (138.77 G) and TIM-Net (71.57 G), and our inference time is substantially lower than TIM-Net (15.50 s) and Multimodal+MTL (1.60 s).

Summary: Our proposed method exhibits an excellent balance between accuracy and computational complexity. It achieves the highest accuracy metrics (UA and WA) while maintaining reasonable FLOPs and inference time requirements. This balance indicates that our method is not only effective but also efficient, making it suitable for practical applications where both performance and computational cost are critical considerations.

5.6. Ablation study

The hyperparameter value denotes the weighting of the auxiliary task's loss function within the overall loss function. In our pursuit of the optimal hyperparameter combination, we conducted the following experiment: by fixing one hyperparameter (α or β) at 0 and varying the other to observe changes in the efficacy of emotion recognition tasks, akin to conducting ablation experiments for SER+ASR and SER+SR. Concurrently, setting $\alpha = \beta = 0$ establishes a control group, transforming the model into a single-task SER model where the auxiliary task remains uninvolved in training and exerts no influence. Through the outcomes of the ablation experiment, we gain insights into how specific auxiliary tasks impact the efficacy of the primary task.

5.6.1. The impact of ASR task

In the MTL framework where SER is the primary task and ASR is the auxiliary task, previous studies have demonstrated that ASR can enhance the prediction accuracy of SER (Cai et al., 2021). In our experiment, we set $\beta = 0$ and varied the value of α as [0, 0.05, 0.1, 0.5, 1] to simulate this condition. Since the weight of the SR task is zero, it does not affect the SER and ASR tasks. The accuracy of SER and ASR was recorded, and the statistical results are shown in Figs. 7 and 8 .

When $\beta = 0$ and $\alpha \neq 0$, the model becomes an MTL model with SER as the primary task and ASR as the auxiliary task. The value of α represents the weight of the ASR task's loss function in the total loss function, while $\alpha = 0$ and $\beta = 0$ serve as the control group. As observed, the WER of the control group did not improve during training, indicating that the ASR task was not trained. As α increased from 0 to 0.1, the accuracy of SER improved significantly compared to the control group, and the ASR task also converged faster. However, when α further increased to 0.5 and 1, the accuracy of SER decreased and showed no significant difference from the control group. Notably,

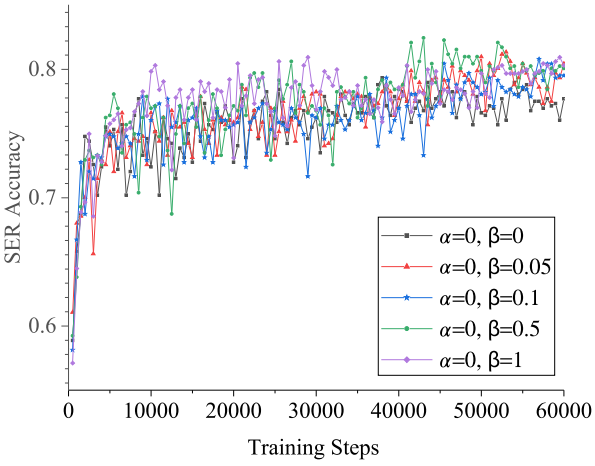


Fig. 9. acc of SER when $\alpha = 0$, $\beta = [0, 0.05, 0.1, 0.5, 1]$.

as α increased from 0.5 to 1, the convergence rate of both SER and ASR slowed down, with similar trends observed for both tasks. This indicates that increasing the value of α does not always enhance the performance of SER. An excessively large α value will reduce the effectiveness of both SER and ASR. There is an optimal α value that maximizes the accuracy of SER.

In conclusion, when ASR is the only auxiliary task, the performance of the primary SER task improves. The effect of ASR on SER varies with changes in the hyperparameter α. In this ablation experiment, both ASR and SER achieved the best performance when $\alpha = 0.1$.

5.6.2. The impact of SR task

The identity information of the speaker can also influence the determination of speech emotion. This is because certain speakers may consistently exhibit certain emotions in their speech, allowing us to predict the likely emotion based on the speaker's identity. Therefore, SR is a related task to SER. In this subsection, we explore how the SR task affects SER. We set $\alpha = 0$ and varied the value of β as [0, 0.05, 0.1, 0.5, 1]. The accuracies of SER and SR are shown in Figs. 9 and 10.

When $\alpha = 0$ and $\beta \neq 0$, the model becomes an MTL model with SER as the primary task and SR as the auxiliary task, where β represents the weight of the SR task's loss function in the total loss function. The ASR task is not involved in training. $\alpha = 0$ and $\beta = 0$ serve as the control group. As β varied, it can be seen from the endpoints

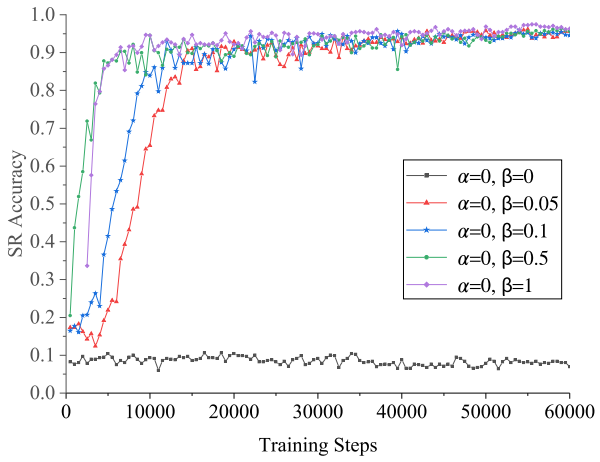


Fig. 10. acc of SR when $\alpha = 0$, $\beta = [0, 0.05, 0.1, 0.5, 1]$.

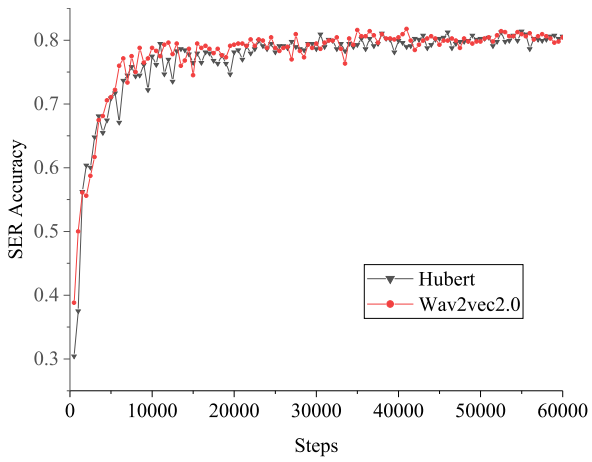


Fig. 11. Comparison of SER accuracy between Wav2vec2.0 and Hubert.

of the experimental group's lines in Fig. 9 that the accuracy of SER in all experimental groups was higher than that of the control group, indicating that the inclusion of the SR task enhanced the SER task's performance. Specifically, the experimental group with $\beta = 0.5$ showed the most significant improvement in SER, while the differences among the other experimental groups were minor. When β increased from 0.5 to 1, the performance of SER declined. In the SR task, the group with $\beta = 1$ achieved the best performance after task convergence, and the convergence rate of the SR task increased with β. However, as β approached 1, the rate of increase in convergence slowed down. This indicates that there is also an optimal β value for the SR task that maximizes its enhancement effect on SER.

Therefore, it can be observed that when training the SR and SER tasks simultaneously, the former positively influences the latter. Overall, $\beta = 0.5$ is the optimal setting.

5.6.3. Selection of pre-trained acoustic models

Firstly, as observed from the baseline methods, simple features like MFCC and spectrograms do not perform well. Using pre-trained acoustic models as feature extractors is a more effective approach. In our proposed model, the selection of a pre-trained acoustic model is crucial, and choosing an appropriate one is of utmost importance. Secondly, among the current acoustic models, wav2vec2.0 has demonstrated outstanding performance across various tasks. To demonstrate that wav2vec2.0 is also suitable for our proposed method, we conducted a comparative experiment with the pre-trained model Hubert. The two experimental groups differ only in the pre-trained model used, with all other settings remaining the same. The experimental results are shown in Figs. 11–13.

The experiment results indicate that wav2vec2.0 and Hubert exhibit similar recognition accuracy in SER, ASR, and SR tasks. However, wav2vec2.0 converges faster and has a lower training loss compared to Hubert. In the future, the pre-trained acoustic model in our proposed MTLSER framework can be replaced with other models that demonstrate better performance to achieve even greater results in SER tasks.

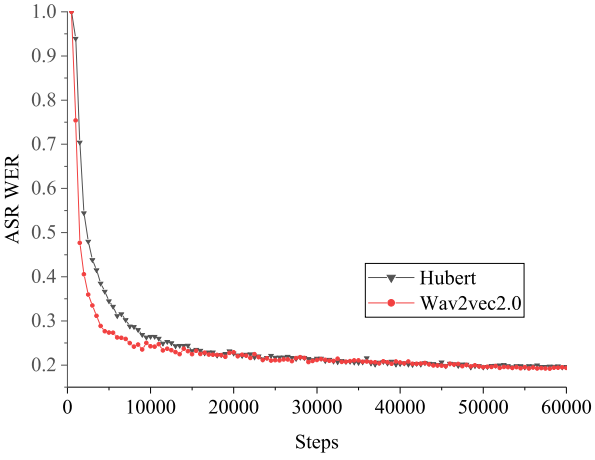


Fig. 12. Comparison of WER between Wav2vec2.0 and Hubert.

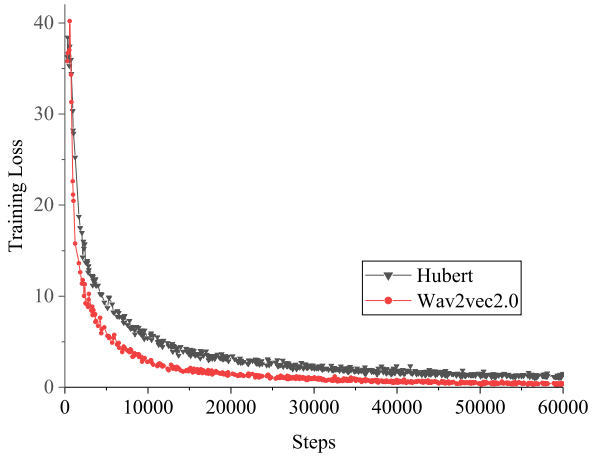
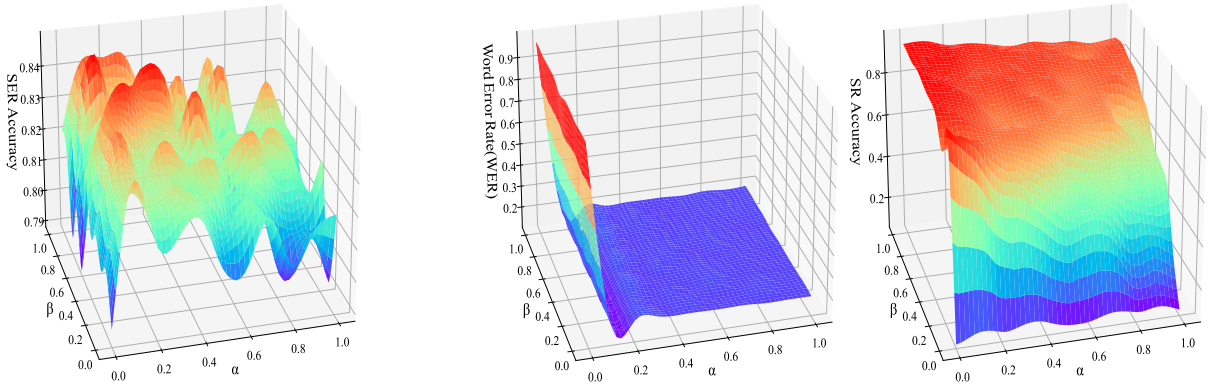


Fig. 13. Comparison of training loss between Wav2vec2.0 and Hubert.

5.7. Performance in noisy environments

In real-world applications, background noise can significantly impact the performance of SER systems. To evaluate the robustness of our model under various noise conditions, we added babble noise samples from the Noisex-92 database (Varga & Steeneken, 1993) to the audio in our dataset, simulating real-world audio input. We conducted experiments at five different signal-to-noise ratio (SNR) levels. The experimental parameters used are the same as those summarized in Table 3. Table 5 summarizes the performance metrics for each SNR level.

The results indicate that as the SNR increases, the model's performance improves across all metrics. At 0 dB, the model achieved a WA of 64.3%, a UA of 66.2%, and an F1 score of 56.1%. As the SNR increased



(a) Sensitivity analysis of SER. (b) Sensitivity analysis of ASR. (c) Sensitivity analysis of SR.

Fig. 14. Grid search results for three tasks.

Table 5
Performance metrics across different SNR levels.

SNR	WA (%)	UA (%)	F1_score (%)
0 dB	64.34 ± 0.83	66.22 ± 0.82	56.17 ± 0.83
5 dB	71.81 ± 0.85	73.45 ± 0.76	68.53 ± 0.78
10 dB	76.33 ± 0.78	77.73 ± 0.73	75.28 ± 0.75
15 dB	79.26 ± 0.82	80.31 ± 0.76	79.34 ± 0.77
20 dB	81.42 ± 0.80	81.93 ± 0.79	81.43 ± 0.83

to 20 dB, these metrics rose to 81.4%, 81.9%, and 81.4%, respectively. This trend suggests that the model is more effective in environments with less noise interference, achieving higher accuracy and consistency (as reflected by the reduced standard deviation) under cleaner audio conditions.

Overall, the model demonstrated robustness in noisy environments, with its performance significantly improving as the SNR increased. This underscores the importance of noise reduction techniques in real-world applications to ensure the optimal performance of SER systems.

5.8. Sensitivity analysis

Through the above experiments, we found that when SR and ASR are used as auxiliary tasks, they will both have an impact on the main task of emotion recognition. Under different choices of hyperparameters, when $\alpha = 0.1$, ASR has the best promoting effect on the main task, and the performance of the auxiliary task is also quite outstanding, while the SR task is in the case of $\beta = 0.5$. However, when the value of the hyperparameter increases to 1, the effect of the main task of SER is reduced. Therefore, we believe that for different auxiliary tasks, there exists an optimal choice for the hyperparameters within the range of 0 to 1, which maximizes the performance of SER.

In subsequent experiments, we jointly trained the aforementioned two auxiliary tasks with the main task to mutually influence and enhance each other. This approach aimed to extract more emotional information from speech signals to identify the optimal hyperparameter combinations for SER. Consequently, our model evolved into a multitask training model encompassing SER, ASR, and SR. To determine the optimal hyperparameter combinations, we conducted sensitivity analyses by performing grid searches on α and β. Results from ablation experiments indicate that both ASR and SR tasks significantly impact the recognition performance of SER, with different hyperparameter values showing considerable variation in SER's final effectiveness. The optimal hyperparameters for the SER+ASR and SER+SR frameworks were found to be $\alpha = 0.1$ and $\beta = 0.5$, respectively. However, as the values of α and β increase, the performance of the main SER task under the multi-task framework tends to decrease. When α and β increase to 1, the SER performance in the multi-task framework becomes nearly equivalent to the control group, i.e., the single-task scenario. Therefore, in our grid search experiments, we defined the search ranges for both hyperparameters α and β from 0 to 1, with a step size of 0.1, resulting in a total of 121 experiments. After completing the grid search, we plotted the search results as a surface plot, as shown in Fig. 14.

For the SER task, Fig. 14(a) demonstrates that the selection of hyperparameters α and β significantly impacts recognition performance. The surface plot reveals distinct peaks and valleys. Unlike the ablation experiments, we simultaneously trained SER with two auxiliary tasks – ASR and SR – resulting in mutual influence among the three tasks. Consequently, the optimal hyperparameter choices differ from previous experiments. Notably, we observed that smaller α values led to better SER performance, while increasing α caused performance deterioration. Conversely, β exhibited the most favorable impact on SER when approaching 1, indicating different sensitivity levels of the SER task to the two auxiliary tasks. Specifically, SER shows higher sensitivity with smaller α and larger β values. The optimal hyperparameter combination, $\alpha = 0.1$ and $\beta = 0.8$, corresponding to the peak in the surface plot, yields the best SER performance.

Additionally, Figs. 14(b) and 14(c) illustrate the grid search results for the ASR and SR tasks, respectively. Unlike the SER task, these plots exhibit no prominent peaks or valleys, suggesting that the performance of ASR and SR is primarily governed by their own hyperparameters and is minimally influenced by the other task. ASR achieves optimal performance when $\alpha = 0.1$, while SR reaches its peak at $\beta = 0.7$. This similarity in hyperparameter settings with those yielding the best SER performance suggests that, in this multi-task setup, achieving optimal SER performance also allows the auxiliary tasks to independently approach their best outcomes.

5.9. Discussion

This study presents a SER model based on an MTL framework. Our model achieves outstanding performance on the publicly available IEMOCAP dataset. We conducted ablation experiments to investigate the impact of using ASR and SR as auxiliary tasks on the SER task. The results show that both auxiliary tasks enhance the accuracy of the primary SER task, validating our hypothesis presented in Section 3.2. ASR captures textual information that can be associated with specific emotional content, while different speakers may exhibit preferences for certain emotions when expressing themselves vocally. Therefore, both ASR and SR are closely related to the SER task. Joint training of these tasks within a MTL framework enables mutual enhancement, facilitating the extraction of richer emotional information from speech. Previous studies by Cai et al. (2021) demonstrated that incorporating ASR as an auxiliary task improves SER performance, indicating that textual information carries emotional content. Zhao, Li, and Zhang

(2023) pointed out that the speaker's emotional continuity influences the expression of current utterance emotions. Zhang et al. (2019) modeled speaker-sensitive dependencies using graph neural networks, reflecting that different speakers may exhibit varied emotional responses due to personality differences. These studies align with our findings. Additionally, we observed that there is a threshold for the hyperparameters controlling the weights of the two auxiliary tasks that optimally enhances the primary SER task. In the sensitivity analysis, we used grid search to identify the hyperparameter combination that yields the best SER performance. Finally, we evaluated our proposed model under different levels of SNR, and the results indicate that our model exhibits robustness in noisy environments.

6. Conclusions

This paper presents an MTL framework designed to enhance SER by integrating ASR and SR as auxiliary tasks. These interrelated tasks contribute to improving SER accuracy within the framework. (1) A pre-trained model is utilized to extract shared features, which are subsequently processed for multiple tasks. The model is refined using an MTL strategy, integrating the loss functions of all tasks. (2) Specifically, the auxiliary task loss functions are weighted and linearly combined with the main task's loss function to construct a comprehensive final loss function for backpropagation and model training.

Additionally, we conducted several experiments to evaluate the performance of the proposed method: (1) ablation studies were conducted to assess the impact of auxiliary tasks on SER. (2) The results demonstrate that multi-task training outperforms traditional single-task frameworks, highlighting the critical influence of task weighting within our MTL framework. (3) Through extensive hyperparameter tuning via grid search, we identified the optimal combination that maximizes recognition performance. (4) We validated the model's performance in noisy environments, demonstrating its robustness.

The MTL approach, by extracting richer feature sets compared to single-task strategies, significantly advances the field of SER. However, the proposed method in this paper has certain limitations: (1) Our model was trained and validated on only one public dataset, necessitating experiments on additional datasets to verify the model's broader effectiveness. (2) The performance of our model is significantly affected by noise, requiring further optimization to enhance its robustness in noisy environments. In future work, we will focus on the following research directions: (1) Testing the proposed model on more datasets to demonstrate its broad applicability and robustness. (2) Optimizing the model architecture to improve its performance in noisy conditions, thereby enhancing its robustness in such environments. (3) Exploring multi-modal experiments to enrich the feature space and further boost performance. (4) Assessing the feasibility of the model in real-time applications to address practical deployment challenges. (5) Future research may explore alternative feature extraction techniques and additional related tasks for inclusion in multi-task training frameworks. (6) Furthermore, enhancing the methodology for processing shared features in downstream tasks could potentially yield superior recognition outcomes.

Abbreviations

The following abbreviations are used in this manuscript:

ACC	Accuracy
ASR	Automatic Speech Recognition
CNN	Convolutional Neural Network
CTC	Connectionist Temporal Classification
DBN	Deep Belief Networks
FLOPs	Floating Point Operations Per Second
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
LPC	Linear Predictive Coefficients
LSTM	Long Short-Term Memory networks
MFCC	Mel-frequency cepstral coefficients
MLP	Multi-Layer Perceptron
MTL	Multi-Task Learning
RF	Random Forests
RNN	Recurrent Neural Networks
SER	Speech Emotion Recognition
SNR	Signal-to-noise ratios
SR	Speaker Recognition
SVM	Support Vector Machines
UA	Unweigthed Accuracy
WA	Weighted Accuracy
WER	Word Error Rate

CRediT authorship contribution statement

Zengzhao Chen: Methodology, Investigation, Writing – original draft, Funding acquisition. **Chuan Liu:** Conceptualization, Methodology, Investigation, Writing – original draft. **Zhifeng Wang:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing. **Chuanxu Zhao:** Investigation. **Mengting Lin:** Methodology, Investigation. **Qiuyu Zheng:** Investigation, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China (Nos. 62477022, 62177022, 62277026,), Self-determined Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE, China (No. CCNU24JC033) and Research Project of National Collaborative Innovation Experimental Base for Teacher Development of Central China Normal University, China (No. CCNUTEIII 2024-01).

Data availability

Data will be made available on request.

References

Alu, D., Zoltan, E., & Stoica, I. C. (2017). Voice based emotion recognition with convolutional neural networks for companion robots. *Science and Technology*, *20*(3), 222–240.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*(2), 256.

Atmaja, B. T., & Sasou, A. (2022). Evaluating self-supervised speech representations for speech emotion recognition. *IEEE Access*, *10*, 124396–124407.

Atmaja, B. T., Sasou, A., & Akagi, M. (2022). Speech emotion and naturalness recognitions with multitask and single-task learnings. *IEEE Access*, *10*, 72381–72387.

Atmaja, B. T., Zanjabila, & Sasou, A. (2022). Jointly predicting emotion, age, and country using pre-trained acoustic embedding. In *2022 10th international conference on affective computing and intelligent interaction workshops and demos* (pp. 1–6).

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *vol. 33, Advances in neural information processing systems* (pp. 12449–12460). Curran Associates, Inc..

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*, 335–359.

Cai, X., Yuan, J., Zheng, R., Huang, L., & Church, K. (2021). Speech emotion recognition with multi-task learning. In *Proc. interspeech 2021* (pp. 4508–4512).

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*, 41–75.

Chen, Z., Li, J., Liu, H., Wang, X., Wang, H., & Zheng, Q. (2023). Learning multi-scale features for speech emotion recognition with connection attention mechanism. *Expert Systems with Applications*, *214*, Article 118943.

Chen, Z., Lin, M., Wang, Z., Zheng, Q., & Liu, C. (2023). Spatio-temporal representation learning enhanced speech emotion recognition with multi-head attention mechanisms. *Knowledge-Based Systems*, *281*, Article 111077.

Chen, W., Xing, X., Chen, P., & Xu, X. (2023). Vesper: A compact and effective pretrained model for speech emotion recognition. arXiv preprint arXiv:2307.10757.

Cheng, X., & Duan, Q. (2012/08). Speech emotion recognition using Gaussian mixture model. In *Proceedings of the 2012 international conference on computer application and system modeling (ICCASM 2012)* (pp. 1222–1225). Atlantis Press.

Costantini, G., Parada-Cabaleiro, E., Casali, D., & Cesarini, V. (2022). The emotion probe: On the universality of cross-linguistic and cross-gender speech emotion recognition via machine learning. *Sensors*, *22*(7).

Fahad, M. S., 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*, 466–489.

Geng, Z., Li, J., Han, Y., & Zhang, Y. (2022). Novel target attention convolutional neural network for relation classification. *Information Sciences*, *597*, 24–37.

Ghosh, S., Tyagi, U., Ramaneswaran, S., Srivastava, H., & Manocha, D. (2023). MMER: Multimodal Multi-task Learning for Speech Emotion Recognition. In *Proc. iNTERSPEECH 2023* (pp. 1209–1213).

Huang, Z., Dong, M., Mao, Q., & Zhan, Y. (2014). Speech emotion recognition using CNN. In *Proceedings of the 22nd ACM international conference on multimedia* (pp. 801–804). New York, NY, USA: Association for Computing Machinery.

Ioannides, G., Owen, M., Fletcher, A., Rozgic, V., & Wang, C. (2023). Towards Paralinguistic-Only Speech Representations for End-to-End Speech Emotion Recognition. In *Proc. iNTERSPEECH 2023* (pp. 1853–1857).

Jahangir, R., Teh, Y. W., Hanif, F., & Mujtaba, G. (2021). Deep learning approaches for speech emotion recognition: State of the art and research challenges. *Multimedia Tools and Applications*, 1–68.

Jain, M., Narayan, S., Balaji, P., Bhowmick, A., Muthu, R. K., et al. (2020). Speech emotion recognition using support vector machine. arXiv preprint arXiv:2002.07590.

Jha, T., Kavya, R., Christopher, J., & Arunachalam, V. (2022). Machine learning techniques for speech emotion recognition using paralinguistic acoustic features. *International Journal of Speech Technology*, *25*(3), 707–725.

Khare, A., Parthasarathy, S., & Sundaram, S. (2020). Multi-modal embeddings using multi-task learning for emotion recognition. arXiv preprint arXiv:2009.05019.

Kim, N. K., Lee, J., Ha, H. K., Lee, G. W., Lee, J. H., & Kim, H. K. (2017). Speech emotion recognition based on multi-task learning using a convolutional neural network. In *2017 Asia-Pacific signal and information processing association annual summit and conference* APSIPA ASC, (pp. 704–707).

Koolagudi, S. G., Murthy, Y. S., & Bhaskar, S. P. (2018). Choice of a classifier, based on properties of a dataset: Case study-speech emotion recognition. *International Journal of Speech Technology*, *21*(1), 167–183.

Kurpukdee, N., Kasuriya, S., Chunwijitra, V., Wutiwiwatchai, C., & Lamsrichan, P. (2017). A study of support vector machines for emotional speech recognition. In *2017 8th international conference of information and communication technology for embedded systems* IC-ICTES, (pp. 1–6).

Kwon, O.-W., Chan, K., Hao, J., & Lee, T.-W. (2003). Emotion recognition by speech signals. In *Eighth European conference on speech communication and technology*.

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Epps, J., & Schuller, B. W. (2022). Multi-task semi-supervised adversarial autoencoding for speech emotion recognition. *IEEE Transactions on Affective Computing*, *13*(2), 992–1004.

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B. (2023). Survey of deep representation learning for speech emotion recognition. *IEEE Transactions on Affective Computing*, *14*(2), 1634–1654.

Le, D., & Provost, E. M. (2013). Emotion recognition from spontaneous speech using hidden Markov models with deep belief networks. In *2013 IEEE workshop on automatic speech recognition and understanding* (pp. 216–221).

wook Lee, S. (2023). Diverse feature mapping and fusion via multitask learning for multilingual speech emotion recognition. In *Proc. iNTERSPEECH 2023* (pp. 3944–3948).

Li, Y., Bell, P., & Lai, C. (2022). Fusing ASR outputs in joint training for speech emotion recognition. In *ICASSP 2022 - 2022 IEEE international conference on acoustics, speech and signal processing* (pp. 7362–7366).

Li, L., Zhao, Y., Jiang, D., Zhang, Y., Wang, F., Gonzalez, I., et al. (2013). Hybrid deep neural network–hidden Markov model (DNN-hmm) based speech emotion recognition. In *2013 humaine association conference on affective computing and intelligent interaction* (pp. 312–317).

Li, Y., Zhao, T., & Kawahara, T. (2019). Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In *Proc. interspeech 2019* (pp. 2803–2807).

Lieskovská, E., Jakubec, M., Jarina, R., & Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. *Electronics*, *10*(10).

Liu, J., Liu, Z., Wang, L., Guo, L., & Dang, J. (2020). Speech emotion recognition with local-global aware deep representation learning. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7174–7178).

Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., et al. (2023). Speech emotion recognition using machine learning — A systematic review. *Intelligent Systems with Applications*, *20*, Article 200266.

Maji, B., Swain, M., Guha, R., & Routray, A. (2023). Multimodal emotion recognition based on deep temporal features using cross-modal transformer and self-attention. In *ICASSP 2023 - 2023 IEEE international conference on acoustics, speech and signal processing* (pp. 1–5).

Mirsamadi, S., Barsoum, E., & Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE international conference on acoustics, speech and signal processing* (pp. 2227–2231).

Nasiri, H., & Ebadzadeh, M. M. (2023). Multi-step-ahead stock price prediction using recurrent fuzzy neural network and variational mode decomposition. *Applied Soft Computing*, *148*, Article 110867.

Noroozi, F., Sapiński, T., Kamińska, D., & Anbarjafari, G. (2017). Vocal-based emotion recognition using random forests and decision tree. *International Journal of Speech Technology*, *20*(2), 239–246.

Schuller, B., Rigoll, G., & Lang, M. (2003). Hidden Markov model-based speech emotion recognition. *vol. 2*, In *2003 IEEE international conference on acoustics, speech, and signal processing, 2003. proceedings* ICASSP '03, (pp. II–1).

Sharma, M. (2022). Multi-lingual multi-task speech emotion recognition using wav2vec 2.0. In *ICASSP 2022 - 2022 IEEE international conference on acoustics, speech and signal processing* (pp. 6907–6911).

Umamaheswari, J., & Akila, A. (2019). An enhanced human speech emotion recognition using hybrid of PRNN and KNN. In *2019 international conference on machine learning, big data, cloud and parallel computing* cOMITCon, (pp. 177–183).

Varga, A., & Steeneken, H. J. (1993). Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, *12*(3), 247–251.

Wang, Z., Yang, Y., Zeng, C., Kong, S., Feng, S., & Zhao, N. (2022). Shallow and deep feature fusion for digital audio tampering detection. *EURASIP Journal on Advances in Signal Processing*, *2022*(69), 1–20.

Wang, Z., Zhan, J., Zhang, G., Ouyang, D., & Guo, H. (2023). An end-to-end transfer learning framework of source recording device identification for audio sustainable security. *Sustainability*, *15*(14), 11272.

Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021). A comprehensive review of speech emotion recognition systems. *IEEE Access*, *9*, 47795–47814.

Wen, G., Fu, J., Dai, P., & Zhou, J. (2021). DTDE: A new cooperative multi-agent reinforcement learning framework. *The Innovation*, *2*(4).

Xie, Y., Liang, R., Liang, Z., Huang, C., Zou, C., & Schuller, B. (2019). Speech emotion classification using attention-based LSTM. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(11), 1675–1685.

Ye, J., Wen, X.-C., Wei, Y., Xu, Y., Liu, K., & Shan, H. (2023). Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. In *ICASSP 2023 - 2023 IEEE international conference on acoustics, speech and signal processing* (pp. 1–5).

Yunxiang, L., & Kexin, Z. (2023). Design of efficient speech emotion recognition based on multi task learning. *IEEE Access*, *11*, 5528–5537.

Zeng, C., Feng, S., Wang, Z., Wan, X., Chen, Y., & Zhao, N. (2024). Spatio-temporal representation learning enhanced source cell-phone recognition from speech recordings. *Journal of Information Security and Applications*, *80*, Article 103672.

Zeng, C., Feng, S., Wang, Z., Zhao, Y., Li, K., & Wan, X. (2024). Audio source recording device recognition based on representation learning of sequential Gaussian mean matrix. *Forensic Science International: Digital Investigation*, *48*, Article 301676.

Zeng, C., Feng, S., Zhu, D., & Wang, Z. (2023). Source acquisition device identification from recorded audio based on spatiotemporal representation learning with multi-attention mechanisms. *Entropy*, *25*(4), 626.

Zeng, C., Kong, S., Wang, Z., Feng, S., Zhao, N., & Wang, J. (2024). Deletion and insertion tampering detection for speech authentication based on fluctuating super vector of electrical network frequency. *Speech Communication*, *158*, Article 103046.

Zhang, Z., Wu, B., & Schuller, B. (2019). Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In *ICASSP 2019 - 2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6705–6709).

Zhang, D., Wu, L., Sun, C., Li, S., Zhu, Q., & Zhou, G. (2019). Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *IJCAI* (pp. 5415–5421). Macao.

Zhang, Y., & Yang, Q. (2022). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, *34*(12), 5586–5609.

Zhao, H., Li, B., & Zhang, Z. (2023). Speaker-aware cross-modal fusion architecture for conversational emotion recognition. In *Proc. iNTERSPEECH 2023* (pp. 2718–2722).

Zheng, Q., Chen, Z., Liu, H., Lu, Y., Li, J., & Liu, T. (2023). MSRANet: Learning discriminative embeddings for speaker verification via channel and spatial attention mechanism in alterable scenarios. *Expert Systems with Applications*, *217*, Article 119511, URL https://www.sciencedirect.com/science/article/pii/S095741742300012X.

Zheng, Q., Chen, Z., Wang, Z., Liu, H., & Lin, M. (2024). MEConformer: Highly representative embedding extractor for speaker verification via incorporating selective convolution into deep speaker encoder. *Expert Systems with Applications*, *244*, Article 123004.

Zhou, Y., Sun, Y., Zhang, J., & Yan, Y. (2009). Speech emotion recognition using both spectral and prosodic features. In *2009 international conference on information engineering and computer science* (pp. 1–4).

Zhu, L., Chen, L., Zhao, D., Zhou, J., & Zhang, W. (2017). Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN. *Sensors*, *17*(7).

Zou, H., Si, Y., Chen, C., Rajan, D., & Chng, E. S. (2022). Speech emotion recognition with co-attention based multi-level acoustic information. In *ICASSP 2022 - 2022 IEEE international conference on acoustics, speech and signal processing* (pp. 7367–7371).