APSIPA Transactions on Signal and Information Processing, 2025, 14, e20

Original Paper

Multi-attribute Learning for Multi-level Emotion Recognition from Speech

Yuan Gao*, Hao Shi, Chenhui Chu and Tatsuya Kawahara

Kyoto University, Kyoto, Japan

ABSTRACT

In this paper, we propose a multi-level speech emotion recognition (SER) system that captures both acoustic and linguistic emotional features using only speech input. In particular, our approach utilizes an acoustic feature extractor (HuBERT) to process the input waveform, capturing acoustic emotional features while simultaneously performing automatic speech recognition (ASR) to implicitly learn linguistic information. The ASR-decoded transcriptions are then fed into a linguistic feature extractor (BERT) to explicitly encode linguistic emotional features. To combine these features, we introduce a temporal gated fusion method that dynamically modulates the contribution of each modality, addressing modality incongruity issues. Furthermore, integrating multi-attribute learning for emotion-related attributes such as gender and speaking style, further enhances SER performance. To address gradient conflicts inherent in multi-attribute learning, we propose a two-stage fine-tuning framework employing adapters. Additionally, to mitigate the negative impact of ASR errors, we introduce an error correction module and a contrastive learning method to align representations learned from ground-truth text and the decoded transcriptions. Comprehensive experimental results on the IEMOCAP and MELD datasets validate that our method enhances SER

*Corresponding author: gao.yuan.75x@st.kyoto-u.ac.jp. This work was supported by JST SPRING (JPMJSP2110), JST Moonshot R&D (JPMJPS2011).

Received 08 May 2025; revised 10 June 2025; accepted 03 July 2025

ISSN 2048-7703; DOI 10.1561/116.20250025

performance without requiring textual input. Compared to the acoustic model baseline, our approach achieves a 10.85% improvement in unweighted accuracy on the IEMOCAP dataset.

Keywords: Speech emotion recognition, automatic speech recognition, self-supervised learning, multi-attribute learning, human-computer interaction.

1 Introduction

Human-computer interactions have become pervasive in our daily lives, and understanding human emotion is crucial for the development of artificial intelligence [11, 3]. Therefore, research on sentiment analysis and emotion recognition has attracted increasing attention in both industry and academia [23, 61]. Speech emotion recognition (SER) aims to identify emotional attributes in human speech [13], and a robust SER system can promote the development of empathetic chatbots [8] and enrich customer services [44]. This research also has other applications, such as monitoring the mental state during interviews [14].

As with other classification tasks, previous SER systems primarily rely on supervised learning models to learn emotion information from spectrograms [4, 66]. However, the scarcity of labeled emotional speech data poses a significant challenge [1]. Collecting and annotating such data is both time-consuming and costly, leading to insufficient datasets for training robust supervised deep learning models. To address this issue, researchers focus on self-supervised learning (SSL) approaches. SSL models, such as HuBERT [19], have been widely adopted in recent years for various speech processing tasks [65, 34]. These models are pre-trained on vast amounts of unlabeled speech data through contrastive learning, capturing underlying acoustic features and linguistic patterns. Then they can be fine-tuned using smaller labeled datasets for downstream tasks, including SER. This approach mitigates the data sparsity problem by transferring the prior knowledge from the pre-training dataset, allowing the models to achieve better performance even with limited labeled data on the target task. Therefore, this work incorporate SSL models as feature extraction modules. However, learning solely based on discrete emotion labels may not be sufficient to capture the emotion expressions.

Human emotional expressions are closely related to linguistic and acoustic attribute information [48, 56]. Linguistic content conveys semantic meaning and can be decoded by ASR systems. Different speakers may display unique acoustic features when expressing emotions in various elicitation contexts [64,

37]. However, in many real-world applications, speaker information is unavailable. Therefore, we focus on speaker attributes, such as gender and nationality. Previous studies [27] have confirmed that gender differences generate specific acoustic patterns in emotional speech. Consequently, incorporating gender identification can enhance the accuracy of emotion recognition systems. Other attributes, such as speaking style, may also influence SER but remain underexplored. For instance, spontaneous speech is generally more challenging for SER systems than acted speech. Accounting for differences in emotion expression style may help improve feature extraction and model performance. Considering that pre-trained SSL models excel in ASR and other tasks, we propose incorporating multi-attribute learning to improve SER.

Multi-attribute learning involves simultaneously learning multiple emotion-related attributes, which can be achieved through multitask learning (MTL). However, MTL poses challenges, particularly concerning gradient conflicts [67]. For instance, ASR typically requires features invariant to speaker characteristics like identity and speaking style, as these are not directly relevant to transcription tasks. Conversely, these features are crucial for tasks like gender classification or speaking style recognition, which are often beneficial for SER. Moreover, the varying complexities of different tasks can result in imbalanced learning rates, further exacerbating the gradient conflict problem. Addressing these conflicts is crucial for leveraging the potential of MTL in SER systems. Although incorporating ASR can potentially improve SER performance by implicitly learning linguistic information, speech-only models still face two main limitations. First, most existing SER systems lack a linguistic feature extraction module (such as BERT), which can explicitly learn emotional cues from the text input. Secondly, the ASR transcription inevitably includes errors of differing severity in each utterance, resulting in degraded feature extraction compared to ground-truth text transcription. While multimodal emotion recognition (MER) systems that combine speech and ground-truth transcriptions as input features do not suffer from these issues, the availability of multimodal data in practical applications is often limited [5, 30]. Additionally, previous works often overlook the problem of modality incongruity [33, 46]. For instance, when negative emotions are expressed sarcastically (e.g.,"That's great"), analyzing emotion based solely on linguistic information becomes challenging [63].

In this paper, we propose a multi-level SER approach that addresses these issues found in previous studies. Our approach processes speech input to extract emotional features at multiple levels: the acoustic level and the linguistic level. Initially, we use an acoustic feature extractor to derive acoustic emotional features directly from the speech signal, capturing prosodic and paralinguistic cues. Concurrently, we perform ASR to obtain transcriptions of the speech. The generated transcriptions are then fed into a linguistic feature extractor to obtain linguistic emotional features, capturing semantic

cues. By combining these acoustic and linguistic features, our system effectively captures emotional cues at multiple levels. Moreover, to mitigate the gradient conflict problem inherent in multitask learning, we introduce a two-stage fine-tuning approach. Building upon the initial training of ASR and SER tasks, we advance to the second stage by freezing feature extractors and incorporating adapter structures as auxiliary learnable parameters. These adapters are trained for additional attribute information recognition, such as gender and emotion expression style (acted vs. spontaneous), alongside SER. This proposed fine-tuning method not only resolves the gradient conflict by sequentially adapting the model for different tasks but also prevents information loss that can occur when updating all model parameters in the second stage. For the modality incongruity during feature fusion, we propose an attention-based gated fusion method that learns complementary information from multimodal features. This module is adopted to dynamically modulate the contribution of each feature representation. Moreover, recognition errors in the generated transcriptions can cause deviations in semantic feature extraction; therefore, we conduct error correction and incorporate a contrastive learning method, which encourages feature learning from transcriptions to align with those learned from ground-truth text to mitigate the impact of these errors on SER.

The main contributions of this work are as follows:

1. **Propose a Multi-level SER System**: We propose a multi-level system that enhances SER by effectively capturing both acoustic and linguistic emotional features using only speech input.

2. **Introduce a Multi-attribute Learning with Two-Stage Fine-Tuning Framework**: To allow for effective multitask learning, we introduce a two-stage fine-tuning framework using adapters in the multi-level SER system. This approach avoids gradient conflicts by training SER, ASR, and other related objectives in separate stages.

3. **Develop an Attention-based Gated Fusion Model**: To handle modality incongruity, we propose an attention-based gated fusion model that dynamically modulates the contribution of multimodal features.

4. **Enhance Linguistic Features**: To mitigate the negative impact of ASR errors on SER, we employ post-processing techniques (error correction) and introduce a contrastive learning method to enhance the linguistic emotional features.

Building upon our previous work [16] that focused on improving SER using a linguistic feature extractor, this paper further explores the effectiveness of multi-attribute learning in multi-level SER. We provide an in-depth examination of the feature fusion method through visualization and enhance the

system by mitigating the impact of ASR errors on emotional feature extraction. While we have demonstrated that adapter tuning can address gradient conflicts in multi-attribute learning for HuBERT [15], this work extends the investigation by applying adapter tuning in multi-level SER, incorporating both HuBERT and BERT models. Compared with previous speech-based emotion recognition systems, our approach effectively incorporates multi-attribute emotion-related information and leverages both acoustic and semantic features jointly. Experimental results demonstrate that the proposed multi-level system significantly outperforms previous SER methods and achieves comparable performance with multi-modal systems that use the ground-truth text.

2 Related Work

2.1 Self-supervised Learning for Affective Computing

Traditional SER approaches, which rely on handcrafted features and deep learning models, face significant challenges such as dependence on expert knowledge and the scarcity of labeled emotional speech data. The advent of SSL offers a promising solution to the data scarcity problem in SER. Researchers have widely applied SSL models to various downstream tasks [53, 55], such as ASR [52], speaker verification [59], and speech enhancement [51]. As for SER, significant improvements have been achieved by fine-tuning these models on limited labeled data. For instance, Pepino *et al.* [40] leveraged SSL embeddings from Wav2vec 2.0 to outperform traditional CNN and RNN models on the IEMOCAP and RAVDESS datasets. Building upon this foundation, researchers have explored cross-lingual applications of SSL models in SER. Pastor *et al.* [38] utilized HuBERT for cross-lingual emotion recognition, demonstrating the model's ability to capture language-independent emotional features and generalize across different languages. Similarly, in text-based sentiment analysis, models like BERT [20] have demonstrated superior performance in capturing semantic information [10], which can be applied to downstream tasks such as emotion recognition [43]. By integrating SSL models into SER, the field can overcome limitations posed by data scarcity and domain mismatch. SSL enables models to learn from vast amounts of unlabeled data, capturing intricate patterns in speech that are crucial for emotion recognition. This leads to more robust and accurate SER systems, advancing the capabilities of affective computing.

2.2 Multitask Learning

Although SSL models have significantly enhanced feature extraction capabilities in affective computing, leveraging emotion related attribute information

can further improve model performance. This leads us to the exploration of MTL, which is a widely used approach that enhances deep learning model by leveraging the attribute information [9, 57, 70]. By concurrently optimizing multiple related tasks, MTL enables the model to learn a shared representation across tasks while preserving task-discriminative information [68, 49].

In the past decade, this approach has proven effective in SER [7, 50, 24], as it allows the model to benefit from the information provided by the attribute tasks. Given the strong correlation between linguistic content and emotional expression, Cai *et al.* [7] proposed an MTL approach that combines ASR and SER using the wav2vec 2.0 model. By jointly training ASR and SER, the model leverages linguistic information to enhance emotional feature extraction. Their approach achieved state-of-the-art results on the IEMOCAP dataset using 10 fold cross validation. Ablation studies confirm that incorporating ASR through appropriate weight parameters can generate optimal performance. Furthermore, speaker attributes have been extensively studied in previous works. On the one hand, different speakers have different ways of expressing emotions, leading to differences in both acoustic and linguistic content for the same emotion. Therefore, speaker-dependent SER systems perform much better than speaker-independent ones [17]. On the other hand, incorporating gender identification can enhance the feature extraction process in speaker-independent SER [37]. More recently, Sharma *et al.* [50] proposed a multi-lingual MTL approach for SER using pre-trained SSL feature extractor. The study further highlights the effectiveness of MTL by integrating other auxiliary tasks including language classification and regression tasks related to pitch and energy. These additional sources of information have been shown to be effective on 25 datasets across 13 locales and 7 emotion categories. Their work underscored the advantages of MTL in enhancing generalization and performance in SER.

Previous studies on MTL have demonstrated that using auxiliary tasks can improve SER performance. However, these approaches often encounter the problem of gradient conflicts between tasks, leading to suboptimal learning. Unlike existing methods that jointly optimize all tasks simultaneously, our work addresses this problem by splitting the training objective into two stages, enabling more effective fine-tuning of multi-attribute information learning.

2.3 Multimodal Emotion Recognition

In human emotional expression, speech conveys nuances through prosody, tone, and rhythm [58, 21, 61], while text captures semantic content and context crucial for analyzing sentiment [22, 32]. MER leverages the simultaneous expression of emotions across multiple channels, each contributing distinct yet complementary information [69]. Consequently, extensive research has focused on combining these modalities to learn more discriminative features for emotion recognition.

One of the pioneering studies on MER by Ngiam *et al.* [35] proposed training a bimodal deep belief network fine-tuned to minimize reconstruction errors in both speech and video modalities. Their approach outperformed single-modality models, demonstrating the effectiveness of multimodal systems. With the advancements in deep learning, Poria *et al.* [41] introduced a model integrating speech and text data for sentiment analysis in user-generated videos. They employed CNN and RNN networks to extract speech and textual features, achieving promising results in MER. Building on these foundational studies, Tsai *et al.* [60] proposed a multimodal Transformer employing cross-modal attention to process features from different modalities. This approach enables the model to capture interactions between modalities without requiring aligned input. The method was validated on the MOSI and MOSEI datasets, demonstrating superior performance through the use of attention mechanisms.

However, despite these advancements, the issue of modality incongruity, where different modalities provide conflicting or misaligned information, remains unresolved [36, 28]. This incongruity limits models' ability to effectively integrate multimodal information, resulting in diminished performance in MER systems. Moreover, practical applications often face single-modality constraints where only speech is available. Unlike existing approaches that either ignore modality conflicts or require multiple input modalities, our work addresses these problems by extracting both acoustic and linguistic information from speech alone, while developing effective fusion strategies to mitigate potential feature conflicts.

3 Proposed Method

In this section, we introduce our proposed SER system. Our approach leverages multi-level feature learning to capture both acoustic and linguistic emotional cues from speech without requiring textual input. Figure 1 illustrates our proposed approach, which learns emotional information across multiple hierarchical levels. Initially, an acoustic feature extractor based on HuBERT processes the input waveform to capture acoustic emotional features while performing an ASR task, thereby implicitly learning linguistic information. The transcriptions decoded from the ASR output are then input into a linguistic feature extractor (BERT) to explicitly encode linguistic emotional features. By effectively combining acoustic and linguistic emotional features using our proposed attention-based feature fusion method, our approach offers a comprehensive solution for SER without textual input. The subsequent subsections provide a detailed exploration of each component of our framework.

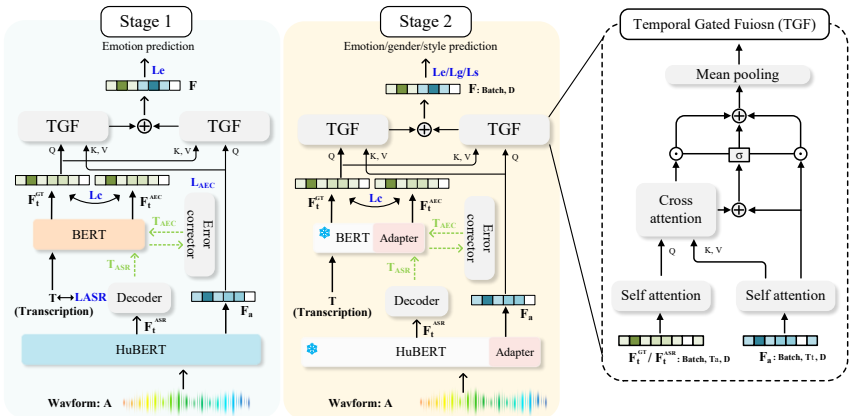


Figure 1: Architecture of the proposed multi-level feature extraction system for SER. The system consists of two primary stages: First, the input waveform $\mathbf{A}$ is processed by an acoustic feature extractor (HuBERT) to obtain acoustic features $\mathbf{F}_a$. These features are used for the ASR task to generate the predicted transcription $\mathbf{T}_{\text{ASR}}$. Simultaneously, a linguistic feature extractor (BERT) processes the ground-truth transcription $\mathbf{T}$ (or $\mathbf{T}_{\text{ASR}}$ during inference) to obtain linguistic features $\mathbf{F}_t$. In this stage, the BERT model is not only fine-tuned for emotion recognition but also pre-trained for ASR error correction. Both $\mathbf{F}_a$ and $\mathbf{F}_t$ are time-pooled and combined using the proposed attention-based gated fusion method to produce the final fused features $\mathbf{F}$ for SER. Furthermore, a contrastive loss $\mathcal{L}_c$ is used to align representations learned from output texts from AEC ($\mathbf{F}_t^{AEC}$) and ground-truth texts ($\mathbf{F}_t^{GT}$). In the second stage, we apply adapter tuning to the feature extractors (HuBERT and BERT); these adapters are double-layer linear layers inserted into each Transformer layer of HuBERT and BERT to learn attribute information (gender, emotion expression style) and avoid gradient conflict.

3.1 Multi-attribute Learning with Multi-stage Fine-tuning

To enhance the proposed SER approach, we incorporate emotion related attribute information learning including gender and emotion expression style. However, multi-attribute learning faces problem due to diverse learning objectives and varying gradient magnitudes across tasks, potentially compromising SER performance during optimization. To address this problem, we propose a two-stage fine-tuning approach that balances the learning of linguistic and paralinguistic features in our multi-level SER.

3.1.1 Joint Acoustic-linguistic Representation Learning

In the first stage, we fine-tune the HuBERT model to perform both SER and ASR simultaneously. Given an input waveform $\mathbf{A}$, we extract the latent acoustic feature representation $\mathbf{F}_a \in \mathbb{R}^{T_a \times D}$ from the final Transformer layer

of the HuBERT model, where T_a denotes the temporal length of the acoustic input and D represents the hidden dimension of the Transformer layer. For the ASR task, we process $\mathbf{F}_a$ through a fully connected (FC) layer to produce $\mathbf{F}_{\text{asr}}$, and apply the Connectionist Temporal Classification (CTC) loss function. The loss is defined as:

$$\mathcal{L}_{\text{ASR}} = \text{CTC}(\text{softmax}(\mathbf{F}_{\text{asr}}), \mathbf{T}), \tag{1}$$

where $\mathbf{T}$ is the ground-truth transcription. For the emotion recognition task, we consider the input text $\mathbf{T}$ (or the ASR decoded transcription $\mathbf{T}_{\text{ASR}}$ during inference). A BERT model is fine-tuned to extract the latent linguistic representation $\mathbf{F}_t \in \mathbb{R}^{T_t \times D}$ from its last Transformer layer, where T_t denotes the sequence length of the textual input. We then apply time pooling to both $\mathbf{F}_a$ and $\mathbf{F}_t$ to obtain fixed-length representations $\mathbf{F}_a, \mathbf{F}_t \in \mathbb{R}^D$ for each modality, which are subsequently fused to create a multimodal feature representation $\mathbf{F} \in \mathbb{R}^D$. Finally, we conduct SER using the cross-entropy (CE) loss function:

$$\mathbf{F} = \text{FC}(\mathbf{F}_a \oplus \mathbf{F}_t), \tag{2}$$

$$\mathcal{L}_{\text{e}} = \text{CE}(\text{softmax}(\mathbf{F}), \mathbf{y}_e), \tag{3}$$

where $\mathbf{y}^e$ is the ground-truth emotion label. The total loss function for the first stage is then formulated as a weighted sum of the ASR loss and the emotion recognition loss:

$$\mathcal{L}_{1st} = \lambda_{ASR}\mathcal{L}_{ASR} + \lambda_e\mathcal{L}_e, \tag{4}$$

where λ_{ASR} and λ_e are hyperparameters that control the relative importance of each loss term.

3.1.2 Adapter-based Refinement

Adapters have recently emerged as a novel approach in transfer learning [31], initially developed for natural language processing tasks and now applied to speech processing. Adapters have enabled a significant advancement in the fine-tuning of large pre-trained models, offering a method for parameter-efficient fine-tuning. In the second stage of fine-tuning, we introduce adapter tuning to mitigate gradient conflicts between tasks and prevent potential information loss from full parameter updates. Adapters introduce a small number of task-specific parameters that allow the model to specialize for different tasks while sharing most of the pre-trained parameters across tasks. This parameter-efficient approach directly addresses gradient conflicts by task-specific adaptations, thereby maintaining the integrity of shared representations while enabling effective multitask learning. Then we process the fused feature $\mathbf{F}$ using

task-specific FC layers. The output features for each task denoted as $\mathbf{F}_e$, $\mathbf{F}_g$, and $\mathbf{F}_s$, are used for emotion, gender, and style recognition, respectively. The individual loss functions for these tasks are formulated as:

$$\mathcal{L}_{task} = \mathrm{CE}\left(softmax(\mathbf{F}_{task}), \mathbf{y}^{task}\right), \quad task \in \{e, g, s\} \tag{5}$$

where $task$ represent the emotion (e), gender (g), and style (s) recognition tasks, respectively. $\mathbf{y}^e$, $\mathbf{y}^g$, and $\mathbf{y}^s$ are the ground-truth labels for these tasks.

3.2 Temporal Gated Fusion Method

To address modality incongruity when integrating acoustic and linguistic features, previous works [29] have introduced the gated fusion method to learn the importance of the input features. Specifically, given the speech features $\mathbf{F}_a$ and text features $\mathbf{F}_t$ ($\mathbf{F}_t^{\mathrm{GT}}$ during training, $\mathbf{F}_t^{\mathrm{ASR}}$ during inference), each modality is independently pooled using mean pooling:

$$\bar{\mathbf{F}}_a = \mathrm{MeanPool}(\mathbf{F}_a), \quad \bar{\mathbf{F}}_t = \mathrm{MeanPool}(\mathbf{F}_t). \tag{6}$$

These pooled features are concatenated and passed through a sigmoid-activated gating mechanism to obtain the gated features:

$$\tilde{\mathbf{F}}_a = \sigma\left(\bar{\mathbf{F}}_a \oplus \bar{\mathbf{F}}_t\right) \odot \bar{\mathbf{F}}_a, \quad \tilde{\mathbf{F}}_t = \sigma\left(\bar{\mathbf{F}}_t \oplus \bar{\mathbf{F}}_a\right) \odot \bar{\mathbf{F}}_t, \tag{7}$$

where $\sigma(\cdot)$ denotes the sigmoid function, $\oplus$ represents concatenation, and $\odot$ denotes element-wise multiplication. However, this approach suffers from significant information loss due to the early mean pooling operation, which discards temporal dynamics.

To overcome these limitations, we propose an attention based temporal gated fusion (TGF) method that preserves temporal information and enables effective cross-modal interactions. Specifically, given the speech features $\mathbf{F}_a$ and text features $\mathbf{F}_t$, we first apply self-attention mechanisms $\mathcal{A}_{\mathrm{self}}(\cdot)$ to learn emotion-salient information within each modality. Subsequently, we employ cross-attention mechanisms $\mathcal{A}_{\mathrm{cross}}(\cdot,\cdot)$ to capture the mutual information between speech and text as follows:

$$\mathbf{F}^c_{a\leftarrow t} = \mathcal{A}_{\mathrm{cross}}\left(\mathcal{A}_{\mathrm{self}}(\mathbf{F}_a), \mathcal{A}_{\mathrm{self}}(\mathbf{F}_t)\right), \tag{8}$$

$$\mathbf{F}^c_{t\leftarrow a} = \mathcal{A}_{\mathrm{cross}}\left(\mathcal{A}_{\mathrm{self}}(\mathbf{F}_t), \mathcal{A}_{\mathrm{self}}(\mathbf{F}_a)\right). \tag{9}$$

Here, the output of cross-attention $\mathbf{F}^c_{a\leftarrow t}$ captures the intrinsic properties of speech along with complementary textual cues, while $\mathbf{F}^c_{t\leftarrow a}$ captures the intrinsic properties of text combined with complementary auditory cues. The self-attention mechanisms preserve discriminative acoustic features F^s_a and

linguistic features F_t^s. Next, we incorporate the fine-grained gated fusion mechanism to dynamically weigh the contributions of F^s and F^c:

$$\mathbf{g}_a = \sigma\left(\mathbf{F}^c_{a\leftarrow t} \oplus \mathbf{F}^s_a\right), \quad \mathbf{g}_t = \sigma\left(\mathbf{F}^c_{t\leftarrow a} \oplus \mathbf{F}^s_t\right), \tag{10}$$

Then the gated features are computed as:

$$\tilde{\mathbf{F}}_{a\leftarrow t} = \left(\mathbf{g}_a \odot \mathbf{F}^c_{a\leftarrow t}\right) \oplus \mathbf{F}^s_a \tag{11}$$

$$\tilde{\mathbf{F}}_{t\leftarrow a} = \left(\mathbf{g}_t \odot \mathbf{F}^c_{t\leftarrow a}\right) \oplus \mathbf{F}^s_t, \tag{12}$$

By applying gated fusion directly to the cross-attended features, our approach effectively preserves the temporal dimension throughout the integration process. We then apply mean pooling to obtain fixed-length representations:

$$\bar{\mathbf{F}}_{a\leftarrow t} = \text{MeanPool}(\tilde{\mathbf{F}}_{a\leftarrow t}), \quad \bar{\mathbf{F}}_{t\leftarrow a} = \text{MeanPool}(\tilde{\mathbf{F}}_{t\leftarrow a}). \tag{13}$$

The proposed TGF ensures that the pooled features $\bar{\mathbf{F}}_{a\leftarrow t}$ and $\bar{\mathbf{F}}_{t\leftarrow a}$ retain rich temporal and multimodal information and better learn emotional cues without significant loss of critical details. The final fused feature $\mathbf{F}$ is obtained by concatenating these representations:

$$\mathbf{F} = \text{FC}(\bar{\mathbf{F}}_{a\leftarrow t} \oplus \bar{\mathbf{F}}_{t\leftarrow a}), \tag{14}$$

Finally, we fed F through a fully connected layer followed by a softmax for SER.

3.3 Bridging ASR Errors in Multi-level SER

We address a critical challenge in our multi-level approach: the potential degradation of linguistic features due to errors introduced by ASR. The textual representation $\mathbf{F}_t^{\text{ASR}}$, derived from ASR transcriptions, may contain errors that lead to suboptimal performance in emotional feature extraction. To mitigate this issue, we introduce an error correction module that utilizes the same BERT model used for emotion recognition. Specifically, we input the ASR transcription $\mathbf{T}_{\text{ASR}}$ into BERT for automatic error correction (AEC), feed the output features into 8 Transformer layers, and use beam search to decode the corrected transcription. The corrected text is then input back into the same BERT model for emotional feature extraction. Additionally, we propose a contrastive learning approach to mitigate the impact of ASR errors. During training, we extract $\mathbf{F}_t^{\text{ASR}}$ from the decoded transcription (or the transcription after AEC) and conduct contrastive learning for $\mathbf{F}_t^{\text{ASR}}$ and $\mathbf{F}_t^{\text{GT}}$. We apply contrastive learning during both fine-tuning stages to encourage the model to generate similar representations from both the ground-truth

text and the ASR transcriptions, even in the presence of transcription errors. The contrastive loss $\mathcal{L}c$ is defined as:

$$\mathcal{L}c = -\log \frac{\exp\left(\cos\left(\mathbf{F}_t^{GT}, \mathbf{F}_t^{\text{ASR}}\right)/\tau\right)}{\sum_{i=1}^{N} \exp\left(\cos\left(\mathbf{F}_t^{GT}, \mathbf{F}_t^{\text{ASR}(i)}\right)/\tau\right)}, \tag{15}$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity, τ is a temperature parameter, N is the number of samples, and $\mathbf{F}_t^{\text{ASR}(i)}$ represents the ASR transcription representation from the i-th sample in the batch. In the first fine-tuning stage, we incorporate AEC pre-training and CL to enhance the robustness of $\mathbf{F}_t^{\text{ASR}(i)}$ for SER. The overall objective function for the first stage is redefined as:

$$\mathcal{L}_{\text{1st}} = \lambda e\mathcal{L}e + \lambda_{ASR}\mathcal{L}_{ASR} + \lambda_{AEC}\mathcal{L}_{AEC} + \lambda c\mathcal{L}c, \tag{16}$$

In the second stage, we freeze the AEC module and fine-tune the adapters in BERT for SER, gender, style recognition, together with $\mathcal{L}c$. The overall objective function for the second stage is defined as:

$$\mathcal{L}_{\text{2nd}} = \lambda e\mathcal{L}e + \lambda g\mathcal{L}g + \lambda s\mathcal{L}s + \lambda c\mathcal{L}c, \tag{17}$$

where λ_e, λ_g, λ_s, and λ_c are hyperparameters that balance the contributions of each loss term.

4 Experimental Setup

4.1 Evaluation Datasets

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [6] is a benchmark dataset extensively used in emotion recognition research. It comprises approximately 12 hours of multimodal data collected from 10 professional actors (5 males and 5 females). Data collection involved five dyadic sessions, each including one male and one female actor performing scripted dialogues and engaging in improvisational scenarios designed to elicit specific emotional expressions. The audio was captured using two microphones at a 48 kHz sampling rate and subsequently downsampled to 16 kHz to align with common audio processing standards. Each speech utterance in the dataset was segmented and annotated by at least three human evaluators who assigned categorical emotion labels based on perceived emotional content. Following established practices in emotion recognition research [45, 25, 54], we merged the "happy" and "excited" categories into a single "happy" category. This consolidation resulted in four primary emotion labels used in our experiments: happy, sad, angry, and neutral. For the IEMOCAP dataset, we employed

the most commonly used metrics in SER: Unweighted accuracy (UA) and Weighted accuracy (WA).

The Multimodal EmotionLines Dataset (MELD) [42] is another benchmark dataset extensively used in MER research. It comprises approximately 13,000 utterances from over 1,400 dialogues, extracted from the popular television series Friends. The dataset includes high-quality audio recordings. Data collection involved extracting multi-party conversations from the TV series, ensuring a diverse range of natural and spontaneous emotional expressions. The audio data is synchronized with the textual transcriptions to provide a comprehensive multimodal representation of each utterance. For consistency with common audio processing standards and our approach with the IEMOCAP dataset, we downsampled the audio to 16 kHz in our study. Each utterance in the dataset was annotated by multiple human evaluators who assigned categorical emotion labels based on perceived emotional content. The annotation process resulted in seven primary emotion categories: anger, disgust, sadness, joy, surprise, fear, and neutral. Due to "disgust" and "fear" each constituting less than 3% of the data, the model achieved 0% accuracy on these classes, leading to a low UA. Therefore, we present the WA and Weighted F1, aligning with previous works [18, 50] that use these metrics.

4.2 Model Configuration

We implemented our proposed models using the PyTorch framework [39] and the Huggingface Transformers library [62]. For the acoustic feature extractor, we utilized the HuBERT-large model, which was pre-trained on the 60,000-hour Libri-Light dataset. This model consists of seven CNN layers that transform raw waveforms into latent representations, followed by 24 Transformer layers that capture underlying representation from speech. The semantic feature extractor employed is BERT-base [12], pre-trained on BooksCorpus and English Wikipedia text passages, comprising 12 Transformer layers to extract semantic embeddings from textual inputs. The hidden layer dimensions D for HuBERT-large and BERT-base are 1,024 and 768, respectively. For the ASR error correction (AEC) module, we used a BERT-base model augmented with 8 additional Transformer layers. Initially, we fine-tuned the AEC module on ASR transcriptions extracted from both the Common Voice and IEMOCAP dataset [2] using the HuBERT model, which was previously fine-tuned on the IEMOCAP dataset. The fine-tuning was performed using the ground-truth text from Common Voice to correct the ASR transcriptions. When integrating the AEC into our multi-level SER framework, we froze the entire AEC module to serve as a post-processing step for the ASR transcriptions during inference.

In accordance with established studies [47, 26], we conducted five-fold speaker-independent cross-validation on the IEMOCAP dataset. All datasets

were downsampled to 16 kHz to unify the sampling rate during data preprocessing. To accommodate varying input lengths for both HuBERT and BERT, we applied sentence padding within each mini-batch. During training, we fine-tuned the pre-trained model for 100 epochs. For the two-stage fine-tuning experiments, the model was trained for 50 epochs in the second stage. We froze the CNN layers in HuBERT while fine-tuning the Transformer layers of both HuBERT and BERT simultaneously. The learning rate was set to 1×10^{-5}, and the mini-batch size was 2 with a gradient accumulation of 8, resulting in a batch size of 16. For the multi-attribute learning setup, we set all auxiliary task weights (λ_{ASR}, λ_{AEC}, and λ_c in stage 1; λ_g, λ_s, and λ_c in stage 2) to 0.1, while maintaining $\lambda_e = 1$ for the primary SER task. This weight configuration was determined empirically by comparing weights of 1, 0.1, and 0.01, where 0.1 achieved the best performance for SER. This setting ensures auxiliary tasks contribute to training without overshadowing the primary objective, consistent with prior findings [7].

To evaluate our models, we employed specific metrics for ASR and SER. For ASR, we used the Word Error Rate (WER). For SER, we adopt weighted accuracy (WA), unweighted accuracy (UA), and F1-score. The WA is calculated as:

$$\text{WA} = \frac{\text{Total Correct Predictions}}{\text{Total Samples}} \times 100\%. \tag{18}$$

The UA is given by:

$$\text{UA} = \frac{1}{C} \sum_{c=1}^{C} \frac{\text{Correct Predictions in Class } c}{\text{Total Samples in Class } c} \times 100\%, \tag{19}$$

where C is the number of emotion classes. The F1 score is the harmonic mean of precision and recall, weighted by the number of samples in each class:

$$\text{F1} = \sum_{c=1}^{C} w_c \times \text{f1}_c, \tag{20}$$

where w_c is the proportion of samples in class c, and f1_c is the F1 score for class c.

5 Results and Analysis

In this section, we present a comprehensive series of experiments designed to evaluate the contributions of each component within our SER system. We first established baselines by evaluating multi-attribute learning performance using speech input alone, text input alone, and a combination of both speech and text inputs. Next, we conducted an overall ablation study to assess the

effectiveness of each component of our approach. In the third experiment, we compared our two-stage fine-tuning method with conventional MTL, where ASR, SER, gender, and style tasks are trained simultaneously. The fourth experiment involved comparing our TGF method with traditional fusion techniques such as simple concatenation. Finally, we analyzed the impact of AEC and contrastive learning (CL) on SER, as both methods aim to mitigate transcription errors introduced by the ASR module. The following subsections provide detailed descriptions of each experiment and discuss the results obtained.

5.1 Multi-attribute Learning for Emotion Recognition

We first analyze the impact of different auxiliary tasks on emotion recognition using various input modalities using the IEMOCAP dataset. Table 1 presents the results of speech input only (SER), which utilizes HuBERT exclusively for feature extraction. To provide a more comprehensive comparison, Table 2 shows the results using ground-truth text input for text emotion recognition (TER), while Table 3 details the performance of models that integrate both speech and ground-truth text inputs for MER.

Table 1: Multi-attribute learning results of speech input on the IEMOCAP dataset.

Exp	Task				SER		ASR	Gender	Style
	SER	ASR	Gender	Style	UA	WA	WER	UA	UA
1	✓				70.32	70.84	-	-	-
2	✓	✓			75.28	75.13	13.57	-	-
3	✓		✓	✓	71.43	71.18	-	85.22	80.63
4	✓	✓	✓	✓	**77.02**	**76.79**	13.75	99.14	87.45

Table 2: Multi-attribute learning results of text input on the IEMOCAP dataset.

Exp	Task			TER		Gender	Style
	TER	Gender	Style	UA	WA	UA	UA
5	✓			67.13	66.85	-	-
6	✓	✓		66.82	65.74	68.41	-
7	✓		✓	**68.79**	**68.27**	-	91.94
8	✓	✓	✓	68.35	67.31	67.96	90.80

From Table 1, we observe that including the ASR task significantly enhances SER performance: comparing the SER results of Exp.-1 and Exp.-2, the UA improved 4.96%, indicating that learning linguistic information for

Table 3: Multi-attribute learning results of multimodal input on the IEMOCAP dataset.

Exp	Task				MER		ASR	Gender	Style
	MER	ASR	Gender	Style	UA	WA	WER	UA	UA
9	✓				72.15	73.07	-	-	-
10	✓	✓			77.64	77.36	13.64	-	-
11	✓		✓	✓	73.44	72.58	-	85.81	82.53
12	✓	✓	✓	✓	**79.18**	**78.86**	14.12	97.52	89.29

HuBERT is crucial for improving SER. Interestingly, even for tasks unrelated to linguistic content, such as gender recognition, incorporating the ASR task yields notable gains. We reason that fine-tuning the HuBERT for simple tasks (four-way classification for SER, binary classification for gender and style recognition) makes the model prone to overfitting; incorporating ASR helps mitigate this issue. Comparing Exp.-4 with Exp.-2, we confirm that learning gender and style information further improves SER performance, which aligns with previous studies [7, 50]. Moreover, in Table 2, comparing Exp.-5 and Exp.-7 shows that including style recognition helps in extracting emotional information from text input, increasing the UA for 1.66%; on the other hand, gender recognition does not benefit emotion recognition from text. In Table 3, comparing Exp.-9 and Exp.-10 reveals that even with BERT extracting emotional information from text, incorporating ASR tasks for HuBERT significantly enhances SER performance, increasing the UA by 5.49%. This result underscores the importance of using ASR to prevent overfitting in fine-tuning acoustic SSL models. Exp.-10 and Exp.-12 demonstrate that for multimodal inputs, integrating gender and style recognition tasks also benefits emotion recognition, yielding an additional improvement in UA by around 1.54%.

5.2 Evaluation of Multi-level SER

To enhance SER without relying on ground-truth text, we conducted multi-level SER by leveraging BERT to explicitly extract emotional information from ASR-decoded transcriptions. An ablation study for each component of the proposed system using the IEMOCAP dataset is provided in Table 4.

A comparison between Exp.-13 and Exp.-2 reveals that incorporating emotional information extracted from transcriptions effectively complements acoustic features, significantly enhancing SER. Specifically, the UA improves 1.97%, underscoring the effectiveness of multi-level SER. Furthermore, when comparing Exp.-13 with Exp.-10, which use ground-truth text, the UA remains comparable (77.25% vs. 77.64%). This finding suggests that with a high-quality ASR system (WER of around 13%), the ASR transcriptions can serve as a reliable substitute for ground-truth text in emotion recognition.

Table 4: Evaluation of the proposed multi-level SER on the IEMOCAP dataset.

Exp	TGF	Two-stage	AEC	CL	SER		ASR
					UA	WA	WER
13					77.25	76.86	13.61
14	✓				79.62	79.50	13.90
15		✓			79.15	78.39	14.53
16			✓	✓	76.41	76.83	12.96
17	✓	✓			**81.17**	**80.16**	14.58
18	✓	✓	✓	✓	80.21	79.35	13.27

Starting from Exp.-14, we enhance the multi-level SER using several components. In Exp.-14, we integrate acoustic and linguistic features using the proposed TGF, which addresses modality incongruity by learning the importance of acoustic and linguistic features in SER, leading to a significant improvement in UA to 79.62%. In Exp.-15, we employ the two-stage fine-tuning process. In the second stage, we fine-tune adapters for multi-attribute learning, integrating style and gender recognition tasks. This proposed two-stage fine-tuning approach achieves a UA of 79.15%, which is comparable to the MER results in Table 3. Exp.-16 examines the impact of applying AEC before inputting text into BERT and utilizing CL between the features learned from the real text and ASR decoded transcription. Even though we observe 4.78% relative improvement on the WER, both AEC and CL does not yield a noticeable improvement in SER. According to the previous conclusion, the ASR transcription is sufficiently accurate for learning emotional information, rendering AEC and CL less impactful. In Exp.-17, combining TGF and two-stage fine-tuning results in the highest UA of 81.17%, demonstrating the effectiveness of our proposed SER approach. Compared with the single-modal baseline using only SER task (Exp.-1), we achieved an absolute improvement of 10.85% and 9.32% on UA and WA, respectively. Finally, Exp.-18 includes AEC and CL methods but does not lead to further performance gains, with UA slightly decreasing to 80.89%. To investigate the impact of AEC and CL, we introduce another dataset MELD, which have much lower performance of ASR due to the recording condition. More detailed evaluation and analysis for each module are provided in the following sections.

5.3 Impact of Two-stage Fine-tuning

We compared two fine-tuning strategies for SER using the IEMOCAP dataset: (1) fine-tuning all parameters of the Transformers in HuBERT and BERT, and (2) adapter tuning, where the feature extractors are frozen and only the adapter modules are trained. As shown in Table 4, the effectiveness

of two-stage fine-tuning in addressing gradient conflicts can be evaluated by comparing with Exp.-4 in Table 1.

As shown in Table 5, in Exp.-19 and Exp.-20, we assessed the performance of adapter tuning. The results indicated that training only the adapters led to inferior SER performance compared to fine-tuning all parameters. This suggests that fine-tuning the Transformers is crucial for capturing discriminative emotion information. Building on these findings, we first fine-tuned the entire HuBERT and BERT models, then explored different fine-tuning strategies in the second stage. In Exp.-21 and Exp.-22, we continued to fine-tune all Transformer parameters in the second stage, which did not enhance SER performance. Moreover, Exp.-22 exhibited a decline in performance compared to the initial fine-tuning step, further implying that fine-tuning HuBERT on simpler tasks without ASR can lead to overfitting. In Exp.-23, we adopted adapter tuning in the second stage. This approach enabled the model to effectively learn gender and style information, resulting in the best SER performance among all experiments. Compared to Exp.-24, where all parameters were fine-tuned in both stages, using adapter tuning in the second stage allowed the model to learn emotion-related information and avoid information loss.

Table 5: Comparison of fine-tuning Transformers and adapter tuning in two-stage fine-tuning using IEMOCAP dataset.

Exp	Fine-tuning strategies		Task				SER		ASR
	Stage 1	Stage 2	SER	ASR	Gender	Style	UA	WA	WER
19	Adapter	-	✓	✓			72.15	71.89	25.84
20	Adapter	-	✓		✓	✓	69.74	70.23	-
21	All params	All params	✓	✓			77.16	77.28	13.58
22	All params	All params	✓		✓	✓	73.71	72.19	-
23	All params	Adapter	✓	✓			77.32	77.15	13.52
24	All params	Adapter	✓		✓	✓	**79.15**	**78.39**	-

5.4 Impact of Feature Fusion Methods

In this section, we conduct multi-level SER and explore the impact of fusion approaches for acoustic and linguistic features including simple concatenation, cross-attention, and the proposed TGF.

As shown in Table 6, we apply self-attention followed by concatenation in Exp.-25 as baseline approach. In Exp.-26, incorporating cross-attention leads to a significant improvement, increasing the UA for 0.75%, highlighting the benefits of attention for information interactions. In Exp.-27, introducing conventional gated fusion results in a limited increase over baseline of 78.22%

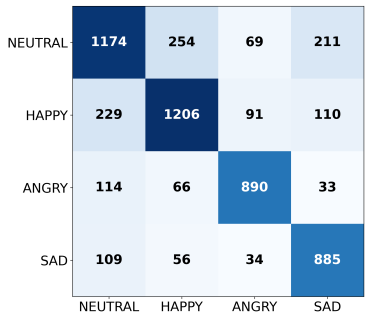
Table 6: Comparison of feature fusion methods on the IEMOCAP dataset.

Exp	Cross-att	Gated	TGF	UA	WA
25	-	-	-	78.10	77.58
26	✓	-	-	78.85	78.54
27	-	✓	-	78.22	78.37
28	✓	-	✓	**79.62**	**79.50**
29	✓	✓	✓	79.35	78.71

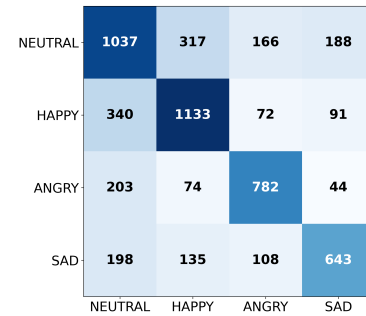
on UA. Notably, Exp.-28, which combines attention with the proposed TGF, achieves the best performance, reaching a UA of 79.62%. This indicates that capturing temporal dynamics during information interaction can benefit SER. Finally, we incorporate additional conventional gated fusion upon the proposed TGF does not yield additional benefits.

Visualization methods are used to further analyze the recognition results and the impact of the proposed TGF in balancing acoustic and linguistic features. In Figure 2, we provide the confusion matrix of a) emotion recognition result using acoustic feature, b) emotion recognition result using linguistic feature, c) multi-level SER result using concatenated acoustic and linguistic features, and d) multi-level SER result, using the proposed TGF method. Comparing a) and b) in Figure 2, acoustic features contain more discriminative features for SER than linguistic feature and generate better performance on all four emotions. Furthermore, concatenating those features leads to more accurate SER in c). As depicted in d), we find that the proposed TGF improved most on happy and angry. Comparing c) and d), we find a significant improvement in the misclassification of happy (78 samples) and angry (35 samples) to neutral. This result validates that TGF benefits the information interaction in feature fusion and results in better classification results in those emotions.

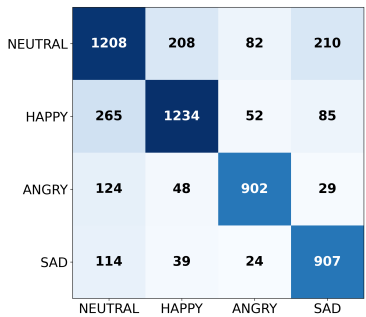
The kernel density estimation distribution of gate values of TGF for each emotion is also provided in Figure 3. Specifically, the blue and green lines denote the gate values of $\mathbf{F}_a^s$ and $\mathbf{F}_t^s$, which contain acoustic and linguistic information, respectively. Since the gate value of $\mathbf{F}_{a\leftarrow t}^c$ and $\mathbf{F}_{t\leftarrow a}^c$, which combine that information, are totally the reverse, we do not show their distribution. In Figure 3, we first find that for happy and angry, the gate value for F_a^s (blue line) gathered around 0.8. This suggests that acoustic features are more important for distinguishing between happy and angry. This is because differences in arousal information are more easily discerned from speech. This conclusion is consistent with the previous analyses. As for neutral and sad, the gate value shows more balancing distribution.



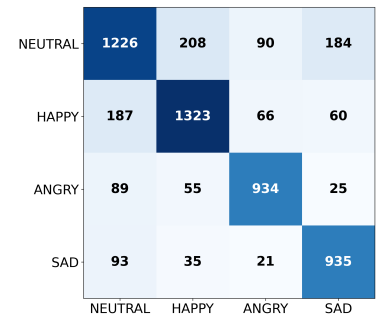
(a) Acoustic feature



(b) Linguistic feature



(c) Multi-level SER

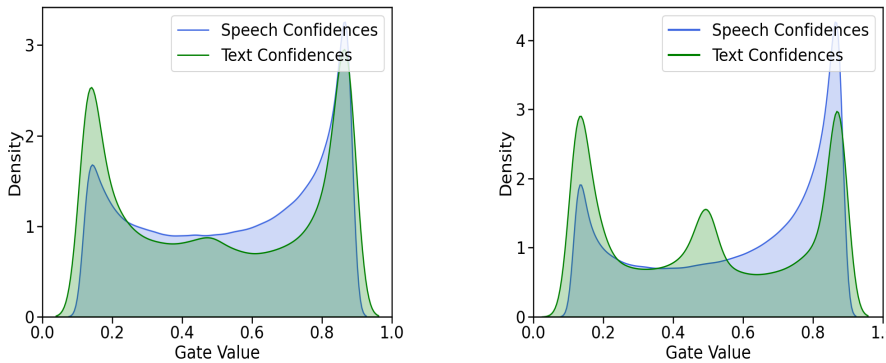


(d) Proposed

Figure 2: Confusion matrix of IEMOCAP dataset. (a) Confusion matrix of emotion recognition result using the acoustic feature. We conduct SER and ASR using HuBERT. (b) Confusion matrix of emotion recognition result using linguistic feature. We conduct emotion recognition using BERT, with ASR decoded transcriptions. (c) Confusion matrix of multi-level SER result, which concatenates acoustic and linguistic features. (d) Confusion matrix of Multi-level SER result, which combines the features using the proposed TGF method.

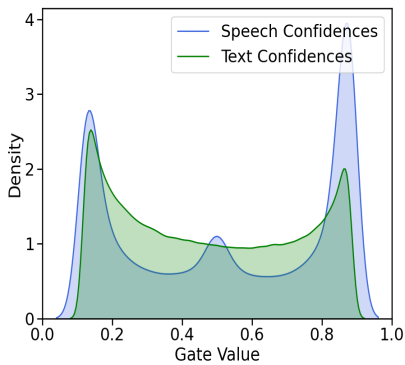
5.5 Impact of Bridging ASR Errors

While the proposed AEC and CL modules for mitigating transcription errors did not show significant improvement on the IEMOCAP dataset, we present experimental results on the MELD dataset to evaluate these methods on noisy emotional speech. As shown in Table 7, baseline model (Exp.-30) incorporat-

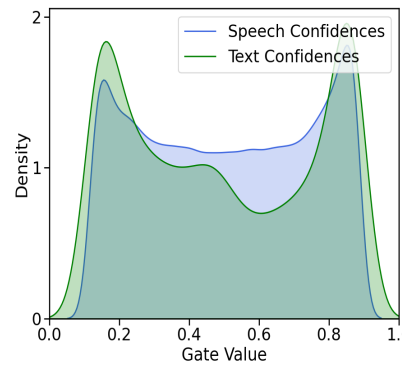


(a) Gate value of neutral

(b) Gate value of happy



(c) Gate value of angry



(d) Gate value of sad

Figure 3: The kernel density estimation distribution of gate value of each emotion. In this figure, blue line plots the gate balancing F_a and $\mathbf{F}_{a \leftarrow t}$, and green line plots the gate balancing F_t and $\mathbf{F}_{t \leftarrow a}$.

ing MTL for SER and ASR results in 48.25% WA. Compared with Exp.-29, introducing BERT to extract linguistic feature from transcription does not improve SER. As the result of TER (Exp.-32) and MER (Exp.-33) is much higher than multi-level SER (Exp.-31), we provide an ablation study for AEC and CL using multi-level SER.

In Table 8, the AEC module achieved 8.52% relative improvement on the WER. Moreover, both AEC and CL results in more than 1.38% improvement over the baseline. Combining those methods by implementing CL to the feature learned from corrected transcription (Exp.-37), we achieved the best

Table 7: Experimental result on the MELD dataset using speech (SER), ground-truth text (TER), and MER with both speech and ground-truth text.

Exp	Input	Approach	WA	F1
30	Speech	SER	48.25	47.29
31	Speech	Multi-level SER	48.75	48.22
32	Text	TER	53.44	52.22
33	Speech & Text	MER	**56.09**	**53.93**

Table 8: Ablation study on the MELD dataset for AEC and CL.

Exp	AEC	CL	SER		ASR
			WA	F1	WER
34	-	-	48.75	48.22	32.45
35	✓	-	50.13	48.75	29.68
36	-	✓	51.78	50.04	32.94
37	✓	✓	**53.35**	**51.02**	30.16

performance on this dataset, attaining a WA of 53.35% and an F1 of 51.02%. This result indicates that our proposed method of bridging ASR errors is effective for SER in scenarios where ASR transcription quality is low.

6 Conclusion

We propose a SER system that directly captures acoustic and linguistic features from speech without relying on ground-truth text. Our approach integrates HuBERT and BERT models: HuBERT extracts acoustic emotional cues and performs ASR, implicitly learning linguistic information; the decoded transcriptions are then input into BERT to explicitly encode linguistic emotional features.

To leverage emotional information in speech, we incorporate multi-attribute learning of gender and emotional expression style through a two-stage fine-tuning process. Initially, HuBERT is fine-tuned jointly for SER and ASR tasks. Subsequently, adapter tuning is applied to both HuBERT and BERT to learn gender and style information, effectively avoiding gradient conflicts and minimizing potential information loss. We introduce a TGF method to combine acoustic and linguistic features, addressing modality incongruity and preserving temporal dynamics through self-attention within each modality and cross-attention between modalities, followed by a fine-grained gating mechanism. To mitigate the impact of ASR errors on linguistic emo-

tional feature extraction, we propose two strategies: 1) an error correction module employing BERT to refine decoded transcriptions before input into the linguistic feature extractor; and 2) a contrastive learning approach that enhances robustness against transcription errors by reducing the distance between features learned from ASR transcriptions and those from ground-truth text.

While incorporating the AEC module did not improve SER on the IEMOCAP dataset, significant improvements were observed on the noisy MELD dataset, indicating that addressing ASR errors is particularly beneficial for SER performance in noisy speech lacking textual input. To address the system complexity concerns, we leverage the same BERT model for both semantic feature extraction and AEC encoding to reduce number of parameters, and employ parameter-efficient adapters rather than full fine-tuning. Using separate models for AEC and semantic feature extraction may yield better performance but would increase model complexity. Furthermore, the modular design allows selective component usage based on application requirementsfor instance, the AEC module can be omitted in clean environments to reduce computational cost.

References

[1] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models", *Sensors*, 21(4), 2021, 1249.

[2] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus", *arXiv preprint arXiv:1912.06670*, 2019.

[3] G. Assunção, B. Patrão, M. Castelo-Branco, and P. Menezes, "An overview of emotion in artificial intelligence", *IEEE Transactions on Artificial Intelligence*, 3(6), 2022, 867–86.

[4] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network", in *2017 international conference on platform technology and service (PlatCon)*, IEEE, 2017, 1–5.

[5] T. Baltruaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy", *IEEE transactions on pattern analysis and machine intelligence*, 41(2), 2018, 423–43.

[6] 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*, 42(4), 2008, 335–59.

[7] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning.", in *Interspeech*, Brno, 2021, 4508–12.

[8] A. P. Chaves and M. A. Gerosa, "How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design", *International Journal of Human–Computer Interaction*, 37(8), 2021, 729–58.

[9] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning", in *Proceedings of the 25th international conference on Machine learning*, 2008, 160–7.

[10] C. Colón-Ruiz and I. Segura-Bedmar, "Comparing deep learning architectures for sentiment analysis on drug reviews", *Journal of Biomedical Informatics*, 110, 2020, 103539.

[11] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction", *IEEE Signal processing magazine*, 18(1), 2001, 32–80.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", *arXiv preprint arXiv:1810.04805*, 2018.

[13] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", *Pattern recognition*, 44(3), 2011, 572–87.

[14] R. Feldman, "Techniques and applications for sentiment analysis", *Communications of the ACM*, 56(4), 2013, 82–9.

[15] Y. Gao, H. Shi, C. Chu, and T. Kawahara, "Enhancing Two-Stage Finetuning for Speech Emotion Recognition Using Adapters", in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, 11316–20.

[16] Y. Gao, H. Shi, C. Chu, and T. Kawahara, "Speech Emotion Recognition with Multi-level Acoustic and Semantic Information Extraction and Interaction", in *Interspeech*, 2024, 1060–4.

[17] I. Gat, H. Aronowitz, W. Zhu, E. Morais, and R. Hoory, "Speaker normalization for self-supervised speech emotion recognition", in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, 7342–6.

[18] L. Guo, L. Wang, J. Dang, Y. Fu, J. Liu, and S. Ding, "Emotion recognition with multimodal transformer fusion framework based on acoustic and lexical information", *IEEE MultiMedia*, 29(2), 2022, 94–103.

[19] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units", *IEEE/ACM transactions on audio, speech, and language processing*, 29, 2021, 3451–60.

[20] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", in *Proceedings of naacL-HLT*, Vol. 1, Minneapolis, Minnesota, 2019, 2.

[21] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review", *IEEE access*, 7, 2019, 117327–45.

[22] V. S. Kodiyala and R. E. Mercer, "Emotion recognition and sentiment classification using bert with data augmentation and emotion lexicon enrichment", in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2021, 191–8.

[23] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review", *International journal of speech technology*, 15(2), 2012, 99–117.

[24] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Multitask learning from augmented auxiliary data for improving speech emotion recognition", *IEEE Transactions on Affective Computing*, 14(4), 2022, 3164–76.

[25] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study", *arXiv preprint arXiv:1712.08708*, 2017.

[26] Y. Li, P. Bell, and C. Lai, "Fusing ASR Outputs in Joint Training for Speech Emotion Recognition", in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, 7362–6.

[27] Y. Li, T. Zhao, T. Kawahara, *et al.*, "Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning.", in *Interspeech*, 2019, 2803–7.

[28] B. Liang, C. Lou, X. Li, L. Gui, M. Yang, and R. Xu, "Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs", in *Proceedings of the 29th ACM international conference on multimedia*, 2021, 4707–15.

[29] P. Liu, K. Li, and H. Meng, "Group Gated Fusion on Attention-Based Bidirectional Alignment for Multimodal Emotion Recognition", in *Proc. Interspeech*, 2020, 379–83.

[30] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "Smil: Multimodal learning with severely missing modality", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, No. 3, 2021, 2302–10.

[31] R. K. Mahabadi, S. Ruder, M. Dehghani, and J. Henderson, "Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks", *arXiv preprint arXiv:2106.04489*, 2021.

[32] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection", *IEEE transactions on affective computing*, 14(3), 2022, 1743–53.

[33] W. McCallum, S. Farmer, and P. Pocock, "The effects of physical and semantic incongruites on auditory event-related potentials", *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 59(6), 1984, 477–88.

[34] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, *et al.*, "Self-supervised speech representation learning: A review", *IEEE Journal of Selected Topics in Signal Processing*, 16(6), 2022, 1179–210.

[35] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning", in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, 689–96.

[36] H. Pan, Z. Lin, P. Fu, Y. Qi, and W. Wang, "Modeling intra and inter-modality incongruity for multi-modal sarcasm detection", in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, 1383–92.

[37] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition", in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, 7169–73.

[38] M. A. Pastor, D. Ribas, A. Ortega, A. Miguel, and E. Lleida, "Cross-corpus speech emotion recognition with HuBERT self-supervised representation", in *IberSPEECH 2022*, ISCA, 2022, 76–80.

[39] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library", *Advances in neural information processing systems*, 32, 2019.

[40] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings", in *Proc. Interspeech*, 2021, 3400–4.

[41] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos", in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2017, 873–83.

[42] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations", *arXiv preprint arXiv:1810.02508*, 2018.

[43] X. Qin, Z. Wu, T. Zhang, Y. Li, J. Luan, B. Wang, L. Wang, and J. Cui, "Bert-erc: Fine-tuning bert is enough for emotion recognition in conversation", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, No. 11, 2023, 13492–500.

[44] S. Ramakrishnan and I. M. El Emary, "Speech emotion recognition approaches in human computer interaction", *Telecommunication Systems*, 52(3), 2013, 1467–78.

[45] V. Rozgi, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, "Ensemble of svm trees for multimodal emotion recognition", in *Proceedings of the 2012 Asia Pacific signal and information processing association annual summit and conference*, IEEE, 2012, 1–4.

[46] A. C. Samson, C. F. Hempelmann, O. Huber, and S. Zysset, "Neural substrates of incongruity-resolution and nonsense humor", *Neuropsychologia*, 47(4), 2009, 1023–33.

[47] A. Satt, S. Rozenberg, and R. Hoory, "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms.", in *Proc. Interspeech*, 2017, 1089–93.

[48] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture", in *2004 IEEE international conference on acoustics, speech, and signal processing*, Vol. 1, IEEE, 2004, I–577.

[49] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization", *Advances in neural information processing systems*, 31, 2018.

[50] M. Sharma, "Multi-lingual multi-task speech emotion recognition using wav2vec 2.0", in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, 6907–11.

[51] H. Shi, Y. Gao, Z. Ni, and T. Kawahara, "Serialized Speech Information Guidance with Overlapped Encoding Separation for Multi-Speaker Automatic Speech Recognition", in *2024 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2024, 193–9.

[52] H. Shi and T. Kawahara, "Dual-path Adaptation of Pretrained Feature Extraction Module for Robust Automatic Speech Recognition", in *Interspeech*, 2024, 2850–4.

[53] H. Shi, M. Mimura, and T. Kawahara, "Waveform-Domain Speech Enhancement Using Spectrogram Encoding for Robust Speech Recognition", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, 2024, 3049–60.

[54] X. Shi, S. Li, and J. Dang, "Dimensional Emotion Prediction Based on Interactive Context in Conversation.", in *INTERSPEECH*, 2020, 4193–7.

[55] X. Shi, X. Li, and T. Toda, "Multimodal fusion of music theory-inspired and self-supervised representations for improved emotion recognition", in *Proc. Interspeech*, 2024, 2024–350.

[56] A. P. Simpson, "Dynamic consequences of differences in male and female vocal tract dimensions", *The journal of the Acoustical society of America*, 109(5), 2001, 2153–64.

[57] T. Standley, A. Zamir, D. Chen, L. Guibas, J. Malik, and S. Savarese, "Which tasks should be learned together in multi-task learning?", in *International conference on machine learning*, PMLR, 2020, 9120–32.

[58] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review", *International Journal of Speech Technology*, 21, 2018, 93–120.

[59] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation", *arXiv preprint arXiv:2202.12233*, 2022.

[60] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences", in *Proceedings of the conference. Association for computational linguistics. Meeting*, Vol. 2019, NIH Public Access, 2019, 6558.

[61] 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*, 9, 2021, 47795–814.

[62] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, *et al.*, "Transformers: State-of-the-art natural language processing", in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, 38–45.

[63] Y. Wu, Y. Zhao, X. Lu, B. Qin, Y. Wu, J. Sheng, and J. Li, "Modeling incongruity between modalities for multimodal sarcasm detection", *IEEE MultiMedia*, 28(2), 2021, 86–95.

[64] R. Xia and Y. Liu, "Using i-Vector Space Model for Emotion Recognition.", in *Interspeech*, 2012, 2230–3.

[65] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, *et al.*, "Superb: Speech processing universal performance benchmark", *arXiv preprint arXiv:2105.01051*, 2021.

[66] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech Emotion Recognition Using Spectrogram & Phoneme Embedding.", in *Interspeech*, 2018, 3688–92.

[67] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning", *Advances in Neural Information Processing Systems*, 33, 2020, 5824–36.

[68] B. Zhang, E. M. Provost, and G. Essl, "Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences", *IEEE Transactions on Affective Computing*, 10(1), 2017, 85–99.

[69] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, and X. Zhao, "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects", *Expert Systems with Applications*, 2023, 121692.

[70] Y. Zhang and Q. Yang, "A survey on multi-task learning", *IEEE transactions on knowledge and data engineering*, 34(12), 2021, 5586–609.