Computer Vision and Image Understanding 260 (2025) 104483

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Optimizing ambiguous speech emotion recognition through spatial–temporal parallel network with label correction strategy

Chenquan Gan [a,b], Daitao Zhou [b], Kexin Wang [b], Qingyi Zhu [a], Deepak Kumar Jain [c,d,*], Vitomir Štruc [e]

[a] *School of Cyber Security and Information Law, Chongqing University of Posts and Telecommunications, Chongqing 400065, China*
[b] *School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China*
[c] *Key Laboratory of Intelligent Control and Optimization for Industrial Equipment of Ministry of Education, Dalian University of Technology, Dalian 116024, China*
[d] *Symbiosis Institute of Technology, Symbiosis International University, Pune 412115, India*
[e] *Faculty of Electrical Engineering, University of Ljubljana, Ljubljana SI, 1000, Slovenia*

ARTICLE INFO

Keywords:
Speech emotion recognition
Ambiguous speech
Spatial–temporal parallel network
Label correction

ABSTRACT

Speech emotion recognition is of great significance for improving the human–computer interaction experience. However, traditional methods based on hard labels have difficulty dealing with the ambiguity of emotional expression. Existing studies alleviate this problem by redefining labels, but still rely on the subjective emotional expression of annotators and fail to consider the truly ambiguous speech samples without dominant labels fully. To solve the problems of insufficient expression of emotional labels and ignoring ambiguous undominantly labeled speech samples, we propose a label correction strategy that uses a model with exact sample knowledge to modify inappropriate labels for ambiguous speech samples, integrating model training with emotion cognition, and considering the ambiguity without dominant label samples. It is implemented on a spatial–temporal parallel network, which adopts a temporal pyramid pooling (TPP) to process the variable-length features of speech to improve the recognition efficiency of speech emotion. Through experiments, it has been shown that ambiguous speech after label correction has a more promoting effect on the recognition performance of speech emotions.

1. Introduction

As the core carrier of human–computer interaction, the emotional information contained in speech is promoting the development of a new generation of intelligent auxiliary systems (Jahangir et al., 2021). Speech emotion recognition technology based on deep learning enables machine systems to more accurately perceive and understand human emotional states by analyzing emotional features in speech signals (Leo et al., 2022a). This ability shows great value in various assistive applications. In the field of medical rehabilitation (Zhong et al., 2020), intelligent systems combined with visual emotion recognition technology (Del Coco et al., 2017; Lecciso et al., 2021) can help medical staff to assess patients' status more comprehensively; In the educational assisted scene (Zepf et al., 2020), the emotion-sensing robot can dynamically adjust teaching strategies according to the emotional changes of learners (Kang, 2025). In a smart home environment (Chatterjee et al., 2021), voice assistants equipped with emotion recognition can provide a more empathetic service experience. With the progress of multi-modal affective computing technology, intelligent auxiliary systems will achieve more natural and accurate human–computer affective interaction in the future (Gao et al., 2025).

Speech emotion Recognition (SER) is an important way to improve human–computer interaction. Traditional methods mainly rely on manual extraction of acoustic features (such as spectrum, rhythm, etc.) and combine with machine learning algorithms (such as SVM, GMM) for sentiment classification (Sharma et al., 2021a; Khurana et al., 2021). However, these methods have the problem of feature redundancy and are difficult to comprehensively represent the complex emotional information in speech (Khurana et al., 2021). In recent years, the rapid development of deep learning technology has provided new solutions for speech emotion recognition. Studies show that methods based on deep neural networks (such as CNN, RNN, and LSTM) can automatically learn the deep emotional features in speech, significantly improving the recognition performance (Sharma et al., 2021b). Especially in the research of multimodal emotion recognition, deep learning

[*] Corresponding author.
E-mail addresses: gcq2010cqu@163.com (C. Gan), S220101216@stu.cqupt.edu.cn (D. Zhou), S200101155@stu.cqupt.edu.cn (K. Wang), zhuqy@cqupt.edu.cn (Q. Zhu), dkj@ieee.org (D.K. Jain), vitomir.struc@fe.uni-lj.si (V. Štruc).

https://doi.org/10.1016/j.cviu.2025.104483
Received 22 April 2025; Received in revised form 19 July 2025; Accepted 18 August 2025
Available online 20 August 2025

models have demonstrated a powerful feature fusion ability (Calderon-Uribe et al., 2024). Although the research on emotion theory remains controversial (Sharma et al., 2022), the advancement of speech emotion recognition technology has provided significant support for the field of emotion computing. Therefore, using deep learning methods for speech emotion recognition has become an inevitable trend in current research.

However, most of the real-life speech emotions are characterized by ambiguity, for example, speech expressions with the emotion of sadness are interspersed with emotional expressions of anger and disappointment (Kim and Kim, 2018), and there is an overlapping part of anger and surprise at the embedding level (Kumar et al., 2021). Some literature tries to alleviate the fuzziness of speech from the perspective of hard labels, such as training based on hard labels (Hou et al., 2021), that is, a speech corresponds to a fixed real label, representing that the speech contains only one emotion. However, hard-label emotion determination does not effectively reflect the ambiguity of speech emotion. In addition, individual annotators may present different views of ambiguous emotions depending on their culture and personality, i.e., emotion perception is subjective (Lotfian and Busso, 2019). In combination with the above, using hard labels to express the true emotion of speech not only lacks the ability of mixed expression of emotion but also ignores the subjectivity of the annotator's perception of emotion.

Considering the above problems, approaches based on soft labels (Steidl et al., 2005) or multiple labels (Ando et al., 2018; Li et al., 2023) were proposed to capture the ambiguity and subjective nature of emotion based on label definition. Soft labels describe the ambiguity of emotions by the proportion of each emotion expressed by annotators, but the fixed emotional proportions do not necessarily represent the true proportions recognized by most people. While multi-label classification can be performed without any proportion limitation, merely estimating the presence or absence of emotions, it still relies on the emotional cognition assigned by some annotators and lacks a clear emotional bias.

As a result, some more effective ambiguous processing training methods have been developed, such as joint learning (Chou and Lee, 2019), meta-learning (Fujioka et al., 2020), emotion contour extraction (Mao et al., 2020), multi-classifier interaction (Zhou et al., 2022), etc., aimed at combining the model's knowledge to avoid the problems of relying on annotators labels in soft labels and multi-label methods. These methods only consider the speech samples with dominant labels (hard labels), that is, the samples that can get the emotional consensus of the majority of annotators, and do not take advantage of the speech samples without dominant labels in the dataset. However, the vagueness of speech emotion is mainly reflected in these non-master label samples. Because the no-master label sample is caused by the annotator's inability to reach a consensus on the emotional judgment of the speech, this indicates that the sample is emotionally ambiguous, resulting in human illegibility. Moreover, not every sentence of speech in the actual environment has the emotion that most people agree with Zhang et al. (2021a). Therefore, these methods, which do not use ambiguous speech samples without master labels, do not fully consider the essential problem of speech emotional ambiguity.

To summarize the above discussions, the previous work has the following problems: (1) Multi-label methods are completely dependent on annotators' emotional perception cannot represent everyone's opinions, and do not provide a clear emotional bias to the model. (2) While some improved training methods can avoid the problem of relying on annotator labels, they fail to address the problem of unlabeled speech samples, and the ambiguity of speech emotion remains unaddressed.

To address the above issues, this paper proposes a method of label correction utilizing the model's emotional cognition to handle the ambiguity in speech emotions. For problem (1), we aim to alleviate the excessive reliance on annotators in existing labeling methods by employing a model with clear emotional knowledge. This enables model training to rely proportionally on both partial annotators and the model's inherent emotional cognition. For problem (2), we use a two-stage training approach, where label correction is first performed on ownerless labeled ambiguous speech, followed by emotion training of the model. In addition, this paper builds a spatial–temporal network with a spatial domain assisted by a temporal domain to implement the label correction strategy. Through experimental comparisons with five state-of-the-art methods, it is verified that our proposed method is superior for the processing of speech emotion ambiguity.

The main contributions of we can be summarized as follows.

(1) An effective label correction strategy is proposed, which modifies multiple labels without emotional bias by the pre-training exact model to generate labels, so that model training does not completely rely on the emotional cognition of a few annotators, and the modified labels have clear emotional bias through a balance factor.
(2) This paper provides a reference method for exploring and processing ambiguous speech without master labels, which shows that samples without the affective consensus of most annotators are still helpful for the model to establish affective cognition.
(3) A spatial–temporal auxiliary network that utilizes spatial domain emotional information to assist in the temporal domain has been constructed, offering a new attempt at spatial–temporal information fusion.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 details the proposed ambiguous speech emotion recognition method. Sections 4 and 5 describe the experimental setup and results, respectively. Finally, Section 6 summarizes the work and gives an outlook.

2. Related work

This section reviews research on speech emotion recognition from common hard label methods to previous methods for dealing with speech ambiguity.

Hard labels are used in speech emotion recognition as the majority vote of a group of annotators to determine the emotion label for an entire utterance. Training models using hard labels is a common practice in this field. For instance, Hou et al. (2021) proposed a collective multi-view relational network that analyzes emotion in speech from multiple acoustic feature perspectives, considering the complexity of speech emotion. Fan et al. (2022) considered that differences in individual emotional expression would cause emotional confusion among individuals, so they proposed an individual standardized network to alleviate differences in individual emotional expression. Yin et al. (2021) proposed a progressive collaborative learning approach that divides speech samples into difficulty levels based on the model's output loss, aiming to mitigate the confusion caused by ambiguous emotions during initial training. Wang et al. (2024) combined large language models (LLMs) with effective spatial learning methods to address the limitations of the IEMOCAP dataset. Considering that spatial–temporal networks can effectively capture both spatial and temporal information (Quach et al., 2022), Gan et al. (2023) proposed a spatial–temporal parallel network. This network takes Fbank features from different granularities (with 26 and 40 filter banks used for temporal and spatial domain module inputs, respectively) and feeds them into separate time and spatial domain modules to extract emotional features from speech. They employ a multi-fusion mechanism to combine the temporal and spatial features of the speech. Specifically, the temporal, spatial, and concatenated features are classified separately by classifiers, and the average of the classification results is taken. This allows the network to handle speech data of varying lengths and effectively extract spatial–temporal features. Their spatial feature extraction does not separately extract frame-level feature patterns but only analyzes the spectral features of speech. Although the above works describe the ambiguity and complexity (Thimmaiah et al., 2024) of speech emotion from different perspectives, they rely on hard labels for training, assuming that each speech segment contains only one emotion. This approach ignores the possibility of emotion blending and the subjectivity of annotators' emotion perception, resulting in lower recognition accuracy of the models. Therefore, we do not adopt hard labels to solve the problem of emotional ambiguity.

The soft label and multiple label training methods take into account all the emotions provided by the annotator, representing more than one emotion contained in each speech through label definition. The groundbreaking work of Steidl et al. (2005) is to represent the emotion labels of speech as probability distributions and named them soft labels, aiming to imitate human annotators' ambiguity and confusion regarding emotions. Zhang et al. (2019) proposed a Deep Metric Learning (DML) method supporting soft labels, and used f-Similarity Preservation Loss to guide the model to learn the pairwise label similarity in the feature embeddings. Furthermore, Ando et al. (2019) improved the soft label method through a multi-label presence model, first applying multiple labels to estimate the presence and absence of each emotion, and then using a soft label training model to determine the final emotion. Li et al. (2023) proposed an inter-class difference loss function using multiple labels as real labels, enabling the network to learn the emotion distribution in speech automatically. It is worth noting that multi-labeling sets all the emotions assigned by the annotator to exist at the same level, which does not allow the model to have a definite affective bias, and network learning is still limited to the fixed affective cognition of some annotators. Therefore, these models are still influenced by the subjective emotional cognition of the annotators.

To avoid the above problems associated with soft and multiple labels, more effective ambiguous processing methods were subsequently developed. Chou and Lee (2019) proposed a joint learning approach to incorporate annotator features into label distribution, considering both label uncertainty and individual characteristics of annotators. Based on the idea of meta-learning, Fujioka et al. (2020) fused two trainable parameters into model training for dynamic label correction and sample contribution weight estimation respectively, to repair inaccurate labels and ignore ambiguous samples. Mao et al. (2020) argued that static soft labels cannot capture the dynamic variations of emotions, the method of refined emotion contour for each utterance is applied to refine the emotion contours further, achieving the goal of dynamically generating soft labels. Similarly, Zhang et al. (2021b) utilized emotion contour refinement methods to update the soft labels obtained from the facial expression modality, and utilized the soft labels as a weak label for speech data, thereby transferring facial emotion knowledge into the speech modality. In addition, Zhou et al. (2022), inspired by the optimal interaction theory, built multiple classifiers to imitate the emotional cognitive interaction process of human beings, so as to avoid the inappropriate representation of emotions by the single label and multiple labels. Unfortunately, these methods are only good at handling samples with a dominant label and exclude samples without a dominant label. However, it is often the speech sample without a dominant label that shows the vagueness of speech emotion. Therefore, making full use of these samples is crucial for building a more robust speech emotion recognition system that is close to the true emotion distribution.

Inspired by the above work, we propose a speech ambiguous processing method that utilizes the emotional cognition of the model for label correction. This method successfully extracts favorable emotional information from data without a dominant label and better addresses the problem of speech ambiguity. In the next section, we will provide a detailed introduction to the proposed method for speech emotion recognition.

3. The proposed method

In this section, we first mathematically formulate the problem of emotional ambiguity; subsequently, we outline the thought process of this paper to solve the problem of emotional ambiguity; furthermore, we detail the method of correcting emotionally ambiguous speech labels; and finally, we explain the construction of a network for emotion learning. Next, we will elaborate one by one.

3.1. The ambiguous problem of emotion

In a supervised classification study of K-class speech emotion, the hard label is a commonly used expression for the real label in emotion datasets, which is defined as the category that receives the majority of votes from annotators as the dominant label, and the remaining categories are labeled as 0. It is defined as follows:

$$\begin{aligned}&\mathbf{y}^i_{hard} = (y^i_1, y^i_2, \ldots, y^i_K \mid \textstyle\sum_{j=1}^{K} y^i_j = 1, y^i_j \in \{0,1\}),\\&1 \le i \le N, 1 \le j \le K,\end{aligned} \tag{1}$$

where $y^i_j \in \{0,1\}$ represents whether there is a jth emotion category in the ith sample in the emotion dataset, N represents the total number of samples in the speech dataset, and K represents the number of emotion categories classified. It can be seen that the hard label $\mathbf{y}^i_{hard}$ indicates that a sample only contains one emotion. However, in a real environment, people's emotional expression is always intertwined with multiple emotions, not just one emotion. Moreover, the emotional cognition formed by a small number of annotators does not represent the emotional cognition of the public. Therefore, the emotion labels of most speech samples in the dataset are ambiguous. To alleviate these problems, soft label and multi-label methods have emerged. Soft labels describe the ambiguity of the speech sample through the ratio of various emotional attitudes expressed by the annotator. Its definition is as follows:

$$\begin{aligned}&\mathbf{y}^i_{soft} = (y^i_1, y^i_2, \ldots, y^i_K \mid \textstyle\sum_{j=1}^{K} y^i_j = 1, y^i_j \in [0,1]),\\&1 \le i \le N, 1 \le j \le K,\end{aligned} \tag{2}$$

where $y^i_j \in [0,1]$ represents the proportion of the ith sample in the emotion dataset that contains the jth emotion category. This proportion is pre-calculated based on statistics for all the emotion categories provided by a group of annotators, which increases the workload of emotion prediction. However, Multiple labels are not limited by scale and only represent the presence or absence of all emotion categories annotated by the annotator. The definition is as follows:

$$\begin{aligned}&\mathbf{y}^i_{multi} = (y^i_1, y^i_2, \ldots, y^i_K | y^i_j \in \{0,1\}),\\&1 \le i \le N, 1 \le j \le K,\end{aligned} \tag{3}$$

where $y^i_j \in \{0,1\}$ also indicates whether there is a jth emotion category in the ith sample in the emotion dataset, but is not limited to a particular category. It can be seen that multi-label $\mathbf{y}^i_{multi}$ can express multiple emotions in the sample without pre-calculating the emotion proportion, making it more suitable for practical applications. However, this multi-label approach still cannot represent the emotion cognition of most people, as its establishment still relies solely on a small number of annotators, and the multi-label approach only represents certain emotions in the speech and does not represent the dominant emotion bias in the speech. Therefore, we propose to use the model with exact sample knowledge to correct the labels of emotionally ambiguous samples, to achieve the effect that the corrected multi-labels do not completely depend on the emotional cognition of the labeling so that the model has an emotional bias.

3.2. The basic idea of solving ambiguous problem

To alleviate the issues of multi-label classification being overly dependent on annotators' emotional annotations and lacking emotion bias, this paper proposes a multi-label correction strategy specifically for ambiguous speech samples, based on a multi-label approach. Additionally, a spatial–temporal network designed for variable-length speech is utilized for experimental validation. Fig. 1 illustrates the

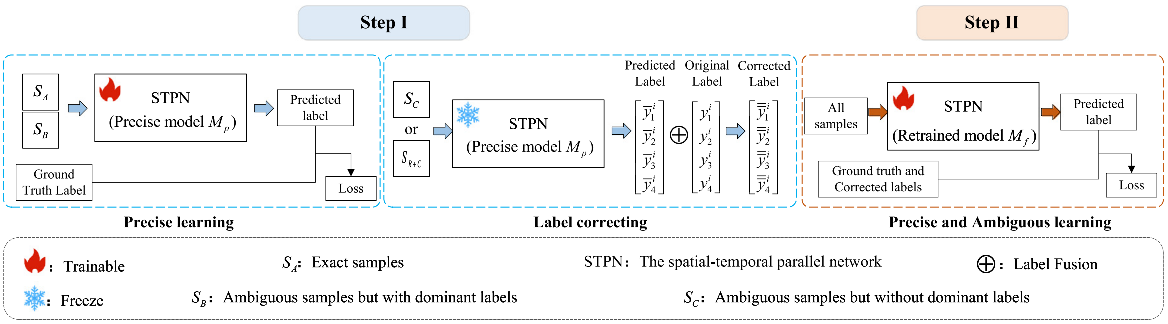


Fig. 1. Flowchart of label correction strategy.

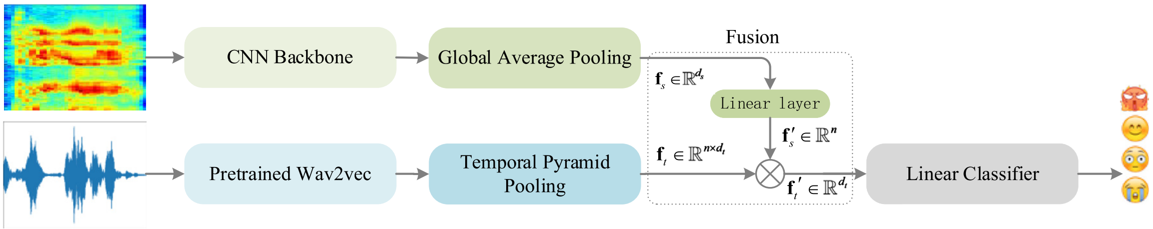

CNN Backbone
Global Average Pooling
Fusion
$\mathbf{f}_s \in \mathbb{R}^{d_s}$
Linear layer
$\mathbf{f}_s' \in \mathbb{R}^{n}$
Pretrained Wav2vec
Temporal Pyramid Pooling
$\mathbf{f}_t \in \mathbb{R}^{n \times d_t}$
$\mathbf{f}_t' \in \mathbb{R}^{d_t}$
Linear Classifier

Fig. 2. The spatial–temporal parallel network.

flowchart of the proposed method, while Fig. 2 showcases the spatial–temporal network model.

The correction strategy is outlined as follows. Firstly, according to the multiple labels for each sample, all the samples in the dataset can be categorized into three types: exact samples S_A, inaccurate samples S_B (with ambiguity but with a dominant label), and ambiguous samples S_C. Next, pretraining a model using an exact sample S_A or a sample with a dominant label ($S_A + S_B$) to generate a model M_p with exact sample knowledge, and input the remaining ambiguous samples, assumed to be $\mathbf{x}_i$, into the exact model M_p to obtain the generated label $\hat{\mathbf{y}}(\mathbf{x}^i|M_p)$ for the ambiguous samples. Then, design a label correction strategy $\boldsymbol{\Phi}(\cdot)$ to modify the original multi-label $\mathbf{y}^i_{multi}$ of the ambiguous sample, obtaining a corrected label $\boldsymbol{\Phi}(\mathbf{y}^i_{multi}, \hat{\mathbf{y}}(\mathbf{x}^i|M_p))$ of the ambiguous sample. Finally, the ambiguous sample after label correction is combined with the pre-training sample to retrain the final emotion classification model, to ensure that the model does not completely rely on the affective cognition of some annotators, alleviate the lack of emotional bias of multi-label, and realize the maximum utilization of ambiguous data.

The spatial–temporal parallel network used is summarized below. Firstly, input the original one-dimensional variable-length speech into the Wav2vec model to generate temporal emotion features that include temporal context information, and enable the model to train speech with different lengths through the TPP module for subsequent feature fusion. Secondly, to compensate for the lack of spatial information in one-dimensional speech, CNN is used as the backbone of the spatial network to extract spatial emotion features of the variable-length speech spectrum, and these features are processed into a fixed size through global pooling. Finally, using the idea of collaborative attention, spatial emotion features are fused with temporal emotion features and spatial information is used to assist temporal information in achieving emotion classification.

Next, we will specifically explain the label correction strategy proposed and the spatial–temporal feature extraction model in this paper.

3.3. Label correction

The proposed label correction strategy requires two stages of training. The first stage is to use exact samples to pre-train the exact model and correct the ambiguous samples; the second stage is to combine the ambiguous samples after label correction with the exact samples to reconstruct the training and complete the emotion classification. For this purpose, we classify all data samples in the dataset into three types based on their multiple labels: S_A (exact sample), S_B (ambiguous but with dominant labels), and S_C (ambiguous but without dominant labels). For example, in a four-class classification task, if eight annotators vote on a speech sample as [8,0,0,0], it is classified as a class S_A sample. If the voting result is [7,1,0,0], it belongs to class S_B. When the votes are evenly distributed as [2,2,2,2], the sample is categorized as class S_C. In practical implementation, the categorization of samples into S_A, S_B, and S_C was performed based on the consistency between the single-label annotation (obtained through majority voting) and the corresponding multi-label set. Specifically, a sample was assigned to S_A if the single-label and the multi-label set were fully consistent, indicating unanimous agreement among annotators. If the single-label existed but differed partially from the multi-label set, the sample was categorized as S_B, reflecting the presence of a dominant emotion along with minor inconsistencies. Finally, if no single-label annotation was available but multi-label information existed, the sample was assigned to S_C, indicating that no clear dominant label could be established and the annotators' votes were highly dispersed. This practical rule ensures a systematic and reproducible classification approach consistent with the voting mechanism of the IEMOCAP dataset. S_A and S_B with main labels are used for the pre-training of the exact model in the first stage. The trained exact model is then modified with multiple labels of S_C or $S_B + S_C$. Finally, the newly constructed labels are combined with S_A type samples for the second stage of model re-training.

For the first stage of training, as shown in Step I in Fig. 1, we use S_A and S_B as pre-trained samples, using one–hot hard labels as true values, and train a model with exact knowledge through traditional multi-classification cross-entropy, defined as follows:

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N} \mathbf{y}^i_{hard} \log\left(\hat{\mathbf{y}}\left(\mathbf{x}^i \mid M_p\right)\right), \tag{4}$$

where N represents the number of samples, $\hat{\mathbf{y}}\left(\mathbf{x}^i \mid M_p\right)$ represents the prediction output of the trained exact model M_p, and $\mathbf{y}^i_{hard}$ represents the one-hot hard labels. As the training samples used are data with dominant labels, the trained model can learn knowledge from exact samples. Further, by using this model to generate labels for ambiguous samples, and modifying the multiple labels for ambiguous samples according to a label correction strategy, the system can label samples from the emotional perspective of the model, rather than relying solely on the emotional cognition of a small number of annotators. In this study, we studied two types of label correction strategies, hard label correction, and soft label correction, summarized below.

(1) Soft label correction

This method directly uses the probability distribution vector output by the model as the generated label:

$$\hat{\mathbf{y}}^{\text{soft}}\left(\mathbf{x}^i \mid M_p\right) = \hat{\mathbf{y}}\left(\mathbf{x}^i \mid M_p\right). \tag{5}$$

Then, the corrected labels are weighted combined with the original multi-labels that are devoid of emotion bias, and the resulting corrected labels are represented as follows:

$$\Phi\left(\mathbf{y}^i_{multi}, \hat{\mathbf{y}}\left(\mathbf{x}^i|M_p\right)\right) = (1-\lambda)\mathbf{y}^i_{multi} + \lambda\hat{\mathbf{y}}^{soft}\left(\mathbf{x}^i|M_p\right), \tag{6}$$

where $\lambda \in [0,1]$ represents the balance factor, which is used to balance the relative importance of the original multiple labels and the model-generated labels.

(2) Hard label correction

This method requires further selection of the most likely class from the probability distribution vector output by the model, and the generation of one-hot hard labels as a result, to more clearly represent the model's label choices. The process of obtaining one-hot hard labels is as follows:

$$\hat{y}^{hard}_k\left(\mathbf{x}^i \mid M_p\right) = \begin{cases} 1, & \text{if } k = \arg\max_j \hat{y}_j\left(\mathbf{x}^i \mid M_p\right), \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where $j, k \in [1, K]$ is an integer, $\hat{y}^{hard}_k\left(\mathbf{x}^i \mid M_p\right)$ represents the value of the kth emotion category, and $\hat{y}_j\left(\mathbf{x}^i \mid M_p\right)$ represents the jth emotion category of $\hat{y}\left(\mathbf{x}^i \mid M_p\right)$.

Correspondingly, the balance factor λ is used to weight the generated labels and the original no emotion bias multi-labels to measure their relative importance. Specifically, the corrective label states the following:

$$\Phi\left(\mathbf{y}^i_{multi}, \hat{\mathbf{y}}^{hard}\left(\mathbf{x}^i \mid M_p\right)\right) = (1-\lambda)\mathbf{y}^i_{multi} + \lambda\hat{\mathbf{y}}^{hard}\left(\mathbf{x}^i \mid M_p\right), \tag{8}$$

where $\hat{\mathbf{y}}^{hard}\left(\mathbf{x}^i \mid M_p\right)$ is the vector form of $\hat{\mathbf{y}}^{hard}_k\left(\mathbf{x}^i \mid M_p\right)$.

After correction with labels, it will enter the second stage of training. For the second stage of training, shown in Fig. 1 Step II, we use the label-corrected samples combined with the S_A samples to reconstruct the model M_f for training. The cross-entropy loss function for this training process is defined as follows:

$$L_{correct} = -\frac{1}{N}\sum_{i=1}^{N}\Phi\left(\mathbf{y}^i_{multi}, \hat{\mathbf{y}}^*\left(\mathbf{x}^i \mid M_p\right)\right)\log\left(\hat{\mathbf{y}}\left(\mathbf{x}^i \mid M_f\right)\right), \tag{9}$$

where $\hat{\mathbf{y}}\left(\mathbf{x}^i \mid M_f\right)$ represents the predicted output vector of the reconstructed model M_f, and $*$ represents either $soft$ or $hard$.

As there is no emotion ambiguity in the sample of S_A, its hard-label and multi-label expressions are consistent, so S_A still uses the one-hot hard labels as the true value. Although S_B has a dominant label, like S_C, there is an emotional ambiguity. Therefore, we investigated two training methods, only correcting S_C and correcting $S_B + S_C$. The details are summarized as follows.

(1) Correction of only S_C: This method acknowledges the emotional perceptions of most annotators and maintains S_B as the dominant target label, using the one-hot hard labels representation. As S_C does not have a dominant label, we use the aforementioned label correction strategy to modify its multiple labels before fusing it into the model training.

(2) Correction of $S_B + S_C$: This method does not agree with the emotion cognition of most annotators, as long as there is ambiguity in the sample, it is considered to be an ambiguous sample. Therefore, the aforementioned label correction strategy is adopted to correct the multiple labels for $S_B + S_C$.

3.4. The spatial–temporal parallel network

The training model adopted in we can be divided into four modules: temporal emotion feature extraction, spatial emotion feature extraction, spatial–temporal emotion feature fusion, and linear classification. Next, we will explain each of them.

(**1**) *Temporal emotion feature extraction*

Speech is a continuous signal with a temporal nature and variable length, and its emotional expression can dynamically change over time (Mao et al., 2021). Benefiting from the progress in speech recognition research, the Wav2vec model (Park et al., 2023), which is pre-trained using a large amount of unlabeled data, can provide us with a contextual representation of speech features. This is because Wav2vec is implemented based on the $Transformer$ (Vaswani et al., 2017) architecture, which predicts certain sampling points in the future by learning the context of the current input. Therefore, the pre-trained Wav2vec model can obtain contextual emotion representations from the original waveform of speech, which has been studied in the speech emotion recognition field (Sharma, 2022). Specifically, the formula can be described as follows:

$$\mathbf{f}_w = \mathrm{Wav2vec}(\Theta_w, \mathbf{x}^i_w), \tag{10}$$

where $\mathbf{x}^i_w$ represents the original speech directly input into the pre-trained Wav2vec model, Θ_w is a series of trainable parameters for the Wav2vec model, $\mathbf{f}_w \in \mathbb{R}^{T_t \times d_t}$ represents the potential emotion representation obtained with contextual features, T_t represents the temporal dimension and its size depends on the length of the input speech, and d_t represents the feature dimension.

Then, a problem worth exploring arose. Due to the variability in the length of input speech, the potential emotion representation processed by the Wav2vec model still has variability in length in the temporal dimension, which is detrimental to subsequent feature fusion and classification. Therefore, we introduced the temporal pyramid pooling to handle this variability. This pooling method is widely used in audio and video processing, extracting multi-level information from the temporal dimension. As shown in Fig. 3, divide the input feature into regions in a coarse-to-fine manner on the time axis, perform pooling operations in each region, and finally concatenate all pooling results along the time axis to obtain a fixed-length feature, which depends on the setting of Pyramid levels (PL). For example, $PL = \{1, 2, 3\}$, it is necessary to perform 3 region segmentations on the feature's time axis and 6 pooling operations, resulting in a feature size of 6xd. Here, it is assumed that n pooling operations are performed, which means that the temporal dimension of the fixed latent emotion representation $\mathbf{f}_w$ is n:

$$\mathbf{f}_t = \mathrm{TPP}(\mathbf{f}_w), \tag{11}$$

where $\mathbf{f}_t \in \mathbb{R}^{n \times d_t}$ represents a fixed size temporal emotion feature obtained through TPP (Yu et al., 2019).

(2) *Spatial emotion feature extraction*

In the natural environment, the sounds we hear are often composed of multiple complex frequencies that overlay each other, and these frequency variations include pitch, phonemes, and other emotional features related to sound attributes. Although information about emotion changes over time can be obtained from the original waveform of the speech, this one-dimensional representation cannot represent the two-dimensional sound properties, lacking frequency variation information in the spatial domain. Therefore, to compensate for the deficiencies of one-dimensional raw speech in the spatial domain, we employ multi-layer CNNs (Leo et al., 2022b) to learn more abstract spatial emotional features from the fixed-resolution spectral features of speech, thereby enhancing the fine-grained temporal emotional features. This approach differs from traditional methods that utilize multi-resolution data as model inputs (Peng et al., 2021). The structure is concretely illustrated

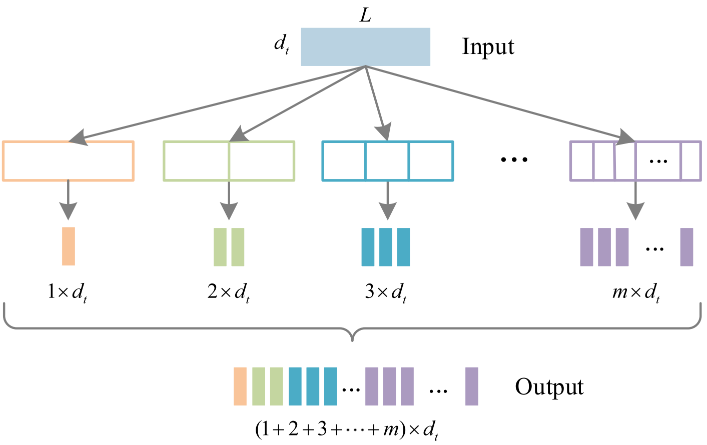


Fig. 3. A schematic diagram of temporal pyramid pooling.

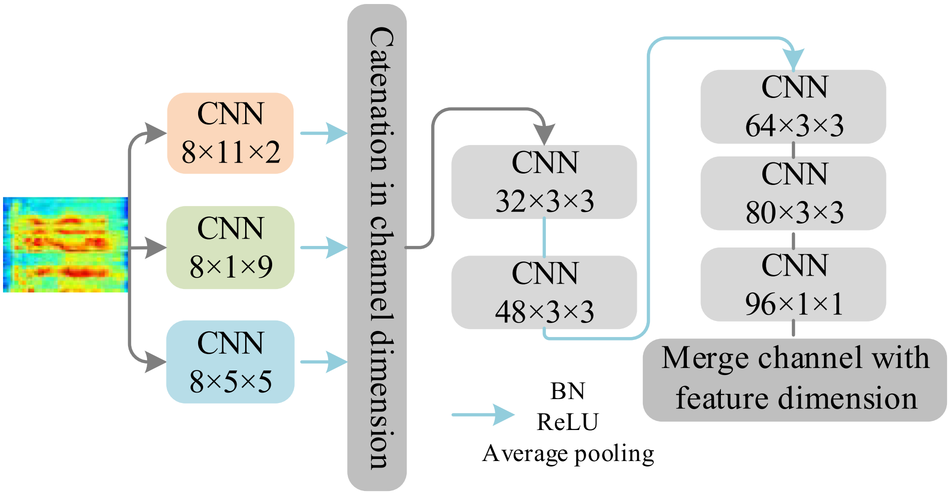


Fig. 4. Spatial structural diagram.

by Fig. 4, and it is worth noting that average pooling is used in the first three layers to downscale the features, with the aim of focusing the model on certain useful features and preventing overfitting triggered by too many parameters. Here, we summarize the overall process described above:

$$\mathbf{f}_c = \mathrm{Spatial}(\Theta_c, \mathbf{x}_s^i), \tag{12}$$

where $\mathrm{Spatial}(\cdot)$ is a brief description of the formula in Fig. 4, Θ_c is a set of trainable parameters for a convolutional network, $\mathbf{f}_c \in \mathbb{R}^{L_c\times d_c}$ represents the extracted spatial features, L_c represents the temporal dimension and its size depends on the input speech length, and d_c represents the feature dimension.

However, since the features reduced by averaging pooling are already very condensed in the temporal dimension, redundant zero-fill information will be introduced by using multi-level information extraction. Therefore, the features extracted from fixed spatial domains in this paper can be globally averaged pooling, which is equivalent to $PL = \{1\}$. At this point, the equivalent of automatically averaging on the time axis using GAP (Chang et al., 2024) is obtained, and the resulting spatial emotion features of fixed dimensions can be expressed as:

$$\mathbf{f}_s = \mathrm{GAP}(\mathbf{f}_c), \tag{13}$$

where $\mathbf{f}_s \in \mathbb{R}^{d_s}$, d_s represent the dimensions of the spatial features.

(3) *Spatial–temporal feature fusion*

The temporal domain encoder described above is based on the temporal domain analysis of the original waveform of the speech, while the spatial domain encoder is based on spatial analysis of the spectral features of the speech. Temporal domain analysis lacks direct insight into the frequency features of speech, while spatial analysis lacks a time-varying relationship between speech and noise. Therefore, combining the advantages of the two fields is a question worth exploring. Inspired by the thinking of Zou et al. (2022) (co-attention), we combine spatial emotion features into temporal emotion features to achieve spatial-domain-assisted emotion classification in the temporal domain.

Firstly, spatial-domain emotion features $\mathbf{f}_s$ are converted into spatial-domain emotion weights $\mathbf{f}_s'$ via a fully connected layer:

$$\mathbf{f}_s' = \delta\left(\mathbf{f}_s\mathbf{W}_s + \mathbf{B}_s\right), \tag{14}$$

where $\mathbf{f}_s' \in \mathbb{R}^n$ and n are the temporal dimension size of temporal emotion feature, $\delta(\cdot)$ indicates ReLU activation, $\mathbf{W}_s \in \mathbb{R}^{d_s\times n}$ and $\mathbf{B}_s \in \mathbb{R}^n$ represents fully-connected parameters.

Then, apply the spatial emotion weight obtained above to the temporal emotion feature to obtain the temporal emotion feature $\mathbf{f}_t'$ with spatial emotion information:

$$\mathbf{f}_t' = \mathbf{f}_s' \cdot \mathbf{f}_t, \tag{15}$$

where $\mathbf{f}_t' \in \mathbb{R}^{d_t}$.

(4) *Feature classification*

The classifier used in this study consists of multiple fully connected layers, which aims to allow the model to more carefully learn the distributed representation of features, and to map the aforementioned emotion features into the classification output. The process is represented as follows:

$$\begin{aligned}\hat{\mathbf{y}}^i &= \delta\left(\delta\left(\mathbf{f}_t'\mathbf{W}_f^1 + \mathbf{B}_f^1\right)\mathbf{W}_f^2 + \mathbf{B}_f^2\right)\mathbf{W}_f^3 + \mathbf{B}_f^3, \\ \hat{\mathbf{y}}\left(\mathbf{x}^i \mid *\right) &= \mathrm{softmax}\left(\hat{\mathbf{y}}^i\right),\end{aligned} \tag{16}$$

where $\mathbf{W}_f^1, \mathbf{W}_f^2 \in \mathbb{R}^{d_t \times 128}, \mathbf{W}_f^3 \in \mathbb{R}^{128 \times K}, \mathbf{B}_f^1, \mathbf{B}_f^2 \in \mathbb{R}^{128}, \mathbf{B}_f^3 \in \mathbb{R}^K$ is a fully connected trainable parameter, K is the number of classes for the classification task, $\delta(\cdot)$ represents the ReLU activation function, and $*$ refers to a model. The $softmax(\cdot)$ (Yin et al., 2024) activation function is commonly used together with the cross-entropy loss function to avoid gradient explosion.

4. Experimental setup

In this section, we describe the data used in conducting the experiments and the details of implementation. All experimental codes can be obtained from the GitHub.[1]

4.1. Data processing

(1) Dataset

Considering that the proposed model is to alleviate the Emotional ambiguity of multi-label speech, the single-label dataset cannot verify the performance of the proposed method. Currently, the only suitable multi-label dataset is the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset (Busso et al., 2008). Therefore, we will test our modelability and comparison on the IEMOCAP dataset.[2] This dataset is an English corpus commonly used in speech emotion recognition, recording multiple behavioral modalities of ten professional actors participating in interactions in improvisational and scripted performance scenarios, such as video, speech, text, and so on. For speech data, the emotion classification label for each speech is annotated by at least three annotators, and each annotator is allowed to annotate multiple emotion labels. Furthermore, the categorization labels of speech samples may contain more than one emotion, reflecting differences in human expression and cognition of speech emotion. To explore the problem of differences, we followed research Li et al. (2023) and Fujioka et al. (2020), using data from improvisational and scripted scenarios, and conducted experimental evaluations using anger, happiness, sadness, and neutral emotions as the basic emotions. Additionally, according to studies Hou et al. (2021) and Yin et al. (2021), excitement and happiness have similar activation and valence states, so excitement is included in happiness.

(2) Preprocessing

The speech processing flow is shown in Fig. 5. The speech signal first undergoes pre-emphasis, then is divided into frames with windowing to meet short-time stationarity requirements. Each frame is transformed into the frequency domain using the Short-Time Fourier Transform (STFT) to obtain spectrograms. To simulate human auditory perception, Mel-scale filter banks are applied to perform nonlinear frequency warping on the spectrum, followed by logarithmic processing to generate FBank features. Although Mel-Frequency Cepstral Coefficients (MFCCs) can be obtained by applying the Discrete Cosine Transform (DCT) to FBank features for dimensionality reduction, this process loses important nonlinear information. To preserve the complete perceptually-relevant nonlinear characteristics, we use FBank features as input to the spatial processing module.

Table 1
Distribution of speech samples across three categories (S_A, S_B, S_C) for each speaker in the IEMOCAP dataset. S_A denotes exact samples, S_B denotes ambiguous samples with a dominant label, and S_C denotes ambiguous samples without a dominant label. These divisions are used in the proposed label correction experiments.

Speaker	S_A	S_B	S_C
Session1M	325	365	392
Session1F	224	402	334
Session2M	292	396	401
Session2F	243	317	400
Session3M	298	444	452
Session3F	258	324	521
Session4M	157	458	602
Session4F	297	304	457
Session5M	372	454	460
Session5F	230	438	430

In the dataset, there are speech samples that are too long (greater than 30 s), which will result in excessive server memory requirements for training and cannot meet practical implementation requirements. In addition, studies (Aftab et al., 2022) have shown that a 7-second speech can contain sufficient emotional information. The session1 in the first column of Table 1 indicates the session number, and the rest are similar, with M representing men and F representing women. Therefore, we divide speech longer than 7 s into segments of 7 s each, while maintaining the original length of speech shorter than 7 s. The details of the data distribution are shown in Table 1. In addition, some of the speech samples are too short for Wav2vec model processing, so the speech samples with a duration of less than 1.5 s are padded with 1.5 s. Due to the need to fix the length of each batch during neural network training, each speech segment needs to be padded with zeroes to represent the maximum length of the input speech for that batch. To minimize the adverse effects of zero padding on training, we sorted all the speech segments by length and divided them into batches to reduce the number of zero padding segments in each speech segment and accelerate training convergence.

After the above processing, further process the speech into suitable input features for model learning. We use two forms of the original signal and the spectral features of the speech. For the original speech signal, it is converted into a digital signal that can be processed by a computer with a sampling frequency of 16 KHz, and then the high-frequency components in the speech signal are enhanced through pre-emphasis, which can be represented as $\mathbf{x}_i^w \in \mathbb{R}^T$. For spectral features of speech, we adopt the Log-Mel filterbank energy feature (MFB) (Chen et al., 2018, 2021), which has human auditory features, applied 40 mel filters and a 40 ms Hamming window with a hop length of 10 ms, and calculated it as $\mathbf{x}_s^i \in \mathbb{R}^{L \times 40}$. Additionally, the features described above are normalized based on the mean and variance of each speaker.

4.2. Implementation details

The construction and training verification of the proposed model is implemented on 4× NVIDIA RTX 2080TiGPU using the PyTorch1.7.0 framework (Paszke et al., 2019), with specific parameter settings shown in Table 2. The parameters of the entire model are 94.7 M, and the number of floating-point operations per second is 6.9 G. During the training process, the initial and secondary iterations are performed under identical settings, with the exact model used to correct multiple labels being the optimal model for the first stage of training. Additionally, for each training step, the inverse value of the frequency of

[1] Code: https://github.com/mychongu/SER-ambiguous-speech.git

[2] Dataset: https://www.kaggle.com/datasets/dejolilandry/iemocapfullrelease

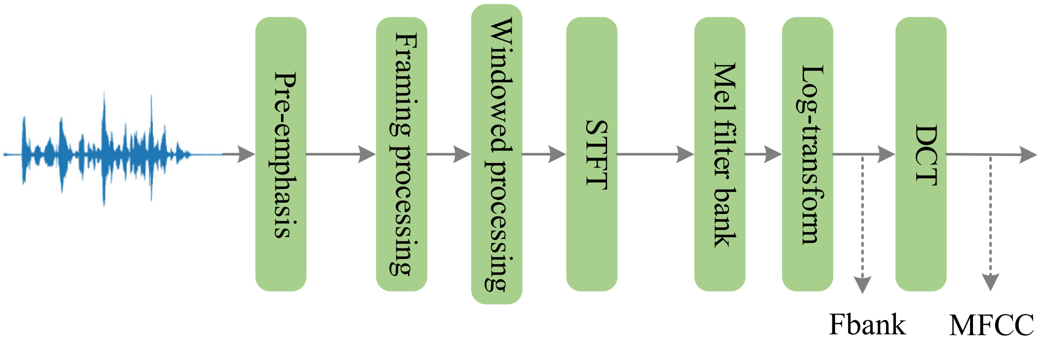


Fig. 5. Speech processing flow .

Table 2
Parameter setting details.

Name	Value
Pyramid levels	2, 4, 8, 16, 32, 64
Batch size	16
Optimizer	Adam
λ	0.5
Learning rate	1e-5
Weight decay	0.0005
Max epoch	100
Early stopping patience	8

the emotion category is used as the loss class weight. The verification process, as per the findings in most research in speech emotion recognition (Li et al., 2023; Ntalampiras, 2021), all our experiments use the LOSO cross-validation method, which preserves the data of one speaker for testing while using the remaining speaker's data for model training. Additionally, the testing sets were speech samples with a dominant label, and this label was the true label for the testing set. The evaluation metric adopted for the experiment employed both the widely applied Weighted Accuracy (WA) and Unweighted Accuracy (UA) metrics for speech emotion recognition. WA is the classification accuracy of all utterances, while UA is the average accuracy of each emotion category.

5. Experimental results and analysis

In this section, we present experimental results of the proposed method, demonstrating the superiority of the proposed method through comparison with state-of-the-art methods and ablation experiments: (1) The spatial domain emotion information is beneficial for temporal domain emotion classification; (2) The method of the corrective label can improve the learning efficiency of the model on ambiguous speech samples; (3) In addition to correcting the labels of ambiguous samples S_C, the label correction strategy can also correct the labels of samples S_B, further enhancing the performance of the model.

5.1. Compared with the state-of-the-art methods

To demonstrate the superior performance of our proposed method in handling speech ambiguity, we conduct a comparative analysis with state-of-the-art blur processing methods. Since our approach operates on a multi-label dataset and requires correction of multiple labels to expand the available dataset of speech samples, models such as Gan et al. (2023), Jalal et al. (2020), which do not utilize multi-label datasets or perform label correction to utilize ambiguous speech samples, as well as multi-modal methods that utilize both text and audio (Atmaja et al., 2019), are not suitable for comparison. The specific details of the comparative methodology are as follows.

Soft-target training (2018) (Ando et al., 2018): A soft label training method is defined, which successfully utilizes unlabeled ambiguous speech by modifying the representation of soft labels.

Emotion existence (2019) (Ando et al., 2019): A method for assessing the presence or absence of emotions is proposed. It first evaluates the presence of emotions using multi-label classification and then uses soft labels to retrain the dominant emotion recognition model, thereby alleviating the emotion ambiguity problem.

Joint-learning (2019) (Chou and Lee, 2019): A method for joint learning soft labels, hard labels, and annotator traits is proposed. It fuses annotator traits into the label distribution to consider both label and annotator uncertainties.

Meta-learning (2020) (Fujioka et al., 2020): A method based on meta-learning is proposed to achieve label correction and sample weight estimation. It uses two trainable parameters to perform label correction and sample weight estimation during training, aiming to update the noise labels of training samples and ignore the contribution of ambiguous samples to model training.

Co-teaching (2021) (Yin et al., 2021): A progressive collaborative training method is proposed, which considers the ambiguity of emotions and uses loss values to identify the difficulty level of speech samples. The model completes the establishment of emotional knowledge through alternating training from simple to difficult samples.

Multi-view (2022) (Hou et al., 2021): A collective multi-view relationship network is proposed, which considers the complex emotion relationships of speech from multiple acoustic feature perspectives to improve recognition efficiency.

Inter-class difference loss (2023) (Li et al., 2023): A class difference loss function that can be used for multi-label training is proposed. It defines sample labels as multi-label forms and designs a class difference loss function to learn the emotion distribution in speech samples automatically.

LLMs (2024) (Wang et al., 2024): A novel methodology integrating Large Language Models (LLMs) with efficient spatial learning, which combines LLMs-synthesized emotionally rich speech data with the IEMOCAP dataset and student speech datasets. Furthermore, the approach extracts effective spatial emotion features from speech signals using a Transformer network architecture.

Table 3 gives the results of the comparison experiments on the IEMOCAP dataset, when only the samples in S_A were used as the test set. Compared to the state-of-the-art methods for speech ambiguous processing described above, our proposed method exhibits superiority in terms of WA and UA (WA = 72.4% and UA = 73.1%). It is important to note that, for comparison with other methods, we represent the dataset as a dominant dataset and an ambiguous dataset. The dominant dataset corresponds to our $S_A + S_B$ dataset mentioned in this study, while the ambiguous dataset is equivalent to S_C. Furthermore, the experimental results on the ambiguous dataset only corrected the labels of S_C samples, and this correction was achieved using the soft labels correction strategy based on the balance factor $\lambda = 0.5$.

Compared to the hard label methods (Hou et al., 2021; Yin et al., 2021), and LLMs (Wang et al., 2024), our method achieves significant improvements in performance indicators. Specifically, compared to the Co-teaching (2021) (Yin et al., 2021) method, we have a 10.1% WA

Table 3
Performance comparison on the IEMOCAP dataset for speech emotion recognition. “Dominant” refers to samples with a dominant label, and “Ambiguous” refers to ambiguous samples. WA (Weighted Accuracy) indicates overall accuracy across all samples, while UA (Unweighted Accuracy) represents average accuracy across emotion classes to mitigate class imbalance.

Method	Label	Training set		Metric	
		Dominant	Ambiguous	WA (%)	UA (%)
Soft-target training (2018) (Ando et al., 2018)	Soft	√	–	58.5	57.4
		–	√	53.6	54.0
		√	√	62.6	63.7
Emotion existence (2019) (Ando et al., 2019)	Multi soft	√	√	66.1	65.4
Emotion existence (2019) (Ando et al., 2019)	Multi soft	√	√	66.1	65.4
Joint-learning (2019) (Chou and Lee, 2019)	Soft Hard	√	–	–	61.5
Meta-learning (2020) (Fujioka et al., 2020)	Update hard	√	–	65.9	61.4
Co-teaching (2021) (Yin et al., 2021)	Hard	√	–	62.3	–
Multi-view (2022) (Hou et al., 2021)	Hard	√	–	–	66.6
Inter-class difference loss (2022) (Li et al., 2023)	Multi	√	–	66.0	63.9
		–	√	60.5	61.7
		√	√	68.3	66.2
Ours	Update multi	√	–	**69.9**	**70.7**
		–	√	**65.2**	**66.5**
		√	√	**72.4**	**73.1**

advantage on the dominant dataset and 2.5% further improvements with the addition of an ambiguous dataset. Similarly, compared to the UA of the Multi-view (2022) (Hou et al., 2021) method, our method has a 6.5% advantage in the dominating dataset, and can further improve 6.5% when ambiguous data is added. Compared to LLMs (Wang et al., 2024) models, our model also achieves substantial performance improvements. This is because the spatial–temporal network model we built extracts efficient emotional features from both the temporal and spatial domains of speech. More importantly, the proposed label correction strategy can effectively utilize additional ambiguous data emotional information to assist in model training, thereby further promoting the establishment of emotional cognition in the model.

Compared to soft label (Ando et al., 2018) and multi-label (Li et al., 2023; Ando et al., 2019) methods, our proposed method performs better. On the dominant and ambiguous mixed dataset, compared to the soft labels method proposed in Ando et al. (2018), WA and UA are 9.8% and 9.4% higher, respectively. This is because our label correction strategy can obtain more reliable emotion cognition from pre-trained exact models, rather than relying solely on the emotion distribution assigned by a few annotators. Compared to the multi-label method used in Ando et al. (2019), Li et al. (2023), the results showed that WA increased by 6.3% and 4.1%, respectively, while UA increased by 7.7% and 6.9%. One reason is that the two-stage training method in this paper fully utilizes ambiguous speech samples, increasing the model’s stability. Another reason is that the correction strategy using a balance factor enables no emotional bias multi-labels to gain emotional bias, allowing the model to have a clearer emotional learning direction during training.

Compared with the improved training methods (Chou and Lee, 2019; Fujioka et al., 2020), the performance of our proposed method is optimal. Compared to the Joint-learning (2019) (Chou and Lee, 2019) method, our method can increase UA by 11.6%. Compared to the Meta-learning (2020) (Fujioka et al., 2020) method, WA has improved by 6.5% and UA has improved by 11.7%. This is because we not only address the ambiguity of samples with dominant labels (dominant dataset) but also consider the possibility of samples without dominant labels (ambiguous dataset) having ambiguity. We adopted a two-stage training method and used a label correction strategy to successfully extract effective emotion features from the ambiguous dataset that can promote emotion learning of the model, achieving further improvement in the recognition rate.

Importantly, since ambiguous datasets contain samples where humans cannot reach an emotional consensus, they are more difficult to use for emotion learning than dominant datasets. This can also be seen from the experimental results, where the emotion recognition rate of the dominant dataset (9.8% and 9.4%) is significantly higher than that of the ambiguous dataset (WA = 65.2% and UA = 66.5%), as is also the case with the experimental results of studies (Ando et al., 2018; Li et al., 2023). However, our method for label correction on ambiguous datasets outperforms the soft label method (Ando et al., 2018) by 11.6% WA and 12.5% UA, and outperforms the multi-label method (Li et al., 2023) by 4.7% WA and 4.8% UA. The reason is that our label correction strategy can adjust the labels of ambiguous speech samples through pre-trained exact models, making the labels not only dependent on the emotion cognition of a small number of annotators but also using a balance factor to indicate the emotion bias of the target labels for model training. In addition, it can be observed from the table that using a mixture of dominant and ambiguous datasets for training can improve the performance of emotion recognition. This is because there are differences in the data distribution of the training set. Training sets that only contain clear data or ambiguous data are one-sided in guiding model training, making the model training incomplete. In the real world, speech cannot be clear or ambiguous, so training sets that consider both clear and ambiguous speech are more appropriate for real-life scenarios. We take into account the mixed condition of clear and ambiguous speech, which is also an advantage of our method.

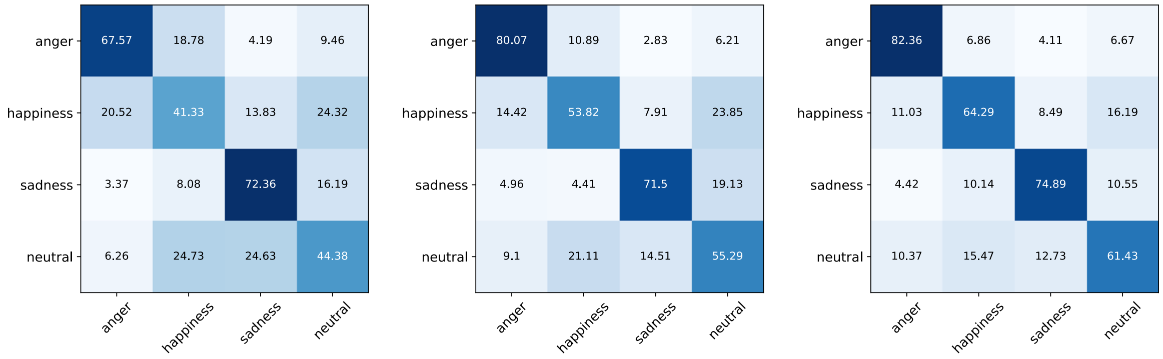
5.2. Ablation analysis

The comparison with hard labels, soft labels, multiple labels, and improved training methods described above proves the superiority of our proposed approach. To further analyze the specific reasons, we will analyze some ablation experiments, including the effects of the spatial-domain-assisted temporal-domain label correction strategy, and the effect of sample distribution on the performance of the model.

(1) The effect of spatial domain assisted temporal domain

As can be seen from Table 4, the temporal domain processing for speech is more effective, so we use the temporal domain as the main focus, and the spatial domain as a complement. To explore the complementary effects of the spatial domain on the temporal domain, Table 4 shows the fusion between spatial and temporal domains and provides further analysis using the confusion matrix as shown in Fig. 6.

In this comparative experiment, the training data is data with dominant labels ($S_A + S_B$), and the test set data are also derived from these two sets. The single-temporal domain method uses average pooling after the temporal pyramid module, while the single-spatial domain method uses a four-classifier after global pooling. The remaining settings remain unchanged. From Table 4, it can be seen that



(a) Spatial method (b) Temporal method (c) Spatial assist temporal

Fig. 6. The confusion matrix of spatial and temporal domain ablation experiments.

Table 4
The ablation of spatial and temporal domains.

Fold	Method					
	Spatial		Temporal		Spatial assist temporal	
	WA (%)	UA (%)	WA (%)	UA (%)	WA (%)	UA (%)
1	55.4	60.0	66.8	68.1	**71.9**	**74.4**
2	56.7	58.2	64.4	64.8	**69.0**	**70.7**
3	56.0	61.1	68.0	73.1	**72.2**	**75.8**
4	53.4	56.7	65.9	68.0	**73.6**	**74.9**
5	52.6	54.4	63.5	64.0	**68.3**	**68.6**
6	54.8	53.1	62.5	61.7	**68.0**	**67.9**
7	50.4	52.8	60.5	62.2	**67.3**	**69.6**
8	54.7	53.6	64.1	61.9	**72.4**	**68.1**
9	47.9	55.0	55.7	59.1	**66.8**	**66.6**
10	56.7	59.1	65.6	68.8	**69.0**	**71.0**
Average	53.9	56.4	63.7	65.2	**69.9**	**70.7**

Table 5
Performance comparison of the label correction.

Fold	Method					
	No correction		Hard label correction		Soft label correction	
	WA (%)	UA (%)	WA (%)	UA (%)	WA (%)	UA (%)
1	70.0	73.0	71.3	**74.1**	**71.7**	73.5
2	67.9	69.9	68.7	**70.2**	**68.7**	69.2
3	70.2	75.2	72.5	74.0	**73.5**	**75.9**
4	**75.2**	**75.9**	73.4	75.8	73.8	75.4
5	**67.9**	68.6	67.7	68.0	67.7	**68.6**
6	**70.5**	**69.9**	67.4	66.0	69.1	67.7
7	64.9	67.0	70.2	71.7	**70.9**	**72.9**
8	**73.4**	66.7	69.1	64.4	71.5	**66.8**
9	**69.0**	**69.3**	66.0	67.6	68.2	69.2
10	69.5	71.3	68.5	71.8	**71.7**	**72.7**
Average	69.8	70.7	69.5	70.4	**70.7**	**71.2**

the single-temporal domain method achieves WA = 63.7% and UA = 65.2% performance, while the spatial domain-assisted temporal domain method can improve WA by 6.2% and UA by 5.5%, indicating that the spatial domain contains effective emotion information that the temporal domain lacks.

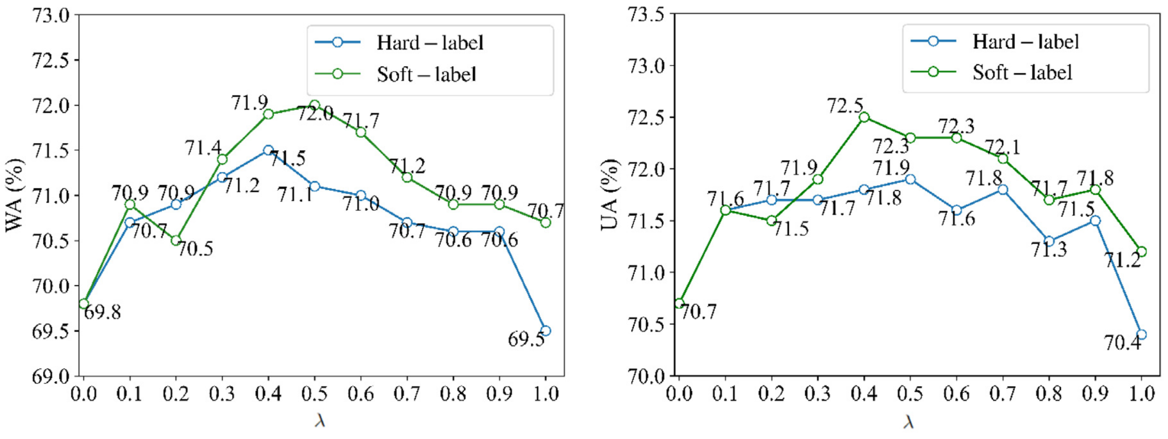
Further, the spatial domain's effect on the temporal domain's emotion classification can be observed in the confusion matrix in Fig. 6. From Fig. 6(a), it can be seen that the single-spatial domain method is easily confused with other emotion categories in terms of happiness and neutrality. This problem also exists in the single-temporal-domain method (Fig. 6(b)). Comparing Fig. 6(a) and (b), the single spatial domain method has a higher recognition rate for sadness, while the single temporal domain method has a higher recognition rate for anger, indicating that the spatial domain and temporal domain have different emotion advantages. After combining spatial-domain emotion features into the temporal domain (Fig. 6(b)), the recognition rates for happiness and neutral categories increased by 10.47% and 6.14%, respectively, on top of the single-temporal-domain approach, alleviating the problem of emotion confusion that often occurs when using a single-temporal-domain approach for happiness and neutral categories. Additionally, the recognition rates of anger and sadness also increased, by 2.29% and 3.39% respectively, indicating that the fusion module can combine the emotion advantages of both spatial and temporal domains.

(2) The effect of the label correction

Table 5 presents the ten-fold cross-validation results for both the hard label correction and soft label correction strategy, as compared to the no label correction. Additionally, because the balance factor in formulae (6) and (8) are critical to label correction strategies, further discussion is provided on the impact of the balance factor on label correction strategies, as illustrated by the line chart in Fig. 7.

In the experiment, the training data includes S_A and S_B with dominant labels, as well as S_C with no dominant label. Among them, S_A and S_B target single labels, and S_C target multiple or generated labels, and test sets are derived from S_A and S_B. This experiment first verifies the impact of soft label correction and hard label correction on the model performance. Therefore, $\lambda = 1$, indicating that, for the class S_C separately generated hard labels or soft labels. Additionally, when unlabeled data is corrected to $\lambda = 0$, it indicates that S_C exclusively focuses on using multiple labels as targets. Table 5 shows the average results of the above three methods in WA and UA. Based on the average results, it can be seen that the hard label correction strategy decreases the performance of non-label correction. This is because the single emotion expression form of hard labels cannot represent the multiple emotion mixtures of ambiguous speech. The soft label correction strategy improves the performance of non-label correction, indicating that the generated soft labels obtained from pre-trained exact models are effective for model emotion learning, and soft labels with emotion bias are more suitable as emotion learning targets for ambiguous speech than multiple labels without emotion bias.

Further, a more detailed comparison of the impact of the balance factor λ on the label correction strategy is shown in Fig. 7. It can be seen that the performance impact of the balance factor λ on WA and UA is similar. Specifically, when the balance factor λ is greater than 0.3, the WA and UA of the soft labels correction strategy outperform the hard labels correction strategy. This is consistent with the above results for soft label and hard label correction when $\lambda = 1$, indicating that the soft label correction strategy is indeed superior to the hard label correction strategy. At the same time compared to $\lambda = 0.3$, the model performance at $\lambda = 0$ (which corresponds to not utilizing the exact model of the first stage, when the labels of the ambiguous speech are not corrected) performs worse, which suggests that utilizing the



(a) The effect of λ on WA

(b) The effect of λ on UA

Fig. 7. The effect of the balance factor.

exact model of this paper helps to improve the model performance. In addition, there is a peak in the curve changes in both Fig. 7(a) and (b), indicating that the balance factor can adjust the relative importance of the original multiple labels and the model-generated labels, resulting in better performance in model learning.

(3) The effect of correcting the number of speech samples

Due to the experimental results above showing better correction effects for soft labels, we experimented to analyze the impact of the correction sample size on the performance of the model using soft label correction strategies. Table 6 presents the experimental results. On this basis, in order to better demonstrate the impact of different speech sample sizes on emotion classification, Fig. 8 shows the t-SNE visualization of the optimal and original results.

Table 6 shows the experimental results of analyzing and correcting the number of speech samples using soft label correction strategies with different balance factor λ. For the selection of the number of corrective speech samples, we adopted two methods. One method is to correct only the multiple labels of the ambiguous speech sample S_C, while maintaining the target label of S_B as the dominant label, and the test set comes from S_A and S_B. The other is to correct the ambiguous multiple labels of the main label sample S_B and the ambiguous label sample S_C, and the test set comes from S_A. The reason is that the labels of S_B and S_C both have differences in human emotion perception, while S_A does not, so experiments were conducted on these two methods. From Table 6, it can be seen that the model performance is optimal when the balance factor $\lambda = 0.8$ is used and S_B and S_C are corrected. This indicates that although the S_B sample has the emotion labels recognized by most people, there are still errors in the emotion label representation given to it.

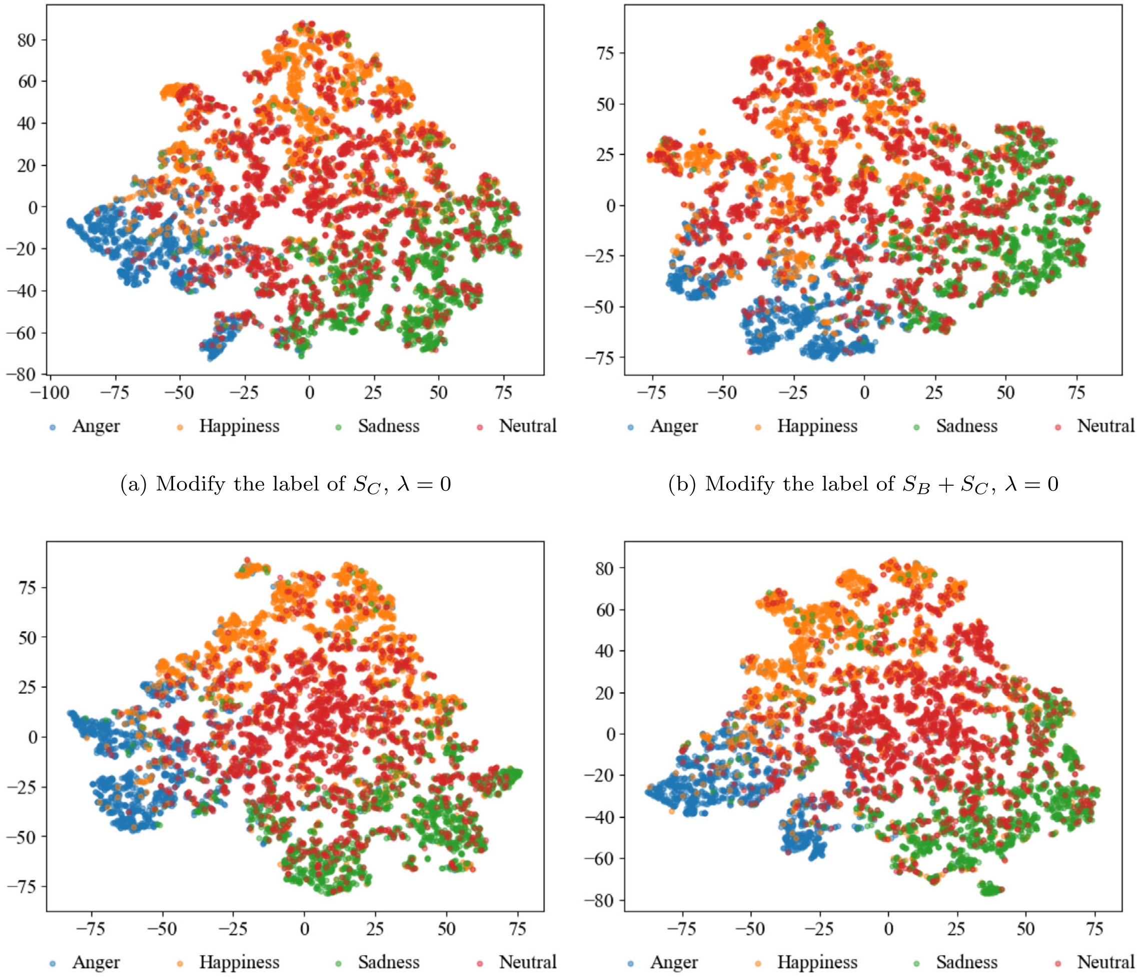
Fig. 8 illustrates the representative emotion clustering results for four scenarios using the t-SNE method, which are (a) only correcting the ambiguous speech sample S_C with the balance factor $\lambda = 0$, at which point the label for S_C is expressed by the original multi-label representation; (b) When the balance factor $\lambda = 0.5$ is set only to correct the ambiguous speech samples S_C, the labels of S_C are expressed by combining the original multiple labels and the soft labels generated by the exact model; (c) Correct the ambiguous samples S_B with dominant label and ambiguous speech samples W3 when the balance factor W0 is applied, and the labels of S_B and S_C are expressed by the original multiple labels. (d) Correct the ambiguous samples S_B with dominant label and ambiguous speech samples without dominant label S_C when the balance factor $\lambda = 0.8$. At this time, the labels of S_B and S_C are expressed by a combination of the soft labels generated by the exact model and the original multiple labels. As can be seen from the figure, the degree of emotion clustering is from small to large (c)<(a)<(b)<(d). When replacing the dominant label of S_B with multiple labels, the degree of emotion clustering is reduced, as multiple labels cannot indicate the explicit emotion preferences of the model training samples, i.e., (c)<(a). When using soft labels generated from pre-trained exact models to adjust the without emotion bias multiple labels, the model has a clear emotion bias when learning the features of the training samples, resulting in a significantly higher degree of emotion clustering for (b) and (d) than for (a) and (c). When soft label correction is added to S_B, the emotion clustering effect is also improved, such as orange happy and green sad clustering. Therefore, (b)<(d).

Table 6
The ablation of correcting samples.

λ	Method			
	Correct S_C		Correct $S_B + S_C$	
	WA(%)	UA(%)	WA(%)	UA(%)
0	69.8	70.7	68.7	69.9
0.1	70.9	71.6	69.5	70.5
0.2	70.5	71.5	70.4	71.3
0.3	71.4	71.9	71.2	72.5
0.4	71.9	72.5	70.6	71.9
0.5	**72.0**	**72.3**	71.7	72.4
0.6	71.7	72.3	72.3	72.8
0.7	71.2	72.1	71.7	72.6
0.8	70.9	71.7	**72.4**	**73.1**
0.9	70.9	71.8	71.4	71.9
1	70.7	71.2	69.9	71.1

6. Summary

The proposed model demonstrates certain advantages in ambiguous speech emotion recognition, and its effectiveness has been verified through systematic experiments. Experimental results on the IEMOCAP dataset show that the method achieves a weighted accuracy of 72.4% and an unweighted accuracy of 73.1%, outperforming current state-of-the-art approaches. These results further confirm the suitability of the proposed model in handling emotionally ambiguous speech. The design of the spatial–temporal auxiliary network enhances the model's ability to capture complex emotional features. Combined with the label correction strategy, it effectively utilizes previously neglected ambiguous samples, thereby improving overall recognition performance. Furthermore, the spatial–temporal fusion mechanism and the two-stage training strategy help the model inherit more stable emotion discrimination capabilities from a pre-trained model, thus reducing reliance on subjective annotations.



(a) Modify the label of S_C, $\lambda = 0$

(b) Modify the label of $S_B + S_C$, $\lambda = 0$

(c) Modify the label of S_C, $\lambda = 0.5$

(d) Modify the label of $S_B + S_C$, $\lambda = 0.8$

Fig. 8. T-SNE diagram.

Despite the aforementioned achievements, this study still has certain limitations. First, the label correction process is currently performed offline, which means the model may not be able to update in real time to adapt to newly emerging cases of emotional ambiguity in practical applications. Future work should explore online label correction mechanisms, enabling the model to continuously learn and optimize in dynamic environments. Second, although the soft label correction strategy demonstrates better performance than hard label correction, precisely adjusting the balancing factor to achieve optimal emotion learning objectives across different scenarios remains a challenge. Moreover, the generalizability of the selected balance factor λ across different datasets or tasks still requires further validation. Furthermore, with the emergence of more diverse and large-scale speech datasets, it will become increasingly important to explore more efficient spatial–temporal information fusion methods to further improve the accuracy and robustness of emotion recognition. These unresolved issues highlight the need for continued research efforts to advance fuzzy speech-based emotion recognition toward broader real-world applications.

CRediT authorship contribution statement

Chenquan Gan: Writing – review & editing, Writing – original draft, Validation, Methodology. **Daitao Zhou:** Writing – original draft, Methodology, Investigation. **Kexin Wang:** Writing – original draft, Software, Conceptualization. **Qingyi Zhu:** Supervision, Formal analysis. **Deepak Kumar Jain:** Supervision, Formal analysis. **Vitomir Štruc:** Supervision.

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

The authors are grateful to the anonymous reviewers and the editor for their valuable comments and suggestions. This work was supported by the Collaborative Tackling Project of the Yangtze River Delta Sci-Tech Innovation Community (Nos. 2024CSJGG01503, 2024CSJG G01500), Guangxi Key Research and Development Program (No. AB240 10317), and Jiangxi Provincial Key Laboratory of Electronic Data Control and Forensics (Jiangxi Police College) (No. 2025JXJYKFJJ002).

Data availability

Data will be made available on request.

References

Aftab, A., Morsali, A., Ghaemmaghami, S., Champagne, B., 2022. LIGHT-SERNET: A lightweight fully convolutional neural network for speech emotion recognition. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 6912–6916.

Ando, A., Kobashikawa, S., Kamiyama, H., Masumura, R., Ijima, Y., Aono, Y., 2018. Soft-target training with ambiguous emotional utterances for dnn-based speech emotion classification. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 4964–4968.

Ando, A., Masumura, R., Kamiyama, H., Kobashikawa, S., Aono, Y., 2019. Speech emotion recognition based on multi-label emotion existence model. In: INTERSPEECH. pp. 2818–2822.

Atmaja, B.T., Shirai, K., Akagi, M., 2019. Speech emotion recognition using speech feature and word embedding. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. APSIPA ASC, IEEE, pp. 519–523.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S., 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Lang. Resour. Eval. 42, 335–359.

Calderon-Uribe, S., Hernández, L.A.M., Guzman-Sandoval, V.M., Dominguez-Trejo, B., Albarrán, I.A.C., 2024. Emotion detection based on infrared thermography: A review of machine learning and deep learning algorithms. Infrared Phys. Technol. 105669.

Chang, P.-C., Chen, Y.-S., Lee, C.-H., 2024. IIOF: Intra- and inter-feature orthogonal fusion of local and global features for music emotion recognition. Pattern Recognit. 148, 110200.

Chatterjee, R., Mazumdar, S., Sherratt, R.S., Halder, R., Maitra, T., Giri, D., 2021. Real-time speech emotion analysis for smart home assistants. IEEE Trans. Consum. Electron. 67 (1), 68–76.

Chen, B., Cao, Q., Hou, M., Zhang, Z., Lu, G., Zhang, D., 2021. Multimodal emotion recognition with temporal and semantic consistency. IEEE/ACM Trans. Audio Speech Lang. Process. 29, 3592–3603.

Chen, M., He, X., Yang, J., Zhang, H., 2018. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Process. Lett. 25 (10), 1440–1444.

Chou, H.-C., Lee, C.-C., 2019. Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 5886–5890.

Del Coco, M., Leo, M., Carcagnì, P., Fama, F., Spadaro, L., Ruta, L., Pioggia, G., Distante, C., 2017. Study of mechanisms of social interaction stimulation in autism spectrum disorder by assisted humanoid robot. IEEE Trans. Cogn. Dev. Syst. 10 (4), 993–1004.

Fan, W., Xu, X., Cai, B., Xing, X., 2022. Isnet: Individual standardization network for speech emotion recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 30, 1803–1814.

Fujioka, T., Homma, T., Nagamatsu, K., 2020. Meta-learning for speech emotion recognition considering ambiguity of emotion labels. In: INTERSPEECH. pp. 2332–2336.

Gan, C., Wang, K., Zhu, Q., Xiang, Y., Jain, D.K., García, S., 2023. Speech emotion recognition via multiple fusion under spatial–temporal parallel network. Neurocomputing 555, 126623.

Gao, Y., Wu, D., Song, J., Zhang, X., Hou, B., Liu, H., Liao, J., Zhou, L., 2025. A wearable obstacle avoidance device for visually impaired individuals with cross-modal learning. Nat. Commun. 16 (1), 2857.

Hou, M., Zhang, Z., Cao, Q., Zhang, D., Lu, G., 2021. Multi-view speech emotion recognition via collective relation construction. IEEE/ACM Trans. Audio Speech Lang. Process. 30, 218–229.

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 Appl. 80 (16), 23745–23812.

Jalal, M.A., Milner, R., Hain, T., 2020. Empirical interpretation of speech emotion perception with attention based model for speech emotion recognition. In: Proceedings of Interspeech 2020. International Speech Communication Association (ISCA), pp. 4113–4117.

Kang, X., 2025. Speech emotion recognition algorithm of intelligent robot based on ACO-SVM. Int. J. Cogn. Comput. Eng. 6, 131–142.

Khurana, L., Chauhan, A., Naved, M., Singh, P., 2021. Speech recognition with deep learning. J. Phys.: Conf. Ser. 1854 (1), 012047.

Kim, Y., Kim, J., 2018. Human-like emotion recognition: Multi-label learning from noisy labeled audio-visual expressive speech. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 5104–5108.

Kumar, P., Jain, S., Raman, B., Roy, P.P., Iwamura, M., 2021. End-to-end triplet loss based emotion embedding system for speech emotion recognition. In: 2020 25th International Conference on Pattern Recognition. ICPR, IEEE, pp. 8766–8773.

Lecciso, F., Levante, A., Fabio, R.A., Caprì, T., Leo, M., Carcagnì, P., Distante, C., Mazzeo, P.L., Spagnolo, P., Petrocchi, S., 2021. Emotional expression in children with ASD: A pre-study on a two-group pre-post-test design comparing robot-based and computer-based training. Front. Psychol. 12, 678052.

Leo, M., Carcagnì, P., Signore, L., Benincasa, G., Laukkanen, M.O., Distante, C., 2022b. Improving colon carcinoma grading by advanced cnn models. In: International Conference on Image Analysis and Processing. Springer, pp. 233–244.

Leo, M., Farinella, G.M., Furnari, A., Medioni, G., 2022a. Machine vision for assistive technologies. Front. Comput. Sci. 4, 937433.

Li, X., Zhang, Z., Gan, C., Xiang, Y., 2023. Multi-label speech emotion recognition via inter-class difference loss under response residual network. IEEE Trans. Multimed. 25, 3230–3244.

Lotfian, R., Busso, C., 2019. Curriculum learning for speech emotion recognition from crowdsourced labels. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (4), 815–826.

Mao, S., Ching, P.-C., Lee, T., 2020. Emotion profile refinery for speech emotion classification. In: INTERSPEECH. pp. 531–535.

Mao, S., Ching, P., Lee, T., 2021. Enhancing segment-based speech emotion recognition by iterative self-learning. IEEE/ACM Trans. Audio Speech Lang. Process. 30, 123–134.

Ntalampiras, S., 2021. Speech emotion recognition via learning analogies. Pattern Recognit. Lett. 144, 21–26.

Park, S., Mark, M., Park, B., Hong, H., 2023. Using speaker-specific emotion representations in wav2vec 2.0-based modules for speech emotion recognition. Comput. Mater. Contin. 77 (1), 1009–1030.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32, 8026–8037.

Peng, Z., Dang, J., Unoki, M., Akagi, M., 2021. Multi-resolution modulation-filtered cochleagram feature for LSTM-based dimensional emotion recognition from speech. Neural Netw. 140, 261–273.

Quach, K.G., Le, N., Duong, C.N., Jalata, I., Roy, K., Luu, K., 2022. Non-volume preserving-based fusion to group-level emotion recognition on crowd videos. Pattern Recognit. 128, 108646.

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. ICASSP, IEEE, pp. 6907–6911.

Sharma, T., Diwakar, M., Arya, C., 2022. A systematic review on emotion recognition by using machine learning approaches. AIP Conf. Proc. 2481 (1).

Sharma, T., Diwakar, M., Singh, P., Arya, C., Lamba, S., Kumar, P., 2021a. A review on EEG based Emotion Analysis using Machine Learning approaches. In: 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering. UPCON, IEEE, pp. 1–6.

Sharma, T., Diwakar, M., Singh, P., Lamba, S., Kumar, P., Joshi, K., 2021b. Emotion Analysis for predicting the emotion labels using Machine Learning approaches. In: 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering. UPCON, IEEE, pp. 1–6.

Steidl, S., Levit, M., Batliner, A., Noth, E., Niemann, H., 2005. "Of all things the measure is man" automatic classification of emotions and inter-labeler consistency [speech-based emotion recognition]. In: Proceedings. (ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005, vol. 1, IEEE, pp. I/317–I/320.

Thimmaiah, S., et al., 2024. A review on emotion recognition from dialect speech using feature optimization and classification techniques. Multimedia Tools Appl. 1–34.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS '17, Curran Associates Inc., Red Hook, NY, USA, pp. 6000–6010.

Wang, L., Yang, J., Wang, Y., Qi, Y., Wang, S., Li, J., 2024. Integrating large language models (LLMs) and deep representations of emotional features for the recognition and evaluation of emotions in spoken english. Appl. Sci. 14 (9), 3543.

Yin, Y., Gu, Y., Yao, L., Zhou, Y., Liang, X., Zhang, H., 2021. Progressive co-teaching for ambiguous speech emotion recognition. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 6264–6268.

Yin, Y., Jing, L., Huang, F., Yang, G., Wang, Z., 2024. MSA-GCN: Multiscale adaptive graph convolution network for gait emotion recognition. Pattern Recognit. 147, 110117.

Yu, Z., Xu, X., Chen, X., Yang, D., 2019. Temporal pyramid pooling convolutional neural network for cover song identification. In: IJCAI. pp. 4846–4852.

Zepf, S., Hernandez, J., Schmitt, A., Minker, W., Picard, R.W., 2020. Driver emotion recognition for intelligent vehicles: A survey. ACM Comput. Surv. 53 (3), 1–30.

Zhang, S., Chen, M., Chen, J., Li, Y.-F., Wu, Y., Li, M., Zhu, C., 2021a. Combining cross-modal knowledge transfer and semi-supervised learning for speech emotion recognition. Knowl.-Based Syst. 229, 107340.

Zhang, S., Chen, J., Li, M., Li, T., Lu, P., Wang, Z., 2021b. Segment-level cross-modal knowledge transfer for speech sentiment analysis. In: 2021 IEEE 4th International Conference on Computer and Communication Engineering Technology. CCET, IEEE, pp. 243–247.

Zhang, B., Kong, Y., Essl, G., Provost, E.M., 2019. F-similarity preservation loss for soft labels: A demonstration on cross-corpus speech emotion recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 5725–5732.

Zhong, Y., Hu, Y., Huang, H., Silamu, W., 2020. A lightweight model based on separable convolution for speech emotion recognition. In: INTERSPEECH, vol. 11, pp. 3331–3335.

Zhou, Y., Liang, X., Gu, Y., Yin, Y., Yao, L., 2022. Multi-classifier interactive learning for ambiguous speech emotion recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 30, 695–705.

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. ICASSP, IEEE, pp. 7367–7371.