Improving Emotion Recognition from Ambiguous Speech via Spatio-Temporal Spectrum Analysis and Real-Time Soft-Label Correction

Chenquan Gan, Daitao Zhou, Qingyi Zhu, *Member, IEEE*, Xibin Wang, *Member, IEEE*, Deepak Kumar Jain, *Senior Member, IEEE*, and *Vitomir Štruc, *Senior Member, IEEE*

***Abstract*—Speech represents a fundamental medium for conveying human emotions and, as a result, speech-based emotion recognition (SER) systems have become pivotal in advancing human-computer interaction (HCI) across a range of applications. While significant progress has been made in speech emotion recognition over recent years, existing solutions still face several key challenges, in that they: (i) rely excessively on subjectively annotated (discrete) labels during training, (ii) often overlook the label ambiguity of speech samples that express more than one class of emotions, and (iii) underutilize unlabeled or ambiguous speech, for which typically a label distribution (or so-called soft labels) is available. To address these issues, we propose in this paper a novel SER model that explicitly handles ambiguous speech samples and overcomes the shortcomings outlined above. Central to our approach is a novel real-time soft-label correction strategy designed to refine the annotations assigned to ambiguous speech. The proposed model leverages both, (explicitly) labeled as well as ambiguous samples and applies the dynamic soft-label correction strategy alongside an enhanced inter-class difference loss function to iteratively optimize the label distributions during training. We theoretically demonstrate that our method is capable of approximating the true emotional distribution of speech even in the presence of label noise, suggesting that utilizing ambiguous speech samples without explicit emotion labels still contributes toward more effective emotion recognition. Furthermore, we integrate the representational power of convolutional neural networks (CNNs) with the contextual modeling capabilities of Wav2Vec 2.0 to enable a comprehensive extraction of spatio-temporal speech features. Experimental results on the IEMOCAP multi-label dataset confirm the effectiveness of our approach, achieving state-of-the-art performance with significant improvements in weighted accuracy (WA) and unweighted accuracy (UA) over competing methods.**

***Index Terms*—Speech emotion recognition, ambiguous speech, soft labels, real-time correction, spatio-temporal analysis**

*Corresponding author

This work is supported by the Guangxi Key Research and Development Program (No. AB24010317).

C. Gan and D. Zhou are with School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: gcq2010cqu@163.com; s220101216@stu.cqupt.edu.cn).

Q. Zhu is with School of Cyber Security and Information Law, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: zhuqy@cqupt.edu.cn).

X. Wang is with College of Data Science, Guizhou Institute of Technology, Guiyang 550025, Guizhou, China (e-mail: binxiwang@git.edu.cn)

D. Jain is with the Key Laboratory of Intelligent Control and Optimization for Industrial Equipment of the Ministry of Education, Dalian University of Technology, Dalian 116024, China (e-mail: dkj@ieee.org).

V. Štruc is with Faculty of Electrical Engineering, University of Ljubljana, Trzaska cesta 25, SI-1000 Ljubljana (e-mail: vitomir.struc@fe.uni-lj.si).

I. INTRODUCTION

Speech is one of the most natural ways of human communication that can directly express intentions and even emotional states. The process of using computer technology to analyze sound features from speech signals and infer the speaker's emotional state is commonly referred to as *speech emotion recognition* (SER) [1]–[3]. Speech emotion recognition has transitioned from a specialized research area to a significant component of human-computer interation (HCI) [4], capable of enhancing user experience in various application domains [5], ranging from call-center conversations and in-vehicle vehicle driving systems to smart-home application and smart healthcare among others [6].

Currently, most speech emotion recognition methods mainly rely on explicit (hard) labels for model training [7], [8], where each speech sample is assigned a single, discrete emotion category. While such methods have achieved notable success, they fall short in capturing the complexity of emotional expression in real-world scenarios, where speech often conveys multiple overlapping emotions. For instance, an utterance labeled as "sad" may concurrently also convey feelings of anger and disappointment [9]. This illustrates the inherent ambiguity and subjectivity present in emotional expression through speech [10]. Moreover, emotional perception in speech can also vary significantly across annotators due to individual differences in cultural background, gender, age, and other similar factors, suggesting that emotion perception is inherently subjective [11]. Consequently, SER methods that rely on single-label annotations not only fail to account for the ambiguity in emotional expression but also overlook the impact of subjective cognitive biases among annotators, which is one of the primary sources of label noise in emotion datasets [12].

To address the limitations of single-label annotation, recent research has increasingly explored multi-label approaches for speech emotion recognition [13], [14]. Multi-labeled methods leverage sets of emotion categories identified by multiple annotators to represent the presence of various emotions within a single utterance. While such approaches better capture the multifaceted nature of emotional expression, they still fall short in modeling the relative prominence of each emotion. In practice, speech often conveys a mixture of emotions, with one dominant emotion prevailing within the overall mixture. Traditional multi-label techniques are typically unable to represent these proportions effectively. To mitigate this issue, researchers

have introduced soft-label strategies [15], [16], where the distribution of annotator votes is used to assign weights to each emotion category, offering a more nuanced description of emotional content of the analyzed speech samples. However, such a soft-label approaches still heavily rely on the subjective judgments of a limited pool of annotators, and as a result, may introduce significant statistical noise and inconsistencies, which pose a considerable challenge for reliable training of speech emotion recognition (SER) models.

Furthermore, several studies have investigated the use of ambiguous speech samples for emotion recognition through multi-classifier interaction learning [17] and joint-learning [18] frameworks in an attempt to mitigate the limitations of multiple and soft labels by allowing the model to infer emotional distributions directly from the data. However, such approaches typically overlook speech samples that lack dominant emotions, i.e., samples that inherently carry the most ambiguity. In practice, the subjectivity of emotion perception and the ambiguity of emotional expression are most evident in these unlabeled or weakly labeled instances, where annotators are less likely to agree on a dominant emotion due to unclear affective cues [19]. This leads to inconsistencies in annotations and challenges in model training. Moreover, in real-world scenarios, it is common for speech utterances to lack a clearly dominant emotion that is agreed upon by a majority of annotators. As a result, existing methods that depend solely on speech samples with consensus labels fail to capture the full complexity of emotional ambiguity and do not adequately address the inherent uncertainty present in natural speech.

To address the above challenges, we propose in this paper a novel model for speech emotion recognition that explicitly targets the inherent uncertainty and ambiguities in emotional speech. Our approach simultaneously incorporates both unlabeled examples as well as samples with explicit discrete labels during training. Labeled samples provide supervision to guide the learning process, while ambiguous samples are iteratively refined using a soft label update mechanism in conjunction with an enhanced inter-class difference loss function. This enables the model to dynamically correct and learn from emotionally ambiguous data. Unlike existing SER methods that rely on static soft labels, sample reweighting, or offline label preprocessing, our model performs real-time, model-driven correction of ambiguous label distributions and jointly optimizes label refinement and representation learning within a single end-to-end framework. Moreover, we provide a theoretical analysis showing that this dynamic correction process guides the model toward the underlying true emotional distribution despite noisy or subjective annotations. Experimental results demonstrate that our method outperforms existing state-of-the-art approaches, particularly in its ability to effectively handle the ambiguity and subjectivity inherent in real-world speech emotion recognition. In summary, we make the following contributions in this paper:

1) We propose a novel speech emotion recognition (SER) approach, designed specifically to handle ambiguous speech, that addresses some of the key limitations of existing SER models, including the over-reliance on subjectively annotated labels, neglect of emotional distributions, and the underutilization of ambiguous samples lacking explicit (hard, consensus) emotion labels.
2) We introduce a real-time soft label correction strategy, theoretically validated to guide the model toward learning the true emotional distribution, even in the presence of label noise. This strategy offers a generalizable solution for other tasks involving noisy/ambiguous annotations.
3) We construct a comprehensive spatial-temporal feature extraction pipeline by combining the representational strengths of CNNs and Wav2Vec 2.0. Through a novel multi-level fusion mechanisms, our model effectively integrates time-frequency emotional cues for robust speech emotion recognition.

It should be noted that in this work, the term spatial specifically refers to the frequency dimension of the spectrogram, which can be treated analogously to the spatial axis in image processing when applying 2D-CNNs. The term temporal corresponds to the time dimension, capturing the dynamic evolution of speech signals.

II. Related work

In this section, we now discuss relevant prior work related to the research presented in this paper. For a more comprehensive coverage of the area of speech emotion recognition (SER), the reader is reffered to some of the excellent surveys available in the literature on this topic, i.e., [6], [20]–[23].

Early work in speech emotion recognition (SER) primarily relied on single-label annotations derived through majority voting (i.e., **consensus labels** hereafter) across annotators. To better exploit the spatio-temporal characteristics of speech signals, Ye *et al.* [24] introduced a time-aware bidirectional scaling network designed to integrate information from both past and future contexts, thereby enhancing the model's contextual representation capabilities. Li *et al.* [25] proposed a model that combines a spatiotemporal attention mechanism with a large-horizon learning strategy built on a CNN backbone, effectively localizing emotional regions while mitigating feature overlap. Building on this idea, Wu *et al.* [26] replaced attention modules with a capsule network to improve recognition accuracy by capturing hierarchical relationships among features. Gan *et al.* [27] developed a spatial-temporal network that integrates features through multiple fusion strategies, achieving strong performance in capturing both spatial and temporal cues.

While the studies outlined above have significantly advanced SER through various modeling strategies in the spatial, temporal, and joint domains, they generally do not address the issue of ambiguous emotional expressions in speech. To indirectly tackle this problem, Wang *et al.* [28] leveraged large language models (LLMs) to generate emotionally rich synthetic speech data based on a student speech dataset, while Yu *et al.* [29] applied an Attention-LSTM-Attention architecture to augmented datasets, aiming to enhance label robustness and alleviate ambiguity-related challenges. Several studies have also looked at the model architecure to reduce the impact of speech ambiguity on emotion recognition. Fan *et al.* [8], for example, introduced an Individual Standardized Network (ISNet) that addresses inter-individual variability in

emotional expression and, thus, aims to reduce confusion caused by personalized affective patterns. Yin *et al.* [30] proposed a progressive co-teaching strategy inspired by human and animal learning processes, where the model is trained on samples of increasing complexity (from simple to ambiguous) to mitigate the negative influence of emotionally ambiguous data on training stability. While these approaches acknowledge the challenges posed by emotional ambiguity, they still mostly rely on single-label annotations and, in turn, fail to capture the nuanced, multi-dimensional nature of emotional expression in speech. As a result, these methods suffer from inadequate feature representation, limiting their ability to fully model the complexity of real-world emotional speech.

Recognizing the limitations of single-label approaches, several studies have proposed alternative labeling strategies to better capture the inherent ambiguity in emotional speech. These efforts aim to model the coexistence of multiple emotions within an single utterance and account for inter-annotator variability in emotion perception. Li *et al.* [14], for instance, employed multiple labels as the ground truth for model training and introduced an inter-class difference loss function to reduce the similarity between emotion classes, thereby facilitating more accurate modeling of emotion distributions.

However, while such multi-label approaches can indicate the presence or absence of emotions, they often fail to distinguish between dominant and subordinate emotions in speech samples, which is critical when dealing with ambiguous speech. The seminal work of Steidl *et al.* [15] introduced soft labels, which reflect the proportion of annotator votes across emotion categories to capture perceptual ambiguity more precisely and addressed this problem. The authors also showed that the entropy of these soft labels closely aligns with the entropy of labels generated by an artificial/simulated annotator, suggesting that such soft labels effectively capture the annotators' perception of ambiguous emotions. Despite this advancement, most studies employing soft labels rely directly on observed annotator distributions, failing to account for the underlying subjectivity and variability in human emotion perception. As a result, these approaches risk embedding annotator bias into the training process, which may distort the model's ability to learn the true emotional content. To address this issue, Fayek *et al.* [11] proposed modeling individual annotators separately and integrating their outputs to generate more robust multi-label representations. Building on this line of work, Ando *et al.* [31] successfully incorporated speech samples lacking a single (discrete, consensus) label by modifying the soft-label representation to accommodate ambiguous expressions. Subsequently, Ando *et al.* [13] extended this framework by leveraging soft labels over multi-label annotations in repeated training runs, enabling a more comprehensive use of the full dataset, including emotionally ambiguous samples. Nevertheless, even these methods fall short of fully addressing annotator subjectivity, where individual perceptions may diverge significantly from the consensus, thus, introducing label noise that can mislead the model and compromise its ability to learn accurate emotional distributions.

In response to the challenges posed by label noise, Wang *et al.* [32] proposed a unified two-stage framework, consisting of labels noise modeling and correction training, to address different types of label noise in image classification tasks. Building on this concept, Liu *et al.* [33] introduced a validation-based mechanism to determine whether labels in the training set should be revised and demonstrated improved model robustness through selective label correction. Inspired by these label correction strategies, Fujioka *et al.* [34] applied a meta-learning approach that combines corrected labels with sample weight estimation to update noisy annotations.

In the context of speech emotion recognition, Mao *et al.* [35] observed that static soft labels fail to capture the dynamic nature of emotional expression, and, thus, proposed an emotion profile refinement strategy that generates soft labels in real time to better represent emotional evolution in speech. While these approaches mark significant progress in SER, they still largely neglect samples without consensus labeles, i.e., samples that are most representative of the ambiguity and uncertainty inherent in natural emotional speech. Ignoring such traning samples limits the ability of classification models to learn comprehensive and robust emotional representations. Beyond these general label correction strategies, recent works have introduced meta-learning and co-teaching frameworks directly into speech emotion recognition (SER). For example, Yin *et al.* [36] proposed a progressive co-teaching approach to mitigate the impact of emotionally ambiguous labels by iteratively exchanging reliable samples between peer networks. Chopra *et al.* [37] and Cai *et al.* [38] further explored meta-learning paradigms for low-resource or multi-task SER, showing that adaptive reweighting of uncertain annotations can improve robustness. However, these approaches mainly rely on sample reweighting or selection, while leaving the underlying label distributions unchanged, which may limit their effectiveness for inherently ambiguous emotional speech.

Inspired by the research discussed above, we propose in this paper a novel model for ambiguous speech emotion recognition. At the core of our approach is an innovative real-time soft label correction strategy, specifically designed to handle emotionally ambiguous speech samples. We theoretically and empirically demonstrate that this strategy effectively captures the underlying emotional distribution, even in the presence of noisy or unreliable labels. Furthermore, we leverage the representational power of convolutional neural networks (CNNs) alongside the contextual learning capabilities of Wav2Vec 2.0 to perform a detailed analysis of the speech signal's spatiotemporal characteristics. By employing multi-level feature fusion, our model efficiently integrates these representations, and allows for a robust and comprehensive understanding of speech emotions.

III. Soft-label Correction Strategy

A key component of the speech emotion recognition model proposed in this work is a **novel soft-label correction strategy**, designed specifically to account for the ambiguity of emotional speech and emotion perception from the annotators. In this section, we first motivate the need for soft-label correction and then theoretically show that soft-label correction leads to better ground truth and, in turn, better recognition models.

A. Problem Description

The variability in individual emotion perception typically leads annotators to assign different emotional labels to the same speech sample, resulting in samples that are annotated with multiple labels. In most existing speech emotion recognition (SER) methods, the prevailing approach is to adopt the emotion category with the highest number of votes as the final (consensus) label for the utterance. However, this practice introduces two key limitations:

1) Human speech frequently conveys a mixture of emotions, and majority-vote labeling fails to capture the nuanced and overlapping emotions inherent in natural speech.
2) Due to the subjectivity of human emotion perception, annotators often struggle to reach consensus, leading to uncertainty and ambiguity in soft-label distributions.

We propose a real-time soft label correction strategy that leverages the emotional features extracted through spatiotemporal neural networks. This strategy is designed to more accurately model the subtle emotional variations in speech while reducing the impact of noisy or ambiguous labels that could negatively affect model performance. Additionally, we provide a theoretical foundation to support the effectiveness of the proposed correction mechanism in learning more reliable emotional representations.

B. Real-time Soft-Label Correction Strategy

The proposed real-time soft label correction strategy consists of two key components. First, it employs **a dynamic soft label update mechanism** to iteratively refine the soft labels associated with ambiguous speech segments. This approach reduces overreliance on potentially noisy ground truth annotations and enables more accurate emotion representation. Second, the strategy incorporates **a joint loss function specifically designed to support real-time label refinement**. This loss function combines a standard cross-entropy term to ensure stable model training with an enhanced inter-class difference loss, which encourages greater discrimination between emotion categories. These components jointly improve the quality of soft labels and enhance overall model performance.

Soft-Label Updating Mechanism. The soft label updating process for ambiguous speech samples is formulated as a weighted combination of the observed (annotator-provided) labels and the model-generated predictions. This design is based on two key observations: (i) deep learning models excel at capturing complex patterns in data and can produce reliable emotion predictions from the provided speech samples when trained effectively, and (ii) the observed annotator provided labels, though potentially noisy, are generally close approximations of the true emotional states. Thus, combining both sources through a weighted summation enables the model to benefit from the provided (prior, potentially noisy) annotations while gradually incorporating its own learned representations and thereby improving label quality over time.

Importantly, the soft label refinement process is applied exclusively to ambiguous speech samples. For unambiguous samples, annotated with a consensus label, the original labels are retained without modification. This is based on the premise that clearly expressed emotions are less prone to annotator disagreement and are thus less likely to contain labeling errors. From this perspective, we classify all speech samples in a given dataset into two disjoint subsets:

- **Clear samples (S_A)**, i.e., samples with (single) discrete consensus labels that exhibit strong agreement across annotators w.r.t. the expressed emotion.
- **Ambiguous samples (S_B)**, i.e., samples associated with multiple labels that reflect disagreement among annotators and emotional uncertainty.

Let S denote the full set of speech samples, such that $S = S_A \bigcup S_B$ and $S_A \bigcap S_B = \varnothing$. Furthermore, let N_1, N_2, and N be the number of clear samples in S_A, ambiguous samples in S_B, and samples in the complete set S, respectively. In a supervised K-class classification task, the labels corresponding to the set of clear samples S_A are single (consensus) labels $y_{cons}^{x^i}$, defined as the emotion categories considered by the majority of annotators, i.e.:

$$y_{cons}^{x^i} = (y_1^{x^i}, y_2^{x^i}, \cdots, y_K^{x^i}), y_j^{x^i} \in \{0, 1\}, \tag{1}$$

where $\sum y_j^{x^i} = 1, i \in \{1, \cdots, N_1\}$, $x^i \in S_A$ denotes the i^{th} speech sample. A value of $y_j^{x^i} = 1$ indicates that the majority of annotators assigned the j^{th} emotion class to sample x^i, whereas a value of 0 indicates otherwise. This one-hot encoding reflects the assumption that each clear sample is associated with a single dominant emotion.

However, due to the limited number of annotators and the inherently subjective nature of emotion perception, many samples exhibit label ambiguity. These ambiguous samples (in S_B) are characterized by inconsistent annotator opinions and are labeled using multi-label representations $y_{multi}^{x^i}$, i.e.:

$$y_{multi}^{x^i} = (t_1^{x^i}, t_2^{x^i}, \cdots, t_K^{x^i}), t_j^{x^i} \in \{0, Z_+\}, x^i \in S_B, \tag{2}$$

where, $i \in \{1, \cdots, N_2\}$, Z_+ denotes the set of positive integers, and N_2 is the number of ambiguous samples in S_B. Here $t_j^{x^i}$ represents the number of annotators who assigned the j^{th} emotion label to speech sample x^i, capturing the distribution of annotator responses. Since this form is not directly compatible with standard model training objectives, it is often converted to a binary multi-label form, as in [14]:

$$y_{multi}^{x^i} = (s_1^{x^i}, s_2^{x^i}, \cdots, s_K^{x^i}), s_j^{x^i} \in \{0, 1\}, x^i \in S_B, \tag{3}$$

where $i \in \{1, \cdots, N_2\}$, and $s_j^{x^i} = 1$ (or 0) indicates whether at least one annotator perceived (or did not perceive) the presence of the j^{th} emotion in sample x^i. It is evident that Eq. (1) is a special case of Eq. (3), i.e., when $\sum s_j^{x^i} = 1$, both formulations are equivalent.

Nevertheless, this multi-label representation does not reflect the relative prevalence of each emotion in the given speech sample. To address this limitation, soft-labels $y_s^{x^i}$, which capture the proportion of annotator votes for each emotion class, are commonly used instead. Here, soft-labels are defined as a probability distribution over the K emotion classes: i.e.:

$$y_s^{x^i} = (p_1^{x^i}, p_2^{x^i}, \cdots, p_K^{x^i}), p_j^{x^i} \in [0, 1], \tag{4}$$

where $\sum p_j^{x^i} = 1$, $i \in \{1, \cdots, N_2\}$, and each $p_j^{x^i}$ is computed as:

$$p_j^{x^i} = \frac{t_j^{x^i}}{\sum_m^K t_m^{x^i}}, \tag{5}$$

with $t_j^{x^i}$ denoting the number of annotators who assigned the j^{th} emotion label to sample x^i. This formulation captures the proportion of votes per class and reflects the perceived emotional distribution. However, due to the limited number of annotators and the subjectivity of emotion perception, these soft labels may still contain inaccuracies and may not fully reflect the true emotional composition of the speech sample. To alleviate this problem, we introduce a *soft-label update mechanism* that generates corrected soft-labels $y_c^{x^i}$ as follows:

$$y_c^{x^i} = \begin{cases} y_{cons}^{x^i}, x^i \in S_A, \\ (1-\alpha) y_s^{x^i} + \alpha y_g^{x^i}, \alpha \in [0,1), x^i \in S_B, \end{cases} \tag{6}$$

where $y_c^{x^i}$ denotes the corrected labels for sample x^i, $y_s^{x^i}$ stands for the original soft label and $y_g^{x^i}$ is the label predicted by the emotion recognition model for the same sample. The parameter $\alpha \in [0, 1)$ is a correction coefficient that controls the contribution of the model-generated prediction to the final soft label. For clear samples ($x^i \in S_A$) the original consensus label is retained, while for ambiguous samples ($x^i \in S_B$), the corrected label is obtained through a weighted combination of the original soft label and the model-generated label.

The Joint Loss Function. The overall training objective for our model is defined by a joint loss function L, which combines two components: a cross-entropy loss L_{cor} for optimizing model predictions, and an enhanced inter-class difference loss L_{Ic} designed for real-time soft-label correction.

The cross-entropy loss L_{cor} quantifies the divergence between the predicted emotion distributions and the target labels and is widely used in supervised classification tasks. It is formulated as:

$$L_{cor} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} y_j^{x^i} \log(p_j^{x^i}), x^i \in S, \tag{7}$$

where where the predicted probability $p_j^{x^i}$ for class j is computed via the softmax function:

$$p_j^{x^i} = \frac{\exp(o_j^{x^i})}{\sum\limits_{k=1}^{K} \exp(o_k^{x^i})}. \tag{8}$$

Here, N denotes the total number of training samples, $o_j^{x^i}$ is the model output (logit) for class j, $y_j^{x^i} \in [0, 1]$ is the target label (either consensus or soft label) and $p_j^{x^i}$ enotes the predicted probability for emotion class j in speech sample x^i.

To support the dynamic correction of soft labels for ambiguous samples, an additional loss term L_{Ic} is introduced. This enhanced inter-class difference loss acts as a regularization mechanism, guiding the model-generated labels $y_g^{x^i}$ to remain consistent with the observed soft labels $y_s^{x^i}$, while mitigating overfitting, particularly when the dataset contains a large proportion of ambiguous samples. Unlike the inter-class difference loss originally proposed in [14], which assumes binary labels, the enhanced version here accommodates soft labels $y_j^{x^i} \in \{0, 1\}$. The loss is defined as:

$$L_{Ic} = \begin{cases} 0, & x^i \in S_A, \\ \dfrac{1}{N_2} \displaystyle\sum_{i=1}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \Big(\exp(u) - 1 \Big), & x^i \in S_B, \end{cases} \tag{9}$$

where

$$u = \max\Big\{0, \Big(1 - y_j^{x^i}\Big)\big(p_j^{x^i} + \beta\big) - y_j^{x^i} p_k^{x^i}\Big\}. \tag{10}$$

Here, $x^i \in S_B$ denotes an ambiguous speech sample, $y_j^{x^i} \in [0, 1]$ is the corrected soft label for class j, $p_j^{x^i}$ is the corresponding predicted probability, and β is a margin control coefficient that regulates the separation between class predictions. This loss encourages higher inter-class discrimination, especially in ambiguous contexts. Given the bounded nature of $y_j^{x^i}$, β, and $p_k^{x^i}$, the value of L_{Ic} is constrained to the interval $\big[0, k^2(exp(2) - 1)\big)$.

The final joint loss L used for training is the sum of the two components, i.e.:

$$L = L_{cor} + L_{Ic}. \tag{11}$$

C. Feasibility Analysis

In the previous section, we introduced the real-time soft label correction strategy. In this section we now present a theoretical analysis and a formal proof to validate the effectiveness of the proposed strategy.

Proof Overview. The main idea of the proof is to demonstrate that, under a reasonable setting of the correction coefficient α and margin control coefficients β, the model parameters θ'_M obtained through the real-time soft label correction strategy can converge to a region near the optimal parameters θ_t, which are obtained using the true label distribution. In other words, the strategy effectively guides the training process toward the underlying ground-truth distribution despite the presence of noisy or ambiguous labels. To support this claim, we first present two foundational lemmas:

Lemma 1. *There exists a two-layer neural network with ReLU (Rectified Linear Unit) activation functions and $2n + d$ parameters that can represent any function over a dataset of n samples in a d-dimensional space [39].*

Lemma 1 implies that a sufficiently parameterized deep neural network can approximate any label distribution, regardless of the choice of loss function.

Lemma 2. *Assuming a neural network with sufficient capacity, for any loss function, L, the training dynamics follow the convergence path: $f_{\theta_0} \rightarrow f_{\theta_t} \rightarrow f_{\theta_*}$, where f_{θ_0} is the is the model initialized with random parameters θ_0, f_{θ_t} is is the model trained with true (clean) labels, and, and f_{θ_*} corresponds to the model trained with observed (possibly noisy) labels [33].*

Lemma 2 suggests that while training may begin with random initialization, convergence paths under true labels and observed labels are both attainable and related.

Building upon these lemmas, the following theorem provides the basis for analyzing the behavior of the proposed correction strategy:

Theorem 1. *Let θ_t denote the model parameters after t optimization steps using true labels, and let θ_M denote the parameters after M steps ($M > t$) using observed, potentially noisy labels. Then, the final parameters θ_M lie with a neighborhood of θ_t, i.e., $\theta_M \in [\theta_t - R_M, \theta_t + R_M]$, where R_M denotes the radius of the neighborhood that quantifies the deviation caused by label noise.*

Theorem 1 implies that if the label correction mechanism effectively reduces the noise in observed labels (by leveraging model-generated predictions and controlling the correction dynamics via α and β) then the model can be guided to converge toward a representation close to that obtained under true labels.

Proof: According to Lemma 1, a sufficiently parameterized model can fit any label distribution and will converge after at most M optimization steps. Lemma 2 further states that the model, initialized at θ_0 reaches the parameter state θ_t after t iterations when trained on true grount truth labels y_{true}. Let us consider the standard parameter update rule in deep learning. At iteration $t+1$ the model parameters are updated as follows:

$$\theta_{t+1} = \theta_t - lr \cdot g(\theta)\Big|_{\theta=\theta_t}, \tag{12}$$

where lr is the learning rate, and $g(\theta) = \frac{\partial L}{\partial \theta}$ denotes the gradient of the loss function L with respect to the parameters. By recursively applying this update rule, the model parameters at iteration M can be expressed as:

$$\theta_M = \theta_t - lr \sum_{i=0}^{M-t-1} g(\theta)\Big|_{\theta=\theta_{t+i}}. \tag{13}$$

Let us define the average gradient during the convergence process as: $\psi = \frac{1}{M-t}\sum_{i=0}^{M-t} \frac{\partial L}{\partial \theta_{t+i}}$. Since this expression is bounded for all $t + i \in [t, M-1]$, the average gradient ψ satisfies:

$$\min_{t\le i\le M-1}\{g(\theta)\Big|_{\theta_i}\} \le \psi \le \max_{t\le i\le M-1}\{g(\theta)\Big|_{\theta_i}\}. \tag{14}$$

From Eq. (14), it follows that the final model parameters θ_M lie within a neighborhood around the true label-trained parameters θ_t with a radius defined by the learning rate and maximum gradient norm:

$$\theta_M \in \{\theta | \theta_t - R_M \le \theta \le \theta_t + R_M\}, \tag{15}$$

where $R_M = lr \max\{|g(\theta_t)|, \cdots, |g(\theta_{M-1})|\}$.

According to Theorem 1, when the real-time soft label correction strategy is applied, the gradient $g(\theta)$ is defined as:

$$g(\theta) = \begin{cases} \sum_{i=1}^{N}\sum_{j=1}^{K}\left(\frac{m}{N_2}\sum_{k=1}^{K} v_{ijk}\exp(u_{ijk}) - \frac{(y_c)_j^{x^i}}{Np_j^{x^i}}\right), \\ m = 0, x^i \in S_A, \\ m = 1, x^i \in S_B, \\ (y_c)_j^{x^i} = (y_{cons})_j^{x^i}, x^i \in S_A, \\ (y_c)_j^{x^i} = (1-\alpha)(y_s)_j^{x^i} + \alpha p_j^{x^i}, x^i \in S_B, \\ u_{ijk} = \left(1-(y_s)_j^{x^i}\right)(p_j^{x^i} + \beta) - (y_s)_j^{x^i} p_k^{x^i}, \\ v_{ijk} = \left(1-(y_s)_j^{x^i}\right) - (y_s)_k^{x^i}\frac{\partial p_k^{x^i}}{\partial p_j^{x^i}}. \end{cases} \tag{16}$$

Eq. (16) indicates that the gradient and, hence, the convergence behavior of the model is directly influenced by the correction coefficient α and the margin control coefficient β. These parameters modulate the extent to which model predictions correct or align with the soft labels, thereby controlling the size of the radius R_M in which convergence occurs.

Theorem 2. *Let $M\left(x^i|\theta\right)$ be a training network converging at iteration M iteration using observation labels and yielding model parameters θ_M. Suppose the observation labels are subsequently corrected and the network is retrained. Upon convergence after M iterations with corrected labels, the resulting parameters θ'_M lie within a neighborhood of θ_M, i.e., $\theta'_M \in [\theta_M - R'_M, \theta_M + R'_M]$, where R'_M is the radius of the neighborhood.*

Proof: According to Theorem 1, the parameters θ_M, obtained after training with observation labels, lies within a neighborhood of the true-label parameters θ_t, such that

$$\begin{cases} \theta_M = \theta_t + n_1 R_M, n_1 \in [-1, 1], \\ \theta'_M = \theta_t + n_2 R^c_M, n_2 \in [-1, 1]. \end{cases} \tag{17}$$

where R_M and R^c_M are the neighborhood radii associated with training under observation labels and corrected labels, respectively. n_1 and n_2 are scaling constants reflecting directional distance. It follows that θ'_M can be expressed relative to θ_M as:

$$\theta'_M = \theta_M + nR'_M = \theta_M + n_2 R^c_M - n_1 R_M, n \in [-1, 1]. \tag{18}$$

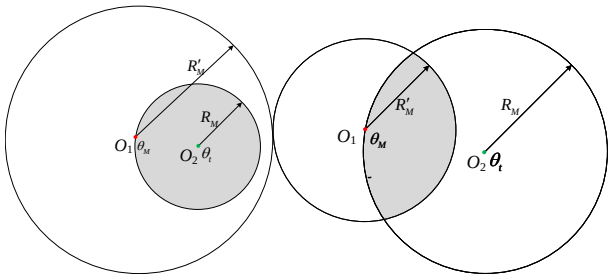
Thus, $\theta'_M \in [\theta_M - R'_M, \theta_M + R'_M]$, confirming that the corrected-label solution lies in the vicinity of the original observation-label solution, where n is a constant and R'_M is the neighborhood radius.

Fig. 1 illustrates two possible relationships between the neighborhoods defined by Theorems 1 and 2. Here:

- O_1 represents the region around θ_M wiht radius R'_M, within which θ'_M is located.
- O_2 represents the region centered at θ_t, defined by the original radius R_M.

In both scenarios, the shaded area denotes regions where θ'_M is closer to θ_t than θ_M. Since observation labels are fixed for a given dataset, the size of neighborhood O_2 remains approximately constant for a given model. Let's define the proportion of the shaded ara S to the are of O_1, denoted as S_{O_1}, as:

$$P = \frac{S}{S_{O_1}}. \tag{19}$$



(a) Relationship 1 (b) Relationship 2

Fig. 1: **Illustration of the relationship** between the two neighborhoods identified in Theorem 1 and Theorem 2

A higher value of P indicates a greater probability that θ'_M lies closer to the true-label parameters θ_t than θ_M. For a constant R_M, the probability P increases as R_M decreases. The correction radius R'_M is influenced by the correction coefficient α and the margin control coefficients β. This control relationship can be expressed as:

$$\big(\alpha, \beta\big) \rightarrow \big(R_M, R^c_M\big) \rightarrow R'_M \rightarrow P \tag{20}$$

where $\rightarrow$ denotes the directional influence of the preceding variables on the subsequent ones.

In summary, by appropriately tuning α and β, it is possible to reduce the neighborhood radius R'_M, thereby guiding θ'_M closer to θ_t than θ_M, which establishes the theoretical feasibility and effectiveness of the proposed real-time soft label correction strategy.

IV. The Proposed Method

In this section, we now present the main contribution of this work, i.e., our **novel model for speech emotion recognition**, designed specifically to handle ambiguous speech samples. We start the section with a brief high-level description of the proposed approach and then proceed with the description of the individual components.

A. Overview

Speech emotion recognition remains challenging due to label ambiguity, subjective annotation noise, and the complex temporal–spectral dynamics of speech signals, as discussed in the Introduction. To address these challenges, we propose an enhanced ambiguous speech emotion recognition model, illustrated in Figure 2. The model architecture consists of five main components that are described in detail in the following sections, i.e.: (i) a spatial feature extraction module (§IV-B), a temporal module (§IV-C), a multi-level fusion module (§IV-D), a real-time soft label correction module (§IV-E), and a classification module that determine the final emotion class of the input speech sample (§IV-F).

B. The Spatial Module

We use Log Mel-Filter Bank (Fbank) features as the input, which form a conventional time–frequency spectrogram widely adopted in speech and audio processing. This representation provides a two-dimensional structure suitable for CNN-based modeling. The resulting Fbank feature maps are fed into a convolutional network, i.e., the spatial domain module, for hierarchical emotion feature extraction. The architecture of this module is illustrated in Figure 3.

To align with the input format expected by 2D convolutional layers, we first reshape the extracted Fbank features into a three-dimensional tensor $X_{\text{in}} \in \mathbb{R}^{1\times f\times M}$, where '1' denotes the input channel dimension required by 2D-CNNs, f denotes the number of frames and M represents the number of Mel filter banks. This representation preserves the two-dimensional time-frequency structure of speech, which is essential for learning emotion-relevant spatial patterns.

Given that each Fbank feature map captures spectral information over a sequence of frames, we design a **parallel convolutional structure** to extract features along different axes of the time-frequency domain. Specifically, we apply three parallel convolutional operations: one emphasizing temporal patterns, another focusing on spectral characteristics, and a third capturing joint time-frequency dependencies. Unlike conventional CNN-based SER approaches that apply uniform 2D kernels directly on spectrograms, our parallel design explicitly separates temporal, spectral, and joint Spatio-temporal modeling through distinct kernel shapes (11×1, 1×9, and 5×5). This decomposition enables the model to capture complementary emotional cues from different perspectives before fusing them into a unified representation. The output of this parallel convolution block is defined as:

$$X_c = C\big(Conv^{1a}(X_{in}),\ Conv^{1b}(X_{in}),\ Conv^{1c}(X_{in})\big), \tag{21}$$

where $X_c \in \mathbb{R}^{24\times \frac{f}{2}\times \frac{M}{2}}$, and $C(\cdot)$ denotes channel-wise concatenation. The operations $Conv^{1a}(\cdot)$, $Conv^{1b}(\cdot)$, and $Conv^{1c}(\cdot)$ correspond to convolution layers designed to capture temporal, spectral, and joint temporal-spectral features, respectively.

To further deepen the representation and aggregate mid-level features, we pass X_c through five consecutive convolutional layers using 3×3 kernels. These layers progressively refine the feature maps while reducing spatial resolution. Notably, no downsampling is applied after the final convolution to preserve critical information. The resulting feature map X_d is given by:

$$X_d = Conv^5(X_c), \quad X_d \in \mathbb{R}^{96\times \frac{f}{32}\times \frac{M}{32}}, \tag{22}$$

where $Conv^5(\cdot)$ represents the stacked five-layer convolutional block. Finally, we apply global average pooling (GAP) to compress the spatial feature map into a fixed-dimensional embedding $X_s \in \mathbb{R}^{96}$, which serves as the output of the spatial domain module:

$$X_s = GP(X_d), \quad X_s \in \mathbb{R}^{96}, \tag{23}$$

where $GP(\cdot)$ denotes global average pooling. This operation summarizes the learned spatial features into a compact representation suitable for downstream fusion and classification.

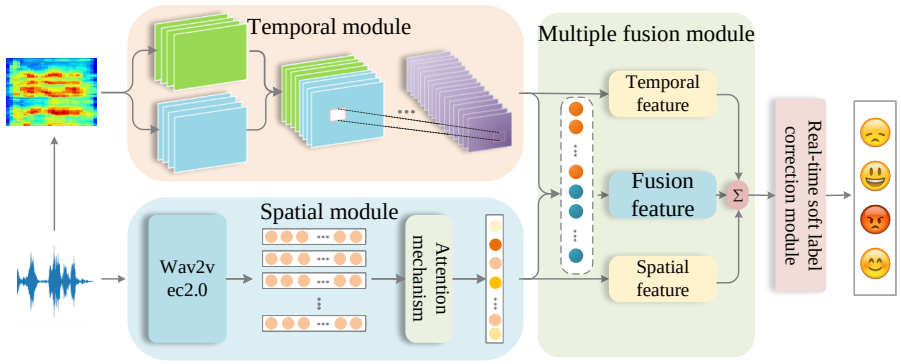


Fig. 2: **High-level overview** of the proposed ambiguous speech emotion recognition model. The model integrates a spatio-temporal feature extraction module based on deedicated CNN and Wav2Vec 2.0, followed by multi-level fusion to capture complementary emotional cues. A real-time soft label correction module is also designed to dynamically refine ambiguous labels during training using a combination of cross-entropy and enhanced inter-class difference losses.

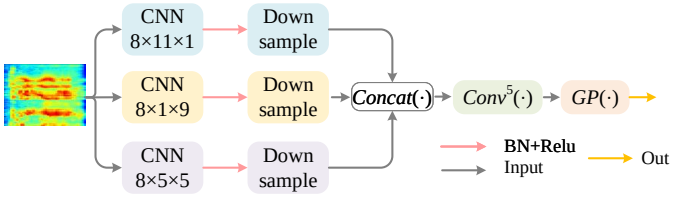


Fig. 3: **Structure of the spatial module.** The module operates on a time–frequency representation of the input speech. It employs three parallel convolutional branches with kernel sizes of 11×1, 1×9, and 5×5, respectively, corresponding to temporal, spectral, and joint Spatio-temporal modeling. Each branch outputs 8 feature maps, where "8" denotes the number of convolutional filters (channels) used to capture diverse patterns.

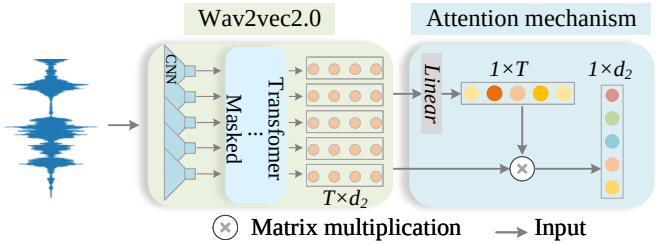


Fig. 4: **Structure of temporal module.** We use Wav2Vec 2.0 to extract high-level temporal features from raw speech waveforms using convolution, masking, and transformer blocks.

C. The Temporal Module

Speech waveforms are continuous-time signals, and crucial emotion-related information is often embedded in their temporal variations. Accurately modeling these dynamic is thus essential for robust speech emotion recognition. To this end, we leverage Wav2Vec 2.0, a powerful self-supervised representation learning framework pre-trained on large-scale speech corpora. Its strong contextual encoding capabilities provide valuable a priori knowledge, allowing it to effectively extract rich temporal emotion features from raw waveforms.

The temporal module, illustrated in Figure 4, is employed to extract such features directly from the input waveform.

Given an input temporal speech waveform x_i we first apply a series of one-dimensional convolutional layers to produce a feature map $G \in \mathbb{R}^{T \times d_1}$, where T corresponds to the number of frames (which varies with input length), and d_1 is the feature dimension. A linear transformation is then applied to project G into a new representation space:

$$G' = Linear(G), \quad G' \in \mathbb{R}^{T \times d_2}, \tag{24}$$

where d_2 denotes the dimension of the transformed features. To improve robustness and simulate masked language modeling objectives during fine-tuning, a masking operation is applied to G', yieldig:

$$F_1 = Mask(G') \in \mathbb{R}^{T \times d_2}, \tag{25}$$

where $Mask(\cdot)$ randomly masks segments of the sequence. The masked features are then passed through the Transformer encoder of Wav2Vec 2.0:

$$F_2 = Transformer(F_1) \in \mathbb{R}^{T \times d_2}. \tag{26}$$

To obtain a fixed-length temporal feature vector, we apply another linear transformation followed by a weighted aggregation using matrix multiplication:

$$F_2' = Linear(F_2), \quad F_2' \in \mathbb{R}^{1 \times T}, \tag{27}$$

$$F_t = F_2' \otimes F_2, F' \in \mathbb{R}^{1 \times T}, F_t \in \mathbb{R}^{1 \times d_2}, \tag{28}$$

where $\otimes$ denotes matrix multiplication. The resulting vector F_t represents the aggregated temporal features of the input speech that captures high-level emotional information across the entire sequence.

D. The Multi-level Fusion Module

Since analyzing temporal or spatial features in isolation limits the model's ability to fully capture the multifaceted nature of speech signals, we introduce a **multi-level fusion module** to integrate both temporal and spatial emotion representations. This design allows the model to comprehensively

learn emotional cues that are distributed across time and frequency dimensions.

First, we obtain an intermediate fused representation F_{st} by concatenating the spatial emotion feature $X_s \in \mathbb{R}^{d_1}$ and temporal emotion feature $F_t \in \mathbb{R}^{d_2}$, followed by a series of fully connected layers with ReLU activations:

$$F_{st} = \delta\left(\delta\left(C\left(X_s, F_t\right) W_f^1 + B_f^1\right) W_f^2 + B_f^2\right) W_f^3 + B_f^3, \quad (29)$$

where $C(\cdot)$ denotes the concatenation operation, $\delta(\cdot)$ is the ReLU activation, and W_f^i, B_f^i (for $i = 1, 2, 3$) are trainable weight and bias parameters. This operation maps the concatenated features into a joint representation space that enables interaction between the temporal and spatial modalities.

Next, to produce the final fused representation F'_{st}, we perform a soft aggregation by applying a softmax function to the sum of three components: linear projections of the spatial and temporal features, and the intermediate fused vector:

$$F'_{st} = \text{softmax}\left(Linear_s(X_s) + Linear_t(F_t) + F_{st}\right), \quad (30)$$

where $Linear_s(\cdot)$ and $Linear_t(\cdot)$ are learnable linear transformations specific to the spatial and temporal features, respectively, and $F'_{st} \in \mathbb{R}^{d_t}$ is the final fusion representation of dimension d_t. The presented multi-level fusion mechanism allows the model to not only learn combined feature representations but also dynamically adjust the contribution of each modality during prediction.

E. The Real-time Soft-Label Correction Module

In Section III, we described the real-time soft label correction strategy in detail and provided a theoretical justification for its effectiveness. In this section, we now apply the proposed strategy directly within our training framework. The complete process is outlined in Algorithm 1.

Algorithm 1 Real-Time Soft Label Correction Strategy

Input: Speech sample set $S = S_A \cup S_B$, where $S_A \cap S_B = \varnothing$
Output: Real-time corrected label $y_j^{x^i}$

```
Fine-tune model M_1(x^i | θ) using clear speech samples S_A to obtain M_1p
Integrate fine-tuned model M_1p with randomly initialized model M_2(x^i | θ) to form model M_p
Randomly shuffle samples and input speech x^i ∈ S into model M_p
for each sample i = 1 to N do
    Extract features: F'_st = M_p(x^i)
    if x^i ∈ S_B then
        Update label: y_c^{x^i} = Ω(y_s^{x^i}, F'_st), where Ω(·) is the label update function
    else
        Assign original label: y_c^{x^i} = y_cons^{x^i}
    end if
    Compute loss: Loss = L(F'_st | M_p, y_c^{x^i})
end for
```

Notation clarification: For clarity, the main notations in Algorithm 1 are summarized as follows: θ denotes the model parameters; N is the total number of training samples; $y_j^{x^i}$ represents the soft-label probability of sample x^i for class j; and H_{st}^i indicates the sequence features extracted from sample x^i.

In this algorithm, we begin by fine-tuning the network $M_1(x^i \mid \theta)$ using clear-labeled samples from S_A, resulting in model M_{1p}, which captures reliable emotional tendencies. This step mitigates the potential negative impact of ambiguous speech during training. The complete dataset S is then passed through the composite model M_p. For each input $x^i \in S$, if the sample belongs to the ambiguous set S_B, its label is updated using the real-time soft label correction function $\Omega(\cdot)$. If the sample belongs to the clear set S_A, the original (hard) consensus label is retained. Finally, the model prediction and the corrected label are used to compute the loss $L(F'_{st} \mid M_p, y_c^{x^i})$.

F. Classification

The final classification component of the proposed model employs a multi-layer fully connected neural network. This structure is designed to perform fine-grained learning over the distributed emotional features and effectively map them to discrete emotion categories. The classification process is formulated as follows:

$$\hat{\mathbf{y}}(x^i \mid M_p) = \text{softmax}(\hat{\mathbf{y}}^{x^i}), \quad (31)$$

where $\hat{\mathbf{y}}(x^i \mid M_p)$ denotes the predicted probability distribution over the emotion classes for the input speech sample x^i, and $\hat{\mathbf{y}}^{x^i}$ represents the output of the final fully connected layer based on the features extracted by the model M_p. The use of the softmax function ensures that the output forms a valid probability distribution over all emotion categories and facilitates effective multi-class classification.

V. Experiments

In this section, we conduct a rigourous experimental evaluation of the proposed speech emotion recognition model and report results that: (i) compare our approach to competing state-of-the-art methods from the literature, (ii) explore the impact of various model components through an ablation study, (iii) investigate the impact of the hyperparameters α and β on classification performance, (iv) study the impact of the model's learning objective, and (v) analyze the generated embedding space.

A. Experimental Setup

The proposed model is implemented using PyTorch, utilizing a 64-bit Ubuntu 22.04 system equipped with an NVIDIA RTX 3090 GPU for training and testing.

Table I outlines the parameter settings for the training procedure. Unless otherwise specified, the soft-label correction (SLC) phase adopts the same optimizer and fixed learning rate (1e-5) as the fine-tuning stage. The masking probability follows the default configuration of the HuggingFace Wav2Vec2-base-960h model. The training begins by fine-tuning the Wav2Vec 2.0 model using the set of clear samples S_A, resulting in an intermediate model M_{1p}. This step initializes the model with a domain-specific emotional representation aligned with the IEMOCAP dataset. Subsequently, M_{1p} is integrated with the null-domain module M_2 to construct the complete soft-label correction model M_p, which is then used to

TABLE I: **Parameter settings** utilized during the training procedure of the proposed speech emotion recognition model.

Name	Value
α	0.3
ε	$1e-5$
β	0.1
Batch size	16
Optimizer	Adam
Learning rate	$1e-5$
Weight decay	0.0001
Max epoch	100
Early stopping patience	16

learn the underlying emotional distribution across all samples, including ambiguous ones. To address class imbalance caused by differing frequencies of emotion categories, the inverse class frequency is employed as the class-wise weight in the loss function during training. This ensures that minority emotion classes are not underrepresented in the learning process.

For evaluation, a Leave-One-Speaker-Out (LOSO) cross-validation strategy is adopted, following standard practice in speech emotion recognition research [14], [18]. In this setup, the speech data from one speaker is reserved as the validation set in each fold, while the remaining data is used for training.

To ensure a comprehensive performance evaluation, two commonly used metrics are reported for our experiments: Weighted Accuracy (WA) and Unweighted Accuracy (UA) [40], [41]. WA reflects the overall classification accuracy across all utterances, accounting for class distribution, while UA measures the average accuracy across all emotion classes, treating each class equally regardless of frequency.

B. Datasets

Given that the primary goal of this study is to address the challenge of label noise and improve the model's ability to learn genuine emotional distributions from potentially noisy annotations, the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [42] is selected for experimentation, as it is (to the best of our knowledge) the only publicly available multi-label dataset suitable for our experiments.

The IEMOCAP dataset consists of recordings of five dyadic sessions involving 10 actors (5 male, 5 female) in two types of scenarios: scripted dialogues and spontaneous improvisations. The dataset includes rich multimodal data such as audio, video, and text, captured during interactive emotional exchanges between participants. To investigate both the ambiguity inherent in emotional expression and the subjectivity of emotional perception, we follows the experimental protocol of [34], which utilizes samples from both improvised and scripted settings. Four commonly studied emotion categories, i.e., anger, happiness, sadness, and neutrality, are selected as the target classes for the evaluation.

In this study, we focus on the IEMOCAP dataset because it aligns well with the objectives of our method and provides the necessary conditions for evaluation. It contains multi-label annotations, sufficiently ambiguous emotional samples, and diverse speakers that enable leave-one-speaker-out (LOSO) validation to examine model generalization. Hence, we adopt IEMOCAP as the benchmark corpus to ensure both experimental feasibility and fair comparison with prior work.

TABLE II: **Dataset (IEMOCAP) partitioning** used in the experiments. S_A represents clear and S_B ambiguous samples.

Session	Session 1	Session 2	Session 3	Session 4	Session 5
S_A	1316	1248	1324	1216	1494
S_B	726	801	973	1059	890

C. Data preprocessing

Inspired by the findings in [43], we observe that speech segments with a duration of 7 seconds typically contain sufficient information for effective emotion recognition. Therefore, we segment utterances longer than 7 seconds into fixed 7-second intervals. For shorter utterances, particularly those less than 1.5 seconds in length, which we found to be suboptimal for processing by the Wav2Vec 2.0 model, we apply zero-padding to extend them to at least 1.5 seconds. To ensure that emotional cues are preserved and properly learned by the model, we apply zero-padding at the end of the audio rather than at the beginning or middle. After applying these preprocessing steps, we partition the dataset as detailed in Table II.

During the feature extraction stage, we convert all original speech signals to digital form at a 16 kHz sampling rate. We apply pre-emphasis filtering to amplify high-frequency components, which are often critical for emotion-related spectral features. We then compute logarithmic Mel filter bank (MFB) energy features [44], [45], using 40 Mel filters, a 40 ms Hamming window, and a 10 ms frame shift. These features serve as input to our models for both training and evaluation.

D. Comparison Methods

To rigorously evaluate the proposed method, we select a range of state-of-the-art models for comparatison, all of which explicitly address the challenge of speech ambiguity. The goal of the comparitive analysis is to highlight the superior performance and notable advantages of our model in handling ambiguous speech.

The baseline models considered in our experiments take different strategies to mitigate the negative impact of emotional ambiguity in speech emotion recognition and can be broadly categorized into three groups: (i) models that enhance dataset reliability [28], [29] or exploiting multiple feature representations [30], (ii) models that directly utilize the observation labels of ambiguous speech without further label refinement [14], [31], and (iii) models that update ambiguous observation labels before training to better model true emotional distributions [13], [34].

It is worth noting that other studies have explored speech emotion recognition based solely on single-label strategies [27], [46]. These works tend to neglect the intrinsic ambiguity and complexity of emotional expression in speech and are, therefore, not considered in this work. Similarly, models leveraging fusion techniques accross different modalities (e.g.,

speech and text) [47], [48] attempt to compensate for the mabiguity by introducing auxiliary information. However, their focus diverges significantly from the present study, which is dedicated to resolving emotional ambiguity strictly within the speech modality. As such, methods that either ignore emotional ambiguity or introduce additional modalities are excluded from our comparisons.

The following models are included in our evaluation:

- **Co-teaching (2021)**: A progressive co-teaching framework that uses loss values to estimate sample difficulty, gradually training the model from simple to hard samples to mitigate issues with early-stage interference caused by emotionally ambiguous speech [30].
- **Attention-LSTM-Attention (2020)**: A SER model combining attention mechanisms and LSTMa to extract emotional features across both temporal and feature dimensions. This approach constructs a derived dataset from IEMOCAP, distinguishing between clear and ambiguous samples to evaluate the dataset's reliability and its impact on model performance [29].
- **LLMs (2024)**: A method that synthesizes emotionally rich speech data using large language models (LLMs) combined with IEMOCAP and student speech datasets. Transformer-based architectures are used for spatial feature extraction to enhance data reliability. [28].
- **Soft-target Training (2018)**: A soft-label training method that adjusts soft label representations to effectively utilize ambiguous speech samples without dominant consensus labels [31].
- **Inter-class Difference Loss (2023)**: A multi-label training approach that introduces an inter-class difference loss function that enabled the network to automatically learn the distribution of emotions by emphasizing differnces between emotion categories. [14].
- **Emotion Existence (2019)**: A method that first estimates the presence or absence of each emotion in speech samples using multiple labels and then refines the estimates with soft labels to resolve emotional ambiguity [13].
- **Meta-learning (2020)**: A meta-learning framework that performs real-time correction of noisy labels and estimates sample contribution weights, aiming to correct ambiguously labeled samples and reduce their negative impact during model training [34].
- **AMSNet (2023)**: A multi-scale attention-based framework designed to enhance the discriminative power of speech emotion representations. The framework employs segment-level feature refinement and hierarchical attention to improve robustness against ambiguous emotional expressions [49].
- **STACN (2025)**: A sparse temporal aware capsule network designed to improve robustness against ambiguous or noisy emotional labels by integrating sparse temporal modeling, multi-head attention, and capsule-based feature routing, achieving stable performance in speech emotion recognition tasks [50].

TABLE III: **Comparison with the state-of-the-art.** The proposed model leads to the best overall performance both in terms of WA and UA.

Method	Label	Train set		Metrics	
		Clear	Ambiguous	WA (%)	UA (%)
Co-teaching (2021) [30]	Consensus	√	-	62.3	-
Attention-LSTM (2020) [29]	Consensus	√	√	67.7	65.1
LLMs (2024) [28]	Consensus	√	-	-	66.6
		√	-	58.5	57.4
Soft-target (2018) [31]	Soft	-	√	53.6	54.0
		√	√	62.6	63.7
		√	-	66.0	63.9
Inter-class (2023) [14]	Multi	-	√	60.5	61.7
		√	√	68.3	66.2
Emotion existence (2019) [13]	Multi & Soft	√	√	66.1	65.4
Meta-learning (2020) [34]	Update Consensus	√	-	65.9	61.4
AMSNet (2023) [49]	Multi	√	√	69.2	70.5
STACN (2025) [50]	Multi	√	√	68.8	-
Proposed model (ours)	Update soft	√	√	70.3	71.3

E. Comparisons with State-Of-The-Art Methods

In the first set of experiments, we compare the proposed model with competing state-of-the-art (SOTA) SER models from the literature. We train our models on either clear, ambiguous or both types of samples (depending on the capabilities of the model), as detailed in Table III. To ensure a fair comparison, the results of other baseline methods are directly cited from the corresponding literature. From the presented results, it can be seen that the proposed model demonstrates superior performance over all considred methods in both weighted accuracy (WA) and unweighted accuracy (UA), achieving a WA score of 70.3% and UA score of 71.3.

SOTA Comparison. Compared to models that focus on enhancing dataset reliability or leveraging different feature representations, our method delivers consistent improvements. Specifically, relative to Attention-LSTM-Attention (2020) [29], we observe a 2.6% point gain in WA and a 6.2% point gain in UA. Compared to LLMs [28], our model achieves a 4.7% point improvement in UA. These results highlight that dataset augmentation alone does not sufficiently resolve the label noise caused by subjective annotation errors. In comparison to Co-teaching [30], we observe an 8.0% point increase in WA. This improvement can be attributed to our model's ability to capture fine-grained spatial-temporal information through a dedicated network structure, as well as the dynamic adjustment of ambiguous labels during training.

When comparing to models that directly use observation labels of ambiguous speech without correction, our model also exhibits clear advantages. For instance, compared with Soft-target training [31], our approach improves WA and UA by 7.7% and 7.6% points, respectively. This is largely due to our real-time correction strategy, which reduces dependence on potentially inaccurate observation labels. Compared with Inter-class Difference Loss [14], we achieve 2.0% and 5.1% points improvements in WA and UA, respectively. These gains demonstrate the effectiveness of balancing the contributions of original soft labels and model-generated labels via the correction coefficient α.

In terms of label correction approaches, our model also

TABLE IV: **Performance statistics** across LOSO folds. Mean, standard deviation, and 95% confidence intervals are reported.

Metric	Mean (%)	Std (%)	95% CI (%)
WA	70.32	±3.02	[68.16, 72.49]
UA	71.30	±3.34	[68.91, 73.69]

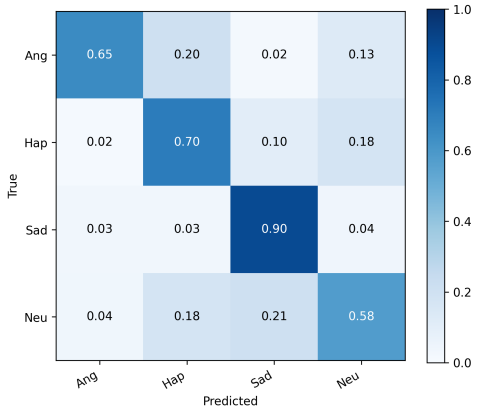


Fig. 5: **Confusion matrix** of the proposed model under one LOSO fold. The figure illustrates the per-class recognition performance, indicating that the model performs reliably across different categories, with noticeable confusions mainly between *Happiness/Sadness* and *Neutrality*.

outperforms existing methods. Compared to Emotion Existence [13], we observe improvements of 4.2% points in WA and 5.9% points in UA. This suggests that our method better captures the emotional nuances in speech samples lacking a discrete consensus label. Similarly, when compared with Meta-learning [34], our model shows substantial improvements, i.e., 4.4% points in WA and 9.9% pints in UA—due to its ability to produce a smoother convergence path and to leverage a broader set of ambiguous samples during training.

Detailed Model Analysis. To further assess the robustness of the reported improvements over the state.of-the-art, Table IV reports the mean, standard deviation, and 95% confidence intervals across the 10 LOSO folds. Specifically, our model achieves WA = 70.32 $\pm$ 3.02% (95% CI: [68.16, 72.49]) and UA = 71.30 $\pm$ 3.34% (95% CI: [68.91, 73.69]), demonstrating that the gains are consistent and statistically significant.

In addition to the standard WA and UA metrics, we also report several supplementary indicators to provide a more comprehensive assessment of the proposed model. These indicators are summarized in Table V, which presents macro and weighted versions of precision, recall, and F1-score. Together, these metrics reflect class-balanced performance, robustness under class imbalance, and the overall discriminative ability of the model.

To further illustrate the per-class recognition performance, we show in Fig. 5 the confusion matrix of the proposed model under a randomly selected LOSO fold. The results indicate that the model performs relatively consistently across most categories, with noticeable confusions mainly between Happiness/Sadness and Neutrality.

TABLE V: **Additional performance metrics of the proposed model.** To provide a more comprehensive evaluation as suggested by the reviewers, we report F1-score, weighted F1-score, macro/weighted precision, and macro/weighted recall of the proposed method.

Metric	Value (%)
F1-score (macro)	70.78
F1-score (weighted)	68.33
Precision (macro)	72.04
Precision (weighted)	70.97
Recall (macro)	71.59
Recall (weighted)	68.31

TABLE VI: **Inference efficiency of our model** measured on a single NVIDIA RTX 3090 GPU (batch size=1). Latency is averaged over 100 runs. RTF = real-time factor.

Input length (s)	FLOPs (G)	Latency (ms)	RTF	Peak Mem (GB)
1.00	6.95	4.44±0.05	0.004	4.37
1.20	8.37	4.76±0.02	0.004	4.39
1.60	11.21	5.16±0.02	0.003	4.41
1.94	13.64	5.66±0.04	0.003	4.45

F. Inference Efficiency

Beyond recognition accuracy, we also evaluate the inference efficiency of our model, as real-time performance is essential for HCI applications. Table VI summarizes the FLOPs, inference latency, real-time factor (RTF), and peak memory consumption across different input lengths. The RTF is simply obtained by dividing the inference time by the input audio duration. Our measurements show that the model achieves efficient inference with moderate GPU memory usage (∼4.4GB), demonstrating its suitability for practical deployment. It should be noted that publicly available code resources for related approaches are very limited, and most prior works do not report detailed efficiency statistics, which further highlights the contribution of our analysis.

G. Ablation Analysis

The comparative analysis presented in the previous section demonstrates that our proposed model outperforms existing approaches in handling ambiguous speech. In the next series of experiments, we conduct a series of ablation studies to better understand the impact of individual model components and learning objectives. While spatio-temporal models have been extensively explored in prior research, we focus our ablation studies on the main contributions of this, such as the real-time soft label correction module and associated learning objectives. Specifically, in the ablation studies, we explore: (i) the impact of the spatio-temporal network module, (ii) the influence of the soft label correction coefficient α and the margin control coefficient β, and (iii) the effect of different loss combinations in the overall learning objective.

Impact of the Spatio-Temporal Module. Table VII presents an analysis of the contributions of different architectural modules on the recognition performance of our model on the

TABLE VII: **Ablation study** exploring the impact of different model components, i.e., features used in the fusion module.

Impact analysis of multiple fusion modules								
Fold	Spatial-only model		Temporal-only model		Spatial-temporal integrated model		Spatial-temporal multi-fusion model	
	WA (%)	UA (%)	WA (%)	UA (%)	WA (%)	UA (%)	WA (%)	UA(%)
1	45.1	44.0	**70.3**	70.4	68.6	69.9	69.6	**71.0**
2	51.1	55.2	66.3	66.7	68.4	68.6	**69.8**	**70.9**
3	44.6	44.9	69.0	72.8	68.8	73.1	**73.0**	**75.5**
4	54.3	53.2	73.4	73.4	70.7	70.1	**74.6**	**76.2**
5	50.1	49.9	64.8	64.7	64.3	64.9	**66.7**	**67.6**
6	48.8	47.3	66.8	66.2	68.0	67.5	**68.6**	**68.1**
7	39.5	35.8	62.9	64.0	62.6	63.5	**65.4**	**66.6**
8	46.6	49.5	70.5	70.0	69.6	70.0	**73.7**	**72.7**
9	47.8	50.4	68.9	69.4	**70.2**	69.8	**69.7**	70.0
10	38.1	41.8	70.1	71.4	70.8	70.6	**72.2**	**74.4**
Average	46.6	47.2	68.3	68.9	68.2	68.7	**70.3**	**71.3**

TABLE VIII: **Sensitivity analysis** with respect to the correction coefficient α and margin control coefficient β.

$\alpha(\beta = 0.1)$	WA (%)	UA (%)	$\beta(\alpha = 0.3)$	WA (%)	UA (%)
0	69.5	70.6	0	69.8	71.0
0.1	69.8	71.0	0.1	**70.3**	**71.3**
0.2	69.8	70.1	0.2	70.0	70.7
0.3	**70.3**	**71.3**	0.3	69.6	70.4
0.4	69.8	71.1	0.4	69.5	70.7
0.5	69.7	70.9	0.5	69.7	70.7
0.6	69.5	70.9	0.6	69.3	70.2
0.7	70.0	70.7	0.7	69.3	70.6
0.8	69.6	70.9	0.8	69.4	70.3
0.9	69.4	70.6	0.9	69.2	70.5

IEMOCAP dataset. The results indicate that the spatial-only model performs significantly worse than other configurations, highlighting its inability to capture temporal dynamics, an essential component for emotion recognition. In contrast, the temporal-only model achieves performance closer to the full spatial-temporal model, suggesting that temporal features play a more dominant role in our overall framework. However, the best performance is observed with the spatial-temporal multi-fusion model, which leverages the complementary strengths of both spatial and temporal representations. This result confirms the effectiveness of jointly modeling spatial and temporal feature interactions for improved emotion classification.

Impact of Hyperparameters α and β. Table VIII shows the influence of different correction coefficients α and margin control coefficients β on model training. To determine the optimal settings for these hyperparameters, we adopt a controlled variable approach by tuning one parameter at a time while holding the other constant. We begin by fixing $\beta = 0.1$ and varying α to observe its influence on performance. Results show that the model achieves the strongest performance when the value of α equals $\alpha = 0.3$.

Subsequently, keeping $\alpha = 0.3$ fixed, we evaluate different values of β and find that $\beta = 0.1$ yields the highest performance. As shown in Table VIII, using $\alpha = 0.3$ and $\beta = 0.1$ results in 0.2% points improvement in both weighted accuracy (WA) and unweighted accuracy (UA) compared to using $\alpha = 0$ with the same β. This performance gain is attributed to the presence of noise in the observation labels of ambiguous speech, and the ability of α to balance the influence between observed and model-generated labels. Additionally, we see that setting $\alpha = 0$ disables the online correction process and forces the model to rely solely on static soft labels. As shown in Table VIII, this leads to a drop in both weighted accuracy (WA) and unweighted accuracy (UA), indicating that updating of ambiguous labels in real-time contributes positively to the overall performance. This confirms that the iterative, online correction mechanism plays an essential role in mitigating the impact of annotation ambiguity. Furthermore, we observe that increasing β beyond 0.1 (with $\alpha = 0.3$) leads to a decline in model performance. This is likely because an excessively large margin coefficient suppresses valuable prediction signals from emotion-positive categories, thereby reducing the effectiveness of the correction mechanism.

Impact of Loss Functions. Table IX presents an analysis of model performance under different combinations of loss functions. To ensure consistency and fairness in the ablation experiments, we fix the correction coefficient to $\alpha = 0.3$ and the margin control coefficient to $\beta = 0.1$. Additionally, Fig. 6 provides t-SNE visualizations, illustrating the distribution of emotional features learned under each loss configuration.

From the results in Table IX, we observe that both the cross-entropy loss L_{cor} and the enhanced inter-class difference loss L_{Ic} contribute most effectively when used in their respective roles, i.e., L_{cor} as the primary loss function for training on clear speech samples, and L_{Ic} as a regularization term for correcting soft labels in ambiguous samples. Compared to using either L_{cor} or L_{Ic} alone for both training and correction, the combined loss function improves weighted accuracy (WA) by 0.1% and 0.7% points, and unweighted accuracy (UA) by 0.3% and 0.9% points, respectively.

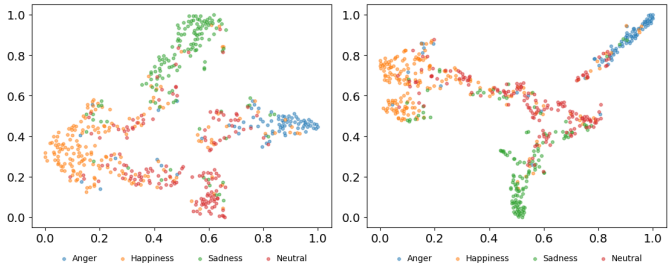
These results can be attributed to the complementary nature of the two loss functions. While L_{cor} is well-suited for optimizing predictions on clearly labeled data, it lacks the ability to effectively handle ambiguity in soft labels. In contrast, L_{Ic} is designed to increase inter-class separability in emotionally ambiguous samples but provides limited benefit for already well-separated clear samples. As previously observed in Table VIII, excessively increasing the boundary margin via β can actually degrade performance, particularly on clear samples where emotional boundaries are already well-defined. Thus, employing a hybrid loss function that combines both L_{cor} and L_{Ic} allows the model to capitalize on the strengths of each: robust learning from clear data and improved label correction for ambiguous cases.

t-SNE Visualizations. Figure 6 presents a t-SNE visualization of the learned emotion representations under the four training configurations. As can be seen, the degree of emotional clustering follows the pattern: (c) $>$ (a) $>$ (b) $>$ (d). This pattern highlights the effectiveness of the soft label correction strategy in mitigating the negative impact of label noise in ambiguous speech, which otherwise misguides model training.

A closer comparison between subfigures (c) and (a) reveals

TABLE IX: **Ablation study** with respect to the components of the overall learning objective.

Fold	L_{cor} and L_{cor}		L_{Ic} and L_{Ic}		L_{cor} and L_{Ic}	
	WA (%)	UA (%)	WA (%)	UA (%)	WA (%)	UA (%)
1	**70.4**	70.7	69.0	70.3	69.6	**71.0**
2	68.5	70.1	68.7	69.5	**69.8**	**70.9**
3	71.5	75.3	71.2	74.8	**73.0**	**75.5**
4	74.3	75.4	70.5	71.8	**74.6**	**76.2**
5	66.7	67.2	65.2	65.8	**66.7**	**67.6**
6	67.7	65.9	**70.3**	**69.6**	68.6	68.1
7	65.0	**67.6**	**65.9**	67.4	65.4	66.6
8	**74.9**	72.4	73.0	71.1	73.7	**72.7**
9	70.5	**71.3**	**71.7**	70.7	69.7	70.0
10	72.2	**74.5**	70.5	72.6	**72.2**	74.4
Average	70.2	71.0	69.6	70.4	**70.3**	**71.3**



(a) L_{cor} and L_{cor} (b) L_{Ic} and L_{Ic}

(c) L_{cor} and L_{Ic} (d) L_{cor} ($\alpha = 0$)

Fig. 6: **t-SNE plots for different loss combinations** with $\alpha = 0.3, \beta = 0.1$: (a) L_{cor} is used for both model training and soft label correction; (b) Enhanced L_{Ic} is used for both model training and soft label correction; (c) Enhanced L_{Ic} is used for soft label correction and L_{cor} is used for model training; (d) No soft label correction is applied ($\alpha = 0$), L_{cor} is used for training, and β is unrestricted.

that emotion clusters in (a) are more scattered and distributed in four distinct directions, consistent with the results in Table IX. This observation supports the conclusion that the cross-entropy loss function alone lacks sufficient discriminative power for ambiguous emotion categories. Further comparisons between (c) and (b) show that although (b) exhibits a larger inter-class margin, the distribution of neutral emotions is more diffuse and significantly overlaps with other categories. This suggests that while L_{Ic} effectively increases inter-class separation, it can simultaneously introduce confusion—particularly for more ambiguous emotion types such as neutrality. Lastly, in comparing (c) and (d), we observe that the clusters corresponding to happiness, anger, and neutral emotions in (c) are more compact and well-separated. This demonstrates that soft label correction on ambiguous speech samples helps guide the model to better capture the underlying emotional distribution.

VI. Summary

In this paper, we proposed an enhanced ambiguous speech emotion recognition model to address several key challenges in the field: the over-reliance on subjectively annotated labels, the disregard for proportional differences among emotions within multi-label annotations, and the underutilization of speech samples lacking dominant consensus labels. Our model integrates Convolutional Neural Networks (CNN) and Wav2Vec 2.0 to jointly capture spatial and temporal characteristics of speech, with a multi-level fusion mechanism that adaptively balances these features for more effective representation. A key contribution of the proposed approach is the introduction of a real-time soft label correction strategy, specifically designed to handle ambiguous labels by dynamically refining them during training. This helps reduce the adverse effects of noisy annotations on model convergence. This novel real-time refinement mechanism distinguishes our approach from existing soft-label or offline correction approaches, thereby strengthening the overall novelty of our method. We also provided a formal mathematical proof of the feasibility and effectiveness of the proposed correction mechanism. Extensive experiments conducted on the IEMOCAP dataset confirmed the superiority of our method, yielding improvements of 2.0% points in weighted accuracy (WA) and 4.7% in unweighted accuracy (UA) over existing state-of-the-art approaches. Furthermore, ablation studies on the real-time soft label correction module highlighted its critical role in enhancing model performance and provided deeper insights into its contribution.

References

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[2] M. Gerczuk, S. Amiriparian, S. Ottl, and B. W. Schuller, "Emonet: A transfer learning framework for multi-corpus speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1472–1487, 2021.

[3] S. Shen, F. Liu, H. Wang, and A. Zhou, "Towards speaker-unknown emotion recognition in conversation via progressive contrastive deep supervision," *IEEE Transactions on Affective Computing*, 2025.

[4] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *Information Fusion*, vol. 49, pp. 69–78, 2019.

[5] W. Zhu and X. Li, "Speech emotion recognition with global-aware fusion on multi-scale feature representation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6437–6441.

[6] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE access*, vol. 7, pp. 19 143–19 165, 2019.

[7] M. Hou, Z. Zhang, Q. Cao, D. Zhang, and G. Lu, "Multi-view speech emotion recognition via collective relation construction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 218–229, 2021.

[8] W. Fan, X. Xu, B. Cai, and X. Xing, "Isnet: Individual standardization network for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1803–1814, 2022.

[9] Y. Kim and J. Kim, "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, 2018, pp. 5104–5108.

[10] R. Lotfian and C. Busso, "Curriculum learning for speech emotion recognition from crowdsourced labels," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 815–826, 2019.

[11] H. M. Fayek, M. Lech, and L. Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *2016 international joint conference on neural networks (IJCNN)*. IEEE, 2016, pp. 566–570.

[12] M. Lukasik, S. Bhojanapalli, A. Menon, and S. Kumar, "Does label smoothing mitigate label noise?" in *International Conference on Machine Learning*. PMLR, 2020, pp. 6448–6458.

[13] A. Ando, R. Masumura, H. Kamiyama, S. Kobashikawa, and Y. Aono, "Speech emotion recognition based on multi-label emotion existence model." in *INTERSPEECH*, 2019, pp. 2818–2822.

[14] X. Li, Z. Zhang, C. Gan, and Y. Xiang, "Multi-label speech emotion recognition via inter-class difference loss under response residual network," *IEEE Transactions on Multimedia*, vol. 25, pp. 1520–9210, 2023.

[15] S. Steidl, M. Levit, A. Batliner, E. Noth, and H. Niemann, "" 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, 2005, pp. I–317.

[16] S. Mao, P. Ching, and T. Lee, "Enhancing segment-based speech emotion recognition by iterative self-learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 123–134, 2021.

[17] Y. Zhou, X. Liang, Y. Gu, Y. Yin, and L. Yao, "Multi-classifier interactive learning for ambiguous speech emotion recognition," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 30, pp. 695–705, 2022.

[18] H.-C. Chou and C.-C. Lee, "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, 2019, pp. 5886–5890.

[19] H.-C. Chou, L. Goncalves, S.-G. Leem, A. N. Salman, C.-C. Lee, and C. Busso, "Minority views matter: Evaluating speech emotion classifiers with human subjective annotations by an all-inclusive aggregation rule," *IEEE Transactions on Affective Computing*, 2024.

[20] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1634–1654, 2021.

[21] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47 795–47 814, 2021.

[22] A. Hashem, M. Arif, and M. Alghamdi, "Speech emotion recognition approaches: A systematic review," *Speech Communication*, vol. 154, p. 102974, 2023.

[23] S. M. George and P. M. Ilyas, "A review on speech emotion recognition: a survey, recent advances, challenges, and the influence of noise," *Neurocomputing*, vol. 568, p. 127015, 2024.

[24] J. Ye, X.-C. Wen, Y. Wei, Y. Xu, K. Liu, and H. Shan, "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 (ICASSP)*. IEEE, 2023, pp. 1–5.

[25] S. Li, X. Xing, W. Fan, B. Cai, P. Fordson, and X. Xu, "Spatiotemporal and frequential cascaded attention networks for speech emotion recognition," *Neurocomputing*, vol. 448, pp. 238–248, 2021.

[26] X. Wu, Y. Cao, H. Lu, S. Liu, D. Wang, Z. Wu, X. Liu, and H. Meng, "Speech emotion recognition using sequential capsule networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3280–3291, 2021.

[27] C. Gan, K. Wang, Q. Zhu, Y. Xiang, D. K. Jain, and S. García, "Speech emotion recognition via multiple fusion under spatial–temporal parallel network," *Neurocomputing*, vol. 555, p. 126623, 2023.

[28] L. Wang, J. Yang, Y. Wang, Y. Qi, S. Wang, and J. Li, "Integrating large language models (llms) and deep representations of emotional features for the recognition and evaluation of emotions in spoken english," *Applied Sciences*, vol. 14, no. 9, p. 3543, 2024.

[29] Y. Yu and Y.-J. Kim, "Attention-lstm-attention model for speech emotion recognition and analysis of iemocap database," *Electronics*, vol. 9, no. 5, p. 713, 2020.

[30] Y. Yin, Y. Gu, L. Yao, Y. Zhou, X. Liang, and H. Zhang, "Progressive co-teaching for ambiguous speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6264–6268.

[31] A. Ando, S. Kobashikawa, H. Kamiyama, R. Masumura, Y. Ijima, and Y. Aono, "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, 2018, pp. 4964–4968.

[32] C. Wang, J. Shi, C. Tao, F. Xu, X. Tang, L. Li, Y. Zhou, B. Tian, S. Wei, and X. Zhang, "Multitype label noise modeling and uncertainty-weighted label correction for concealed object detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–12, 2023.

[33] D. Liu, I. W. Tsang, and G. Yang, "A convergence path to deep learning on noisy labels," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 4, pp. 5170–5182, 2022.

[34] T. Fujioka, T. Homma, and K. Nagamatsu, "Meta-learning for speech emotion recognition considering ambiguity of emotion labels," in *INTERSPEECH*, 2020, pp. 2332–2336.

[35] S. Mao, P.-C. Ching, and T. Lee, "Emotion profile refinery for speech emotion classification," *INTERSPEECH*, pp. 531–535, 2020.

[36] Y. Yin, Y. Gu, L. Yao, Y. Zhou, X. Liang, and H. Zhang, "Progressive co-teaching for ambiguous speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6264–6268.

[37] S. Chopra, P. Mathur, R. Sawhney, and R. R. Shah, "Meta-learning for low-resource speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6259–6263.

[38] R. Cai, K. Guo, B. Xu, X. Yang, and Z. Zhang, "Meta multi-task learning for speech emotion recognition." in *Interspeech*, 2020, pp. 3336–3340.

[39] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[40] G. Li, J. Hou, Y. Liu, and J. Wei, "Mvib-dva: Learning minimum sufficient multi-feature speech emotion embeddings under dual-view aware," *Expert Systems with Applications*, vol. 246, p. 123110, 2024.

[41] A. Derington, H. Wierstorf, A. Özkil, F. Eyben, F. Burkhardt, and B. W. Schuller, "Testing correctness, fairness, and robustness of speech emotion recognition models," *IEEE Transactions on Affective Computing*, 2025.

[42] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[43] A. Aftab, A. Morsali, S. Ghaemmaghami, and B. Champagne, "Lightsernet: 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, 2022, pp. 6912–6916.

[44] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[45] Q. Cao, M. Hou, B. Chen, Z. Zhang, and G. Lu, "Hierarchical network based on the fusion of static and dynamic features for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6334–6338.

[46] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[47] Y. Song and Q. Zhou, "Bi-modal bi-task emotion recognition based on transformer architecture," *Applied Artificial Intelligence*, vol. 38, no. 1, p. 2356992, 2024.

[48] B. T. Atmaja, K. Shirai, and M. Akagi, "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, 2019, pp. 519–523.

[49] Z. Chen, J. Li, H. Liu, X. Wang, H. Wang, and Q. Zheng, "Learning multi-scale features for speech emotion recognition with connection attention mechanism," *Expert Systems with Applications*, vol. 214, p. 118943, 2023.

[50] H. Zhang, H. Huang, P. Zhao, and Z. Yu, "Sparse temporal aware capsule network for robust speech emotion recognition," *Engineering Applications of Artificial Intelligence*, vol. 144, p. 110060, 2025.

Chenquan Gan received the Ph.D. degree from the Department of Computer Science, Chongqing University, Chongqing, China, in 2015. He is currently an Associate Professor with Chongqing University of Post and Telecommunications (CQUPT), Chongqing. His research interests include network propagation and control, sentiment analysis, and blockchain.

Daitao Zhou is currently pursuing the M.S. degree with the Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests is sentiment analysis.

Qingyi Zhu received the Ph.D. degree in computer science and technology from the College of Computer Science, Chongqing University, Chongqing, China, in 2014. He is currently a Professor with the Chongqing University of Posts and Telecommunications, Chongqing. He has published more than 60 academic articles in peer-reviewed international journals. His current research interests include cybersecurity dynamics, complex systems, and blockchain.

Xibin Wang received a PhD degree in computer science from Chongqing University in 2015. He is a professor and vice dean in the School of Data Science at Guizhou Institute of Technology. His research interests include recommendation systems, data mining and service computing.

Deepak Kumar Jain (Senior Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, University of Chinese Academy of Sciences, Beijing, China, in 2018. He is currently an Associate Professor with Dalian University of Technology, Dalian, China. He has presented several papers in peer-reviewed conferences and has authored or co-authored numerous studies in science cited journals. His research interests include deep learning, machine learning, pattern recognition, and computer vision.

Vitomir Štruc is a Full Professor at the University of Ljubljana, Slovenia, and an expert on computer vision and machine learning. He has co-authored over 200 papers in leading international journals and conferences. Vitomir serves as Deputy Editor-in-Chief of IEEE T-IFS, Subject Editor for Signal Processing, and Associate Editor for Pattern Recognition. Dr. Štruc is a Senior IEEE member, current VP Technical Activities of the IEEE Biometrics Council, the secretary of IAPR TC4, and a supervisory board member of the EAB.