






# Multimodal Emotion Recognition With Temporal and Semantic Consistency

Bingzhi Chen , Qi Cao, Mixiao Hou , Zheng Zhang , Senior Member, IEEE, Guangming Lu , Member, IEEE, and David Zhang , Life Fellow, IEEE

**Abstract**—Automated multimodal emotion recognition has become an emerging but challenging research topic in the fields of affective learning and sentiment analysis. The existing works mainly focus on developing multimodal fusion strategies to incorporate different emotion-related features. However, they fail to explore the inherent contextual consistency to reconcile the emotional information across modalities. In this paper, we propose a novel Time and Semantic Interaction Network (TSIN), which concurrently incorporates the advantages of temporal and semantic consistency into the multimodal emotion recognition task. Specifically, a well-designed Speech and Text Embedding (STE) module is devoted to formulating the initial embedding spaces by respectively building the modality-specific representations of speech and text. Instead of separately learning or directly fusing the acoustic and textual features, we propose a well-defined Time and Semantic Interaction (TSI) module to conduct the emotional parsing and sentiment refining by performing the fine-grained temporal alignment and cross-modal semantic interaction. Benefitting from temporal and semantic consistency constraints, both speech-text embeddings can be interactively optimized and fine-tuned in the learning process. In this way, the learnt acoustics and textual features can jointly and efficiently predict the final emotional state. Extensive experiments on the IEMOCAP dataset demonstrate the superiorities of our TSIN framework in comparison with state-of-the-art baselines.

**Index Terms**—Multi-label image recognition, graph convolutional network, semantic-interactive, label co-occurrence, semantic similarity.

## I. INTRODUCTION

AUTOMATED emotion recognition has undoubtedly become one of the most popular research topics in the emerging fields of affective computing and sentiment analysis [1] for many human-computer interaction systems. With the arrival of a large amount of social media data [2], it is of great significance to develop the application of automated emotion recognition for improving the accuracy and efficiency of understanding emotions. It can derive the essence from applying emotion-defined technology to different areas, and make the machine identify and understand humans' emotional states with supportive feedback. While some advanced techniques [3], [4] have been proposed to improve the recognition accuracy of sentiment polarities, i.e., polarizing positive and negative sentiments, identifying finer-grained emotions remains challenging in the absence of definitive criteria. To this end, the current task of automated emotion recognition [5], [6] is to classify the media samples into individual emotional classes, such as anger, happiness, sadness, neutrality, which can provide fine-grained decision-making and higher quality human-computer interaction experiences.

In recent years, data-driven deep learning techniques [7] have made a remarkable breakthrough in a range of emotion recognition applications. The early methods for studying emotional features are generally built on unimodal data sources, e.g., voice, facial expression, or linguistic content. Prior works [8], [9] on speech emotion recognition mainly focus on acoustic characteristics deriving from speech. Although making some progress, only using unimodal emotional information might be insufficient to determine the emotional states since it may not contain a complete semantic entity. Inspired by the ability of human cognition that can make full use of information from multiple perception ways, researchers [10]–[12] have attempted to extend the traditional unimodal emotion analysis to complex multimodal cooperation.

Compared to unimodal methods, multimodal emotion analysis has been demonstrated to its superiority of multi-source knowledge. Both audio and text in the multimodal data can provide important cues to better identify the true affective states of the opinion holder. To this end, a combination of text and audio data has great potential to create a better emotion and sentiment analysis model. However, a key overarching challenge of

Manuscript received January 14, 2021; revised July 30, 2021 and November 11, 2021; accepted November 12, 2021. Date of publication November 19, 2021; date of current version December 5, 2021. This work was supported in part by the National Key Research and Development Program of China under Project 2018AAA0100100, in part by NSFC under Grants 62176077 and 62002085, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2019B1515120055, in part by Shenzhen Key Technical Project under Grant 2020N046, in part by Shenzhen Fundamental Research under Grants JCYJ20210324132210025 and GXWD20201230155427003-20200824103320001, in part by Open Project Fund (AC01202005018) from the Shenzhen Institute of Artificial Intelligence, and Robotics for Society, and in part by the Medical Biometrics Perception and Analysis Engineering Laboratory, Shenzhen, China. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Juan Ignacio Godino-Llorente. (Bingzhi Chen and Qi Cao contributed equally to this work.) (Corresponding authors: Zheng Zhang; Guangming Lu.)

Bingzhi Chen is with the Shenzhen Medical Biometrics Perception and Analysis Engineering Laboratory, Harbin Institute of Technology, Shenzhen 518055, China, and also with the School of Software, South China Normal University, Foshan 528200, China (e-mail: chenbingzhi.smile@gmail.com).

Qi Cao, Mixiao Hou, and Guangming Lu are with the Shenzhen Medical Biometrics Perception and Analysis Engineering Laboratory, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: keikei95@126.com; mixiao-hou@163.com; luguangm@hit.edu.cn).

Zheng Zhang is with the Harbin Institute of Technology, Shenzhen & Peng Cheng Laboratory, Nanshan, Shenzhen 518055, China (e-mail: darrenzz219@gmail.com).

David Zhang is with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518055, China (e-mail: davidzhang@cuhk.edu.cn).

Digital Object Identifier 10.1109/TASLP.2021.3129331

speech-text emotion recognition comes from the large discrepancy between speech and text embedding spaces. Therefore, how to effectively integrate both speech and linguistic heterogeneous embedding is of paramount importance to conduct effective multimodal speech-text emotion recognition. To bridge such a gap, various fusion strategies [12], [13] have been proposed to combine the acoustic and textual representations extracted from the speech and text sources. In the literature, the existing fusion strategies fall into three main streams: (1) modality-independent fusion; (2) modality-dependent fusion; (3) word-level fusion.

As a native way to address the multimodal combination, the modality-independent fusion strategy [14] is designed to learn the speech and text embedding spaces independently and directly combine their holistic feature representations for emotion recognition. Such a simple fusion operation may greatly hamper its performance since it only focuses on the decision-level features but ignores the context associations across modalities. Subsequently, a wealth of works [12] pay attention to the modality-dependent fusion strategy, and introduce various attention mechanisms to learn the common emotional signals from the speech and text embedding spaces. By contrast, some word-level fusion approaches [13], [15] deserves further investigation. They aim to leverage the context of the speech and text embedding spaces to distill the word-level emotional information. Although making some progress, they suffer from the problems of semantic exclusion or constituent redundancy. Due to the lack of emotion parsing and refining, different embedding spaces may be incompatible and conflicting. Thus, predicting the correct emotion remains challenging in these scenarios.

Based on the above analysis, two issues should be solved for reconciling the emotional information from different domains. The first issue is *the embedding discrepancy of the temporal scales*. Since the acoustic and textual information is collected from different sources and represented by different descriptors, their embedding spaces can be discrepant on the time scales. Intuitively, we can find that each speech frame can be attached to the corresponding emotion-relevant word in the text sequence, which can neatly assess the embedding space of speech and text. Therefore, a fine-grained temporal alignment mechanism is certainly more effective to conduct the emotional parsing, instead of using the forced alignment. The second issue is *the semantic inconsistency of the contextual contents*. As shown in Fig. 1, a specific word in the text with strong emotional power may change the entire sentimental state. However, some text might not carry the finer emotional information to accurately express the exact state of mood or opinion of a user. Notably, the speaker's unusual tone and intonation can indicate an opposite emotion despite having the same content of the statement. It is difficult to extract accurate emotional features and contextual information from grammatically erroneous texts, short messages, and sarcasm in written documents. Thus, tones or content of words cannot be used independently to reflect emotional information.

To eliminate the semantic exclusion and constituent redundancy, our work conceives the sentiment refining module to preserve the semantic consistent across different embedding spaces. In this paper, we propose a novel multimodal emotion recognition framework, dubbed Time and Semantic Interaction

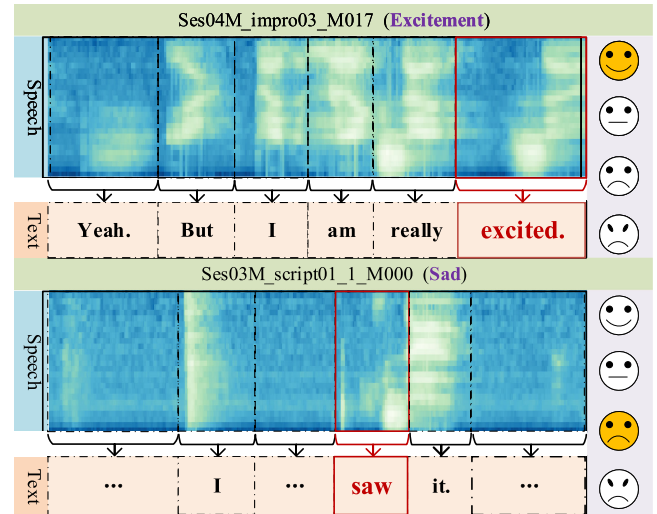


Fig. 1. Illustration of the two multimodal samples selected from IEMOCAP. Emotion-salient speech frames or texts are highlighted in red. Note that the level or intensity of the expressed emotions might differ on a case-to-case basis within a single class [4].

Network (TSIN), which leverages both the temporal and semantic consistency constraints to conduct emotional parsing and sentiment refining. The architecture of our proposed TSIN framework is illustrated in Fig. 2. Specifically, a well-designed Speech and Text Embedding (STE) module is used to construct the corresponding embedding spaces for any input speech-text pair. Moreover, our TSIN framework designs a Time and Semantic Interaction (TSI) module to explore the fine-grained interactions across modalities to jointly optimize these two embedding spaces. In particular, the TSI module first introduces an efficient fine-grained temporal alignment mechanism to reconcile the structural difference on the temporal scale by modeling the time dependence between speech frames and text words. After that, our TSI module employs a novel cross-modal attention mechanism to learn two different groups of semantic-interactive weights, i.e., the text-aware and speech-aware weights, which are used to recalibrate speech and text feature responses, respectively. Benefiting from both the temporal and semantic consistency constraints, the speech-text embedding spaces can be interactively optimized and fine-tuned in the learning process. Furthermore, the optimized speech-text feature representations are integrated into our classification module to jointly predict the emotional states. We verify our proposed TSIN method through extensive experiments and analyses on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [16] dataset. Our main contributions are summarized as follows:

- This paper proposes a novel TSIN framework to simultaneously preserve the temporal and semantic consistency by investigating the fine-grained temporal alignment and cross-modal semantic interaction. To the best of our knowledge, it is the very first attempt to effectively learn the emotional parsing and sentiment refining for multimodal emotion recognition.
- Different from the previous works that directly combine the acoustic and textual features, our TSIN method can efficiently leverage the fine-grained interaction

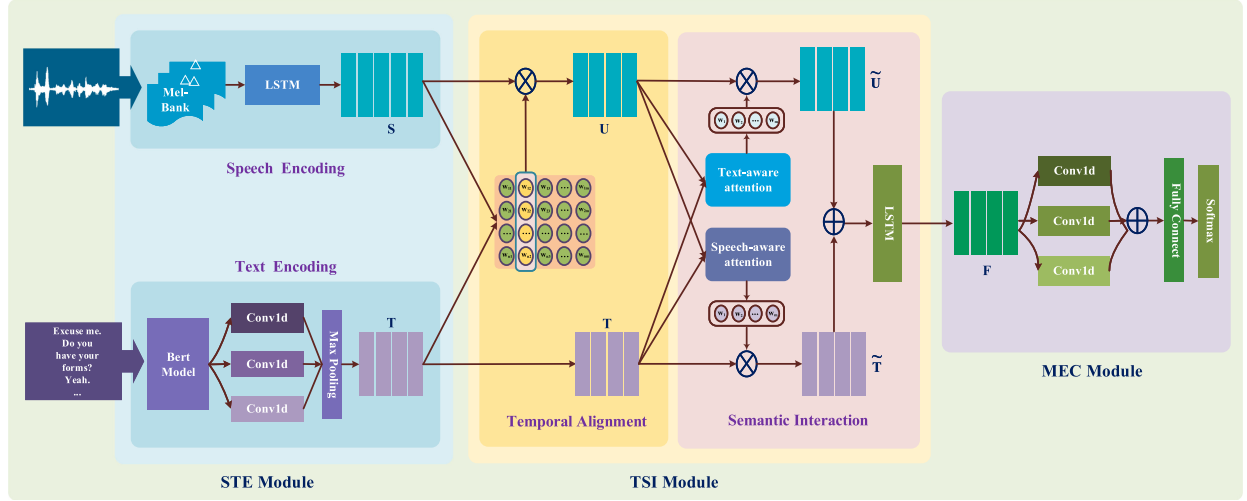


Fig. 2. Illustration of the proposed TSIN framework for multimodal emotion recognition. 1) The STE module utilizes two well-designed encoders to encode the speech and text data into the corresponding embedding spaces  $S$  and  $T$ . 2) Then the TSI module effectively explores the fine-grained interactions across modalities to jointly optimize the speech-text embedding spaces. 3) Finally, the MEC module is applied to the feature fusion representation  $F$  of speech and text embeddings for predicting the emotional states. In this figure, “ $\otimes$ ” and “ $\oplus$ ” represents “cross product” and “concatenation” operations, respectively.

across modalities to solve the temporal discrepancy and semantic inconsistency problems, which guarantees their temporal and semantic consistency.

- Extensive experimental results on IEMOCAP show the superior performance of our TSIN framework in comparison with state-of-the-art methods, which demonstrates the effectiveness of the proposed method.

The rest of this paper is organized as follows. Section II reviews some related works of speech-text emotion recognition, and Section III mainly describes the proposed TSIN method. Next, the comprehensive experiments and visualization analyses are conducted in Section IV. Finally, Section V makes a conclusion of this work.

## II. RELATED WORK

In recent years, many efforts have been devoted to spurring the development of speech and text emotion recognition. In this section, some related works for speech emotion recognition and speech-text emotion recognition are expounded in detail.

### A. Speech Emotion Recognition

Over the past few years, a considerable amount of works have raised to address the task of speech emotion recognition. For example, Mirsamadi *et al.* (WPA) [17] proposed to generate a compact utterance-level representation by using long short-term memory (LSTM) to learn an appropriate temporal aggregation of frame-level acoustic features. Li *et al.* (DRN) [18] leveraged the variants of ResNet and self-attention mechanism to model the dependencies between different space elements in the speech feature sequence. Zhao *et al.* [9] developed an attention-based bidirectional LSTM neural network in combination with a connectionist temporal classification (CTC) objective function for ER. Inspired by the human-human conversation,

Yeh *et al.* (IAAN) [19] presented an interaction-aware attention mechanism to incorporate contextual information into the acoustic representation. Despite obtaining some progress, the performance of these works is still far from satisfactory, due to the problem of the incomplete semantic entity. In most cases, their corresponding average recall rates normally well under 70% on the existing benchmark datasets.

### B. Text Emotion Recognition

In the field of text sentiment analysis, researchers have expended considerable efforts in developing robust models to automatically extract the emotional states by using textual features. For example, Cambria *et al.* [20] utilized top-down and bottom-up learning for text-based sentiment analysis with an ensemble of symbolic and subsymbolic AI tools. Basiri *et al.* [21] presented an attention-based bidirectional CNN-RNN deep model that can effectively assign the weights to different words and sentences for text classification and sentiment analysis. Akhtar *et al.* [22] proposed a multi-layer perception based ensemble technique for predicting the explicit and implicit sentiment in the financial text. Moreover, Song *et al.* [23] introduced the probabilistic linguistic terms sets and the relevant theory into their text representation model (Word2PLTS), improving the performance of short text sentiment analysis. However, the texts may not portray peculiar cues of accurate contextual information to make the decision of emotion states in comparison with multimodal methods.

### C. Speech-Text Emotion Recognition

1) *Modality-Independent Fusion Methods:* The early researches of speech-text emotion recognition mainly focus on modality-independent models, which learn the nonlinear transformations of different modalities separately or directly fuse the acoustic and textual features to predict the emotional states.



For example, Zhou *et al.* (MDNN) [24] proposed a generative model based on semi-supervised variational autoencoder to learn speech features that strengthens the classification ability. Yoon *et al.* (MDRE) [25] attempted to learn speech and text emotional representations separately by using dual branches and then directly combines the emotional information from these sources to predict the emotional states. Moreover, Li *et al.* (PAaAN) [26] built their model by combining different individualized attributes, such as personality, motives, and beliefs. To capture the specific affective factors, Kim *et al.* (DNN-BN) [27] introduced an affective lexicon to capture the emotion-salient words from the text. Although the modality-independent fusion strategy can be considered as a cure-all for feature fusion in the field of multimodal learning, it fails to effectively establish the inner relationship across modalities modes.

2) *Modality-Dependent Fusion Methods*: A variety of attention mechanisms have been introduced into the modality-dependent methods to explore the correlations between speech and text. For example, Lee *et al.* [28] exploited trainable attention mechanisms to capture the nonlinear correlations between modalities, so that the information in the time domain can be retained through the temporal embedding. Huang *et al.* [12] utilized the multi-head attention to produce multimodal emotional intermediate representations from common semantic feature space. Yoon *et al.* (MHA-2) [11] proposed a multi-hop attention mechanism to automatically detect the emotion-relevant segments of the audio and text data for speech-text emotion analysis. Despite some success enabled by these models, these modality-dependent fusion methods fail to take advantage of the inner emotion information of multiple modalities, which plays an important role in emotion recognition tasks.

3) *Word-Level Fusion Methods*: To capture fine-grained emotional features, some methods are designed to explore the context of speech and text embedding to highlight the informative words. For example, Gu *et al.* (FAF) [13] designed a parallel model to employ two concurrent attention mechanisms to select the emotion-salient words in various domains for feature fusion. Chen *et al.* [15] proposed the gated multimodal embedding LSTM with temporal attention model which performs word-level fusion at a finer fusion resolution between input modalities. Aguilar *et al.* (H-MM-4) [29] presented a tandem system to capture important word representations, which first prioritizes one of the modalities at each word step and then combines acoustic and lexical words in a weighted way. Nevertheless, these methods could merely search for emotional words roughly since they take no account of the temporal and semantic consistency constraints. To conduct the appropriate emotion parsing and sentiment refining, the purpose of the proposed TSIN framework is to further implement the fine-grained temporal alignment and cross-modal semantic interaction, which can make the learned acoustic and textual features with the temporal and semantic consistency.

#### D. Multimodal Sentiment Analysis

With the significant increase in the popularity of social media containing multiple modalities of content, sentiment analysis of

multimodal data has attracted increasing interest in recent years. For example, Poria *et al.* [30] utilized multiple kernel learning to combine the emotional information extracted from visual, audio, and textual modalities for predicting the user opinion and emotions. Moreover, Chaturvedi *et al.* [31] incorporated the fuzzy logic into deep convolutional neural networks to predict the degree of a particular emotion in a vector space of affective commonsense. By concurrently exploiting audio, visual, and textual cues. Dashtipour *et al.* [32] presented a context-aware multimodal sentiment analysis framework to accurately predict the emotional states. Moreover, Stappen *et al.* [33] explored a lexica knowledge-based extraction approach that can effectively capture context and emotions to output the emotional valence arousal and speaker topic classes. In this paper, only acoustics and textual features from speech and text modalities are jointly and efficiently designed to effectively enhance the functionality and flexibility of automated multimodal emotion recognition.

### III. METHODOLOGY

In this section, we elaborate a detailed description of the proposed method. The architecture of the proposed TSIN framework is shown in Fig. 2. Our TSIN framework mainly consists of three components, including the STE module, the TSI module, and the MEC module, which would be presented gradually in the following sections.

#### A. Overview of TSIN

Given a pair of speech and text streams as inputs, the main core of the proposed TSIN framework is to perform the emotional parsing and sentiment refining between the acoustics and textual features for improving the performance of automated multimodal emotion recognition. Firstly, the speech and text streams are fed into the speech and text encoders in the STE module to capture the initial feature embedding spaces of  $S$  and  $T$ , respectively. To overcome the above limitations, two well-established emotional parsing processes in the TSI module, i.e., the fine-grained temporal alignment and cross-modal semantic interaction, are performed to achieve the unified representation  $F$  of the speech-text embedding spaces with the temporal and semantic consistency, improving the representation ability of the semantic features. Finally, the unified representation  $F$  is extended by the multi-kernel adaptive weighting block in the MEC module to generate the discriminative emotional representations for predicting the final emotional states. All the components in the TSIN framework will be described in the following sections.

#### B. Speech and Text Embedding Module

It is important to learn discriminative representations of speech and text. In this part, the STE module leverages the well-designed encoders to encode the speech and text data into the corresponding feature vectors, which can be applied to build the feature embedding spaces of speech and text, respectively. Given the multimodal dataset  $D$  with  $n$  speech-text pairs,  $D = \{S_i, T_i\}_{i=1}^n$ , which  $S$  and  $T$  represents the speech and text, respectively.

1) *Speech Encoder*: Benefiting from the development of speech recognition, Mel-filter bank (MFB) [34] has been widely used in speech emotion recognition. While features derived from MFB are quite popular, we argue that they are less representative of real-world speech. To obtain more effective acoustic features, the speech encoder is designed to concurrently integrate the static and dynamic features into a unified representation. In particular, the static features are extracted from log-MFB, which contains sufficient information on the frequency domain. By contrast, the dynamic features leverage the first derivative and the second derivative of log-MFB features to describes the spectral changes between frames and reflect the changing process of emotion. In this way, the obtained acoustics features are flexible and efficient and can meet the complicated situation of the variability in individual expression styles.

Specifically, the hamming window is employed to split the input of speech stream into short frames with 25 ms window width and 10 ms frameshift. Afterward, we utilize the Short-Time Fourier Transform (STFT) to map the time domain signal to frequency. Moreover, we conduct logarithmic operations on the energy spectrum with MFB to capture the static features. Subsequently, a group of triangle filters is designed to obtain the dynamic features. That is, the learned acoustics features are concatenated and represented as  $O = \{o_1, \dots, o_i, \dots, o_n\}$ , where  $n = 500$  is the number of frames. After that, we adopt the LSTM to further encode speech features. The output of the speech encoder is defined as follows:

$$S = LSTM(O) \in \mathbb{R}^{d \times n}, \quad (1)$$

where  $d = 768$  represents the dimensionality of speech and text embedding spaces.

2) *Text Encoder*: Recently, many word embedding methods, such as Glove [35] and BERT [36], have been proved to be powerful in the natural language processing tasks. Follow the previous works, we adopt the pre-trained BERT model to extract word embeddings, which can be represented as  $Q = \{q_1, \dots, q_i, \dots, q_m\}$ , where  $m = 150$  is the number of words.

Considering the phrases in the sentence have various lengths, we propose a simple and effective multi-kernel adaptive weighting block to perceive the words around by explicitly modeling the interdependencies between words and recalibrating the response of keywords. In practice, the proposed block used in TSIN is composed of three one-dimensional convolutions with kernel sizes of 1, 3, and 5. The output of the text encoder is defined as follows:

$$T = f_{max}(C1D_k(Q)) \in \mathbb{R}^{d \times m}, \quad (2)$$

where  $C1D$  indicates one-dimensional convolution and  $f_{max}$  represents the max-pooling layer which is applied for feature extraction.

### C. Time and Semantic Interaction Module

Unlike the previous works that neglect the interaction between speech and text, we aim to leverage these fine-grained interdependences to overcome the limitations of semantic exclusion and redundancy. Here, the TSI module mainly involves

two emotional parsing processes, i.e., the fine-grained temporal alignment and cross-modal semantic interaction, which are used to keep up the temporal and semantic consistency between the speech-text embedding spaces  $S = \{s_1, \dots, s_i, \dots, s_n\}$  and  $T = \{t_1, \dots, t_i, \dots, t_m\}$ .

1) *Learning With Temporal Consistency*: Firstly, we propose a fine-grained temporal alignment mechanism to learn the alignment between speech frames and text words to project the speech frames into the latent space grounded on the text features. Specifically, given an encoded speech embedding  $S$  and text embedding  $T$ , the fine-grained alignment strength between the  $i^{th}$  speech frame and the  $j^{th}$  word is calculated as follows:

$$\alpha_{ij} = t_j^T s_i. \quad (3)$$

Then we can obtain the normalized attention weight over the speech sequence  $\alpha'_{ij}$ ,

$$\alpha'_{ij} = \frac{\exp(\lambda \alpha_{ij})}{\sum_{k=1}^n \exp(\lambda \alpha_{kj})}, \quad (4)$$

where  $\lambda = \frac{1}{\sqrt{d}}$  indicates the scaling factor that is used to prevent the exploding gradient problem. Finally, we compose a speech context feature  $u_j$  for the word  $t_j$  by combining all speech frames,

$$u_j = \sum_{i=1}^n \alpha'_{ij} s_i. \quad (5)$$

In this way, the initial speech embedding space  $S$  is converted to  $U = \{u_1, \dots, u_i, \dots, u_m\} \in \mathbb{R}^{d \times m}$  with temporal consistency.

2) *Learning With Semantic Consistency*: As mentioned above, not all tones or words can reflect emotional information explicitly and consistently. It is known that the emotional similarity between speech and text is essentially dependent on the semantics shared in the different modalities. To ensure the semantic consistency between the speech and text embedding spaces, we are required to further conduct the sentiment refining by exploring the cross-modal semantic interaction. Thus, we proposed a novel cross-modal attention mechanism to learn the text-aware and speech-aware weights  $w = \{w^t, w^s\}$ , which are applied to recalibrate speech and text feature representations, respectively.

Firstly, we extract the hidden semantic states  $h = \{h_t, h_s\}$  from the text and speech embedding spaces,

$$h_t, h_s = f_{mean}(T, U) \quad (6)$$

where  $f_{mean}(\cdot)$  represents the mean-pooling operation.

Then the semantic states are attached to the corresponding acoustic and textual features to obtain the initial weights,

$$\begin{aligned} w_i^t &= V^T \tanh(x_s \cdot u_i + x_t \cdot h_t + b), \\ w_i^s &= V^T \tanh(x_t \cdot t_i + x_s \cdot h_s + b), \end{aligned} \quad (7)$$

where  $V, x_s, x_t$  and  $b$  are trainable parameters. Next, the normalized attention weights  $\tilde{w} = \{\tilde{w}^s, \tilde{w}^t\}$  over the speech sequence can be defined as follows:

$$\tilde{w}_i = \frac{w_i}{\sum_{j=1}^m w_j} \quad (8)$$

Finally, we use the normalized text-aware weights to recalibrate the feature responses,

$$\begin{aligned}\tilde{u}_i &= u_i * \tilde{w}_i^s, \\ \tilde{t}_i &= t_i * \tilde{w}_i^t,\end{aligned}\quad (9)$$

Therefore, the text-aware speech embedding and speech-aware text embedding can be represented as  $\tilde{U} = \{\tilde{u}_1, \dots, \tilde{u}_i, \dots, \tilde{u}_m\} \in \mathbb{R}^{d \times m}$  and  $\tilde{T} = \{\tilde{t}_1, \dots, \tilde{t}_i, \dots, \tilde{t}_m\} \in \mathbb{R}^{d \times m}$ , respectively. Subsequently, both the  $\tilde{U}$  and  $\tilde{T}$  are concatenated and fed into LSTM for feature fusion,

$$F = LSTM(\tilde{U} \oplus \tilde{T}) \in \mathbb{R}^{d' \times m}, \quad (10)$$

where  $d' = 512$  represents the dimensionality of unified representation, and ' $\oplus$ ' indicates the concatenation operation.

#### D. Multimodal Emotion Classification Module

On the basis of the TSI module, the unified representation  $F = \{f_1, \dots, f_i, \dots, f_m\}$  is also built on a word-level form. Thus, the MEC module again utilizes the proposed multi-kernel adaptive weighting block to fully extend the  $F$  and generate more discriminative features, which is defined as follows:

$$f'_i = f_{mean}(C1D_k(f_i)) \quad (11)$$

where  $f_{mean}$  represents the mean-pooling layer. After that, the learned features are integrated into the Fully Connected (FC) layer for predicting the confidence score  $p_c$  of final emotional state. In the training phase, the objective loss  $L$  is computed by using a SoftMax layer with cross-entropy for  $C$ -class classification, which is defined as follows:

$$L = - \sum_{c=1}^C y_c \log(p_c) \quad (12)$$

where  $y_c$  indicates the presence with respect to the ground-truth label.

### IV. EXPERIMENTS

In this section, we evaluate the performances of the proposed TSIN method on the IEMOCAP benchmark dataset by comparing it with state-of-the-art baselines. Next, we make a detailed discussion for the ablation studies. Finally, visualization analyses are presented.

#### A. Dataset and Data Splitting

1) *Dataset*: As one of the most commonly used multimodal emotion datasets, the IEMOCAP dataset is comprised of five sessions performed by 10 unique speakers, where each session contains utterances from two speakers, i.e., one male and one female. Each dialog contains audio, transcriptions, video, and motion-capture recordings, but we only use audio and transcriptions in our work. Consistent with previous works, we merge the excitement dataset with the happiness dataset and conduct 4-class emotion classifications, i.e., anger (1,103), happiness (1,636), sadness (1,084), and neutrality (1,708).

2) *Data Splitting*: The studies [37], [38] have pointed to the sensibility and importance of splitting the dataset, stating that different setting standards might lead to extreme results. To make a fair comparison with previous methods, the proposed TSIN framework should be evaluated on different splitting standards. To the best of our knowledge, the existing commonly used splitting standards generally fall into three categories:

- *Session-Independent* (SesI): One session is divided into the testing set and others are considered as the training set;
- *Speaker-Independent* (SpkI): One speaker is divided into the testing set and others are considered as the training set;
- *Speaker/Session-Dependent* (SpkD): The dataset is randomly divided into ten parts and leaves one as the testing set.

#### B. Experimental Settings

1) *Baselines*: In our experiments, we compared the proposed TSIN method with some state-of-the-art baselines of the unimodal learning methods and multimodal learning methods. Specifically, a series of the unimodal learning methods are introduced to compare with the proposed TSIN framework, including WPA [17], I-CLA [39], IAAN [19], DRN [18], LDF [40], and TFCNN [41]. Moreover, the multimodal learning methods are also considered as the major contenders, which can be classified into three main types: (1) modality-independent fusion methods (i.e., LSTM-CNN [14], MDRE [25], MDNN [24], UWA [42], PAaAN [26], DNN-BN [27], XLNet [37], DCLS [38], EFCS [43]); (2) modality-dependent fusion methods (i.e., CX-LSTM [44], MHA-2 [11]); (3) word-level fusion methods (i.e., FAF [13], AL-LSTM [45], H-MM-4 [29]). Details of these baselines have been introduced in related works in Section II.

2) *Implementation Details*: We implement the proposed TSIN method with the deep learning toolbox Tensorflow on 1 TITAN XP GPU. During the training stage, we optimize the network using Adam optimizer with weight decay 0.001. Besides, our TSIN framework is trained for 80 epochs with a batch size of 64. The initial learning rate is 0.0001, which decays by a factor of 10 for every 50 epochs.

3) *Evaluation Metrics*: It is obvious that the division of test sets can greatly affect evaluation results. To make a fair comparison, we adopt all splitting standards to demonstrate the validity of the proposed TSIN method. To measure the effectiveness of the proposed method, both weighted average accuracy (WA) and unweighted average accuracy (UA) are considered as evaluation metrics. Specifically, WA is a weighted average accuracy over different emotion categories with weights proportional to the number of samples in each class, while UA represents the average accuracy over different emotion categories. They are defined as follows:

$$WA = \frac{\sum_{c=1}^C N_c * Accuracy_c}{\sum_{c=1}^C N_c}, \quad (13)$$

$$UA = \frac{1}{C} \sum_{c=1}^C Accuracy_c, \quad (14)$$

TABLE I  
EXPERIMENTAL RESULTS OF WA AND UA (%) ON IEMOCAP WITH DIFFERENT SPLITTING STANDARDS. ALL RESULTS WERE ANALYZED IN PERCENTAGE (%) TERMS. THE MEAN RESULTS ARE HIGHLIGHTED IN BOLD. “-” DENOTES THE CORRESPONDING RESULT IS NOT PROVIDED

Speech-Text	SesI		Speech-Text	SpkI		Speech-Text	SpkD	
Test set	WA	UA	Test set	WA	UA	Test set	WA	UA
Seesion1	72.1	76.1	Ses01M	75.9	80.2	Group1	78.7	80.2
Seesion2	79.2	82.0	Ses01F	72.2	76.0	Group2	78.3	80.5
Seesion3	72.3	72.4	Ses02M	78.8	84.0	Group3	77.6	78.6
Seesion4	76.9	78.2	Ses02F	82.7	83.0	Group4	77.2	77.5
Seesion5	73.8	74.5	Ses03M	74.9	75.5	Group5	80.1	80.9
-	-	-	Ses03F	72.8	73.3	Group5	80.7	81.7
-	-	-	Ses04M	75.4	76.7	Group6	77.4	78.7
-	-	-	Ses04F	79.9	81.0	Group7	79.9	80.4
-	-	-	Ses05M	77.7	77.0	Group8	79.6	79.9
-	-	-	Ses05F	72.0	74.4	Group9	77.9	79.3
Mean	<b>74.92</b>	<b>76.64</b>	Mean	<b>76.23</b>	<b>78.11</b>	Mean	<b>78.74</b>	<b>79.77</b>

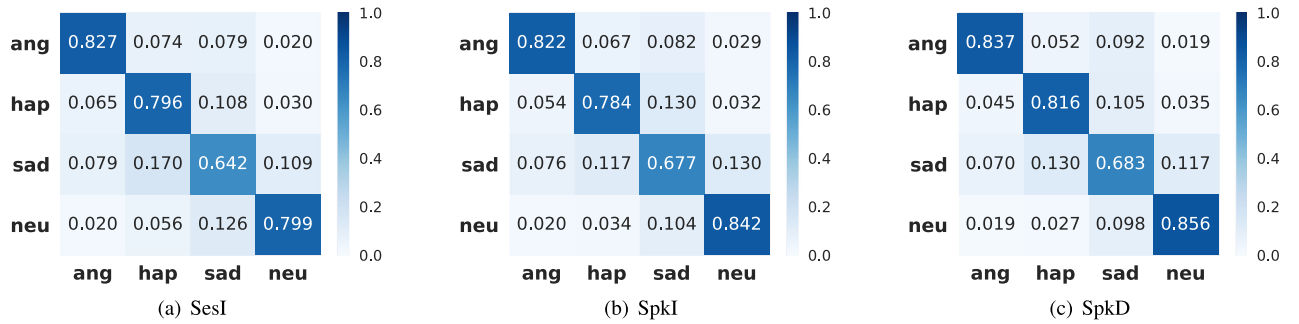


Fig. 3. Confusion matrices of recognition accuracy with the proposed TSIN method on SesI, SpkI, and SpkD, respectively.

where  $C$  denotes the number of emotion categories in the dataset,  $N_c$  denotes the number of samples in the  $c^{th}$  class, and  $Accuracy_c$  denotes the identification accuracy of samples in the  $c^{th}$  class. In our experiments, the UA score is a more reliable evaluation criterion due to the problem of class imbalance.

### C. Comparisons With State-of-the-Art Baselines

1) *Experimental Results on the SesI Splitting*: Based on the SesI data splitting standard, we perform experiments using 5-fold cross-validation to show the effectiveness of our TSIN method. The test scores of all sessions are shown in Table I and the corresponding recognition accuracy of each emotion category is shown in Fig. 3. Moreover, the comparison results with the state-of-the-art baselines are presented in Table II.

From Table II, we have the following observations. 1) The proposed TSIN can achieve the highest scores on both WA and UA metrics, which consistently outperforms other methods. This also verifies that our method is effective in parsing and refining discriminative acoustic and textual features. 2) By comparing with the unimodal models, such as WPA and IAAN, we can discover that the multimodal learning approaches are beneficial to achieve better performance. In particular, the UA score of our TSIN framework is clearly superior to the unimodal models, especially for WPA and IAAN with the improvements of 17.84%

TABLE II  
COMPARISONS OF WA AND UA WITH THE LATEST STATE-OF-THE-ART METHODS ON SESI. ALL RESULTS WERE ANALYZED IN PERCENTAGE (%) TERMS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. “-” DENOTES THE CORRESPONDING RESULT IS NOT PROVIDED

Method	WA	UA
WPA [19]	63.50	58.80
I-CLA [40]	68.10	63.80
EF-CS [44]	-	65.60
IAAN [21]	64.70	66.30
UWA [43]	-	68.18
PAaAN [16]	-	70.30
AL-LSTM [46]	72.50	70.90
FAF [14]	72.70	72.70
Our TSIN	<b>74.92</b>	<b>76.64</b>

and 10.34%. 3) Compared with the modality-independent models (e.g., PAaAN), AL-LSTM can further boost the performance of multimodal emotion recognition (70.90% vs. 70.30%), which proves the validity of exploring the interactions across modalities. 4) Based on the word-level emotional feature learning, FAF can achieve a better performance, improving the UA score from 70.90% to 72.70%. By contrast, our TSIN method is beneficial to



TABLE III

COMPARISONS OF WA AND UA WITH THE LATEST STATE-OF-THE-ART METHODS ON SPKI. ALL RESULTS WERE ANALYZED IN PERCENTAGE (%) TERMS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. “-” DENOTES THE CORRESPONDING RESULT IS NOT PROVIDED

Method	WA	UA
LSTM-CNN [15]	64.97	65.90
DRN [20]	-	67.40
H-MM-4 [30]	-	73.83
DNN-BN [28]	73.70	75.50
CX-LSTM [45]	-	75.60
MDNN [26]	75.20	76.70
Our TSIN	<b>76.23</b>	<b>78.11</b>

learn more fine-grained emotional information from the speech and text embeddings, which results in an improvement of 3.94% on UA metrics. These experimental results demonstrate that the proposed method is able to obtain better performance than those existed techniques.

2) *Experimental Results on the SpkI Splitting*: Moreover, we also perform experiments using 10-fold cross-validation based on the SpkI splitting standard. The test scores of all speakers are shown in Table I and the corresponding recognition accuracy of each emotion category is shown in Fig. 3. Moreover, the comparison results with the state-of-the-art baselines are presented in Table III.

From Table III, we have the following observations. 1) The proposed TSIN framework achieves the highest UA score of 78.11%, which demonstrates the stability and generalization of our proposed TSIN method. 2) Similarly, it is observed that the unimodal model, i.e., DRN, has a lower UA score of 67.40%, which is easily overpowered by multimodal learning approaches. 3) Given the problem of information redundancy, some traditional multimodal learning methods are difficult to make full use of acoustic and textual features, with poor forecast performance. For example, the UA score of LSTM-CNN is 65.9%, which is the worst-case performance in the group. 4) Although MDNN can achieve a reliable performance by using a massive scale of unlabeled data, our TSIN method still outperforms these competitive results, improving the UA score from the 76.70% to 78.11%, which proves the effectiveness of the proposed TSIN framework.

3) *Evaluation on the SpkD Splitting*: In this part, we perform the experiments using 10-fold cross-validation based on the SpkD splitting standard, which tend to achieve higher scores than previous experiments. The test scores of all speakers are shown in Table I and the corresponding recognition accuracy of each emotion category is shown in Fig. 3. Moreover, the comparison results with the state-of-the-art baselines are presented in Table IV.

From Table IV, we have the following observations. 1) Our TSIN framework contributes a new state-of-the-art: it yields the UA score of 79.77%, which is the best performance for multimodal emotion recognition. 2) Compared with unimodal models (e.g., TFCNN and LDF), the proposed TSIN method

TABLE IV

COMPARISONS OF WA AND UA WITH THE LATEST STATE-OF-THE-ART METHODS ON SPKD. ALL RESULTS WERE ANALYZED IN PERCENTAGE (%) TERMS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. “-” DENOTES THE CORRESPONDING RESULT IS NOT PROVIDED

Method	WA	UA
LDF [41]	65.40	66.90
TFCNN [42]	70.34	70.78
MDRE [27]	71.80	-
XLNet [38]	73.50	71.00
DCLS [39]	74.21	-
MHA-2 [12]	76.50	77.60
Our TSIN	<b>78.74</b>	<b>79.77</b>

dramatically improves the performance of emotion recognition, especially for the TFCNN and LDF with the improvement of 8.99% and 12.87%. 3) Although the existed modality-dependent methods are technologically behind our TSIN, their performance can easily outperform the modality-independent models. Notably, MHA-2 achieves the UA score of 77.60%, surpassing XLNet by 6.6%, which proves the effectiveness of multimodal interaction learning. 4) Compared with the previous state-of-the-art competitor, Our TSIN method can clearly enhance the performance of multimodal emotion recognition, which surpasses MHA-2 by over 2% on both WA and UA metrics. Therefore, the technical superiorities of the proposed TSIN method are well established.

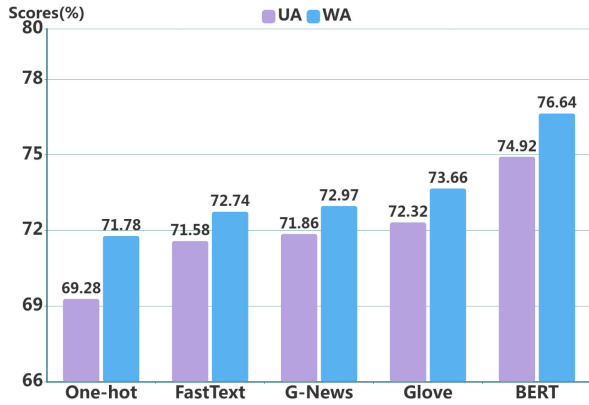
#### D. Ablation Studies

As mentioned above, the TSI module is used to between speech and text embedding spaces, which is indeed the key to the proposed TSIN method. In this part, we perform ablation studies from three different aspects, including the sensitivity of the TSIN framework to different types of word embeddings, effects of the unimodal branches, and effects of the designed modules.

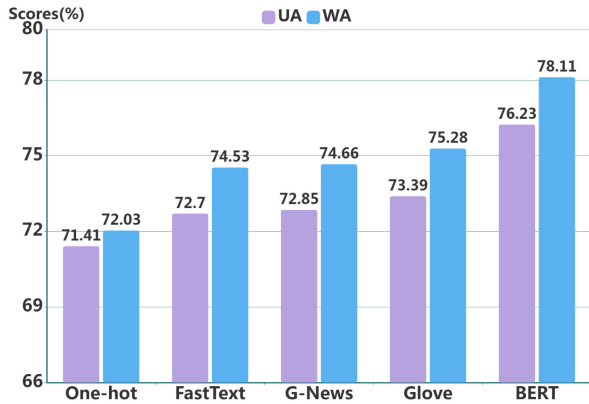
1) *Effects of Different Types of Word Embeddings*: In our experiments, we introduce the pre-trained Bert model as the way of word embedding, which serves as the inputs of the text embeddings for learning the textual feature. Here, we evaluate the performance of the proposed TSIN framework under other available word embedding functions. Specifically, we investigate five different word embedding methods, including Bert, Glove, Google-News [46], FastText [47], and the simple one-hot word embedding, as shown in Fig. 4. It can be seen that the word embedding can significantly affect the evaluation results. When we use more powerful word embedding, e.g., Bert, it can lead to better performance. That may because be that the powerful word embeddings learned from a large amount of text corpus maintain some semantic correlations and implicit dependencies, which can improve the accuracy of emotion recognition.

2) *Effects of Different Unimodal Branches*: In addition, we further compare our TSIN framework with two different types of unimodal settings, i.e., the “Speech” and “Text” models. Specifically, the “Speech” model is only built on the proposed

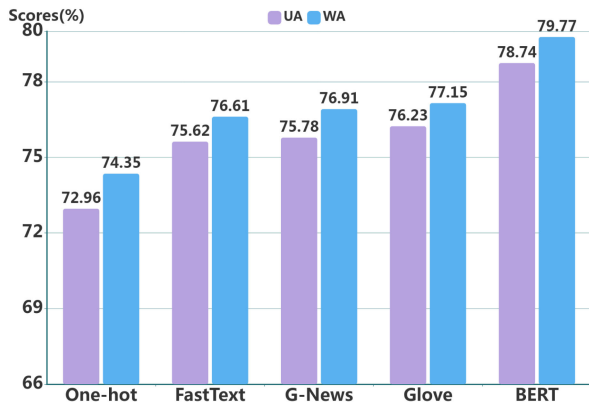




(a) Evaluation on SesI



(b) Evaluation on SpkI



(c) Evaluation on SpkD

Fig. 4. Comparisons of WA and UA for ablation studies with different types of word embeddings on the SesI, SpkI and SpkD splitting, respectively. All results were analyzed in percentage (%) terms.

speech encoder and the MEC module, while the “Text” model is initialized with the proposed text encoder and the MEC module. Table V shows the comparison results with different unimodal branches. As is shown above, the worst-performing model is “Speech,” where it achieves the UA scores of 63.32%, 64.32%, 68.39% on SesI, SpkI, and SpkD, respectively. By contrast, the “Text” model outperforms the “Speech” model, possibly because many people prefer to euphemistically express

TABLE V  
COMPARISONS OF WA AND UA FOR ABLATION STUDIES WITH DIFFERENT SETTINGS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. ALL RESULTS WERE ANALYZED IN PERCENTAGE (%) TERMS

Method	SesI		SpkI		SpkD	
	WA	UA	WA	UA	WA	UA
Speech	60.88	63.32	61.75	64.32	67.08	68.39
Text	69.52	70.48	70.01	71.77	71.50	72.28
TSI w/o TA	73.89	75.52	74.77	76.93	77.47	78.66
TSI w/o SI	74.25	75.74	74.92	77.01	77.68	78.75
Our TSIN	<b>74.92</b>	<b>76.64</b>	<b>76.23</b>	<b>78.11</b>	<b>78.74</b>	<b>79.77</b>

their emotions or rely on logic inferring. With the guidance of temporal and semantic consistency constraints provided by the TSI module, our TSIN framework can effectively explore the fine-grained interaction between speech and text data and further enhance the performance. Therefore, the observations justify that the accuracy improvements achieved by our TSIN method do not absolutely come from the semantic meanings derived from word embedding.

3) *Effects of the Designed Modules*: To evaluate the effectiveness of the designed TSI module, we conduct additional ablation experiments by removing each component in TSI module successively. The experimental results are presented in Table V. In detail, “TSI w/o TA” denotes that the temporal alignment mechanism is removed from the TSI module, while “TSI w/o SI” denotes that the semantic interaction mechanism is removed from the TSI module. “TSIN” is considered as the “full” model. When we remove the component of temporal alignment and forcefully align the speech and text embeddings, the UA scores of the remained model (i.e., “TSI w/o TA”) on SpkD is only 78.66%. Similarly, the UA scores of the “TSI w/o SI” on SpkD is only 78.75%, when it loses the functionality of semantic interaction. It is easy to see that “TSIN” outperforms the incomplete models. Thus, both the temporal alignment and semantic interaction mechanisms in the proposed TSIN module can complement and reinforce each other, which further enhancing the performance of multimodal emotion recognition.

#### E. Visualized Analysis

To further verify the superiority of the proposed TSIN, we perform the t-SNE visualizations [48] to visualize the emotional features extracted by the “Speech” model, the “Text” model, and the proposed TSIN framework. Figs. 5, 6 and 7 plot feature visualizations on four groups of emotion categories. We can clearly see that the distributions of the proposed TSIN framework are more discriminative than the unimodal models. In particular, the distributions of the same category with TSIN also exhibit the cluster pattern, while the distributions of different categories are located more dispersedly. Compared with unimodal learning, the proposed TSIN method can make full use of both acoustics and textual features by exploring the fine-grained temporal alignment and cross-modal semantic interaction. The above phenomenon indicates that our TSIN method with temporal

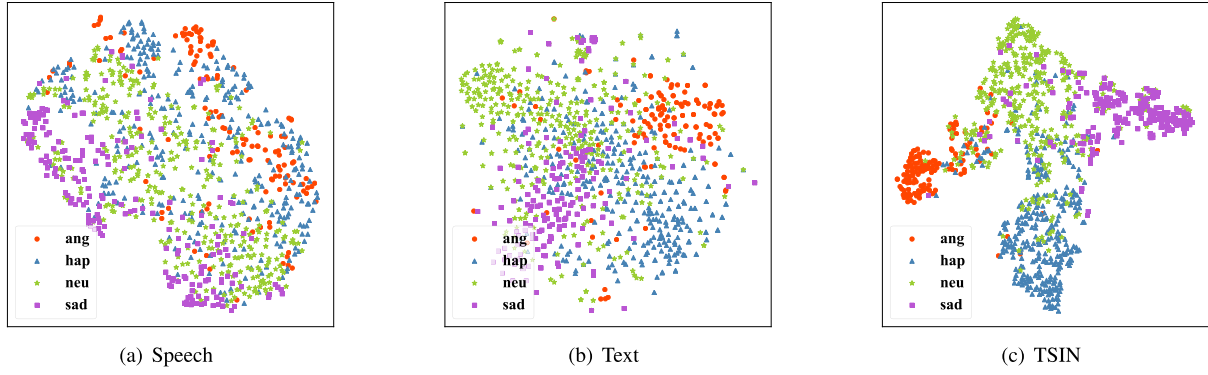


Fig. 5. Visualizations of the “Speech” model, the “Text” model and our TSIN framework by using t-SNE on SesI. Different emotion categories are shown in different colors and shapes, respectively.

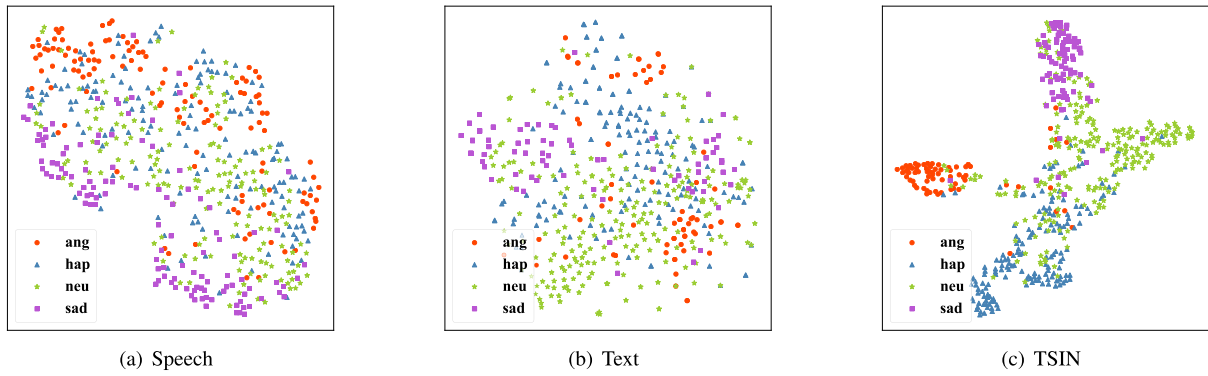


Fig. 6. Visualizations of the “Speech” model, the “Text” model and our TSIN framework by using t-SNE on SpkI. Different emotion categories are shown in different colors and shapes, respectively.

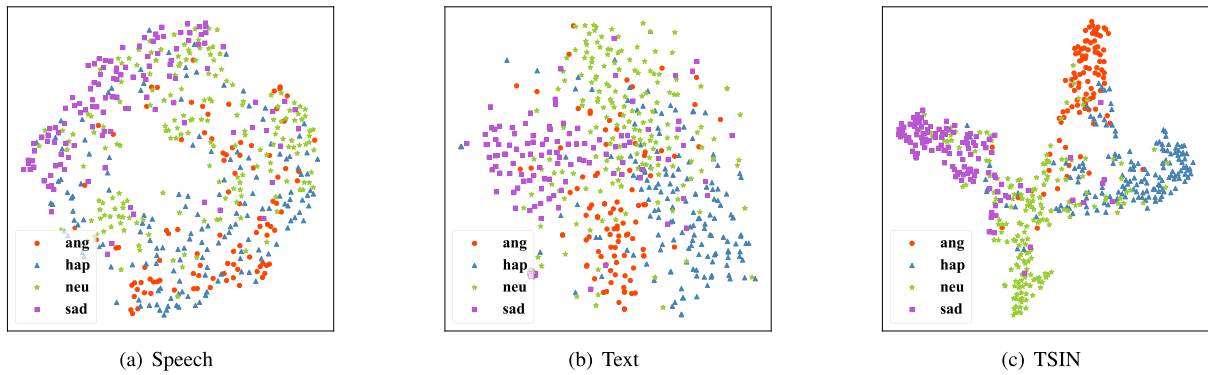


Fig. 7. Visualizations of the “Speech” model, the “Text” model and our TSIN framework by using t-SNE on SpkD. Different emotion categories are shown in different colors and shapes, respectively.

and semantic consistency constraints can effectively classify different emotional states.

## V. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel TSIN framework to address the task of multimodal emotion recognition. The main core of the proposed TSIN method is to concurrently explore the fine-grained temporal alignment and cross-modal semantic

interaction between the speech and text embedding spaces. Further improvements of speech-text emotion classification performance are achieved through the usage of both temporal and semantic consistency constraints for the implementation of emotional parsing and sentiment refining. Under the guidance of multimodal fine-grained interaction, both the speech-text embeddings can be interactively optimized and fine-tuned in an end-to-end manner. Extensive experimental results on the widely used benchmark dataset and the comprehensive analysis have

demonstrated the superiorities of the designed TSIN framework. Although making significant achievements, it is still challenging to make full use of emotional information across modalities. In our future work, we would attempt to combine more modalities data, such as image and video, to further boost the performance of automated emotion recognition.

## REFERENCES

- [1] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 32, no. 2, pp. 102–107, Mar./Apr. 2016.
- [2] B. Chen, Z. Zhang, Y. Lu, F. Chen, G. Lu, and D. Zhang, "Semantic interactive graph convolutional network for multi-label image recognition," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published, doi: [10.1109/TSMC.2021.3103842](https://doi.org/10.1109/TSMC.2021.3103842).
- [3] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, "Comparing and combining sentiment analysis methods," in *Proc. 1st ACM Conf. Online Soc. Netw.*, 2013, pp. 27–38.
- [4] I. Perikos and I. Hatzilygeroudis, "Recognizing emotions in text using ensemble of classifiers," *Eng. Appl. Artif. Intell.*, vol. 51, pp. 191–201, 2016.
- [5] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 12, pp. 2423–2435, Dec. 2018.
- [6] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech emotion classification using attention-based LSTM," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 11, pp. 1675–1685, Nov. 2019.
- [7] B. Chen, Z. Zhang, Y. Li, G. Lu, and D. Zhang, "Multi-label chest X-ray image classification via semantic similarity graph embedding," *IEEE Trans. Circuits Syst. Video Technol.*, to be published, doi: [10.1109/TCSVT.2021.3079900](https://doi.org/10.1109/TCSVT.2021.3079900).
- [8] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, no. 9, pp. 2697–2709, Sep. 2020.
- [9] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. W. Schuller, "Attention-enhanced connectionist temporal classification for discrete speech emotion recognition," in *Proc. Interspeech*, 2019, pp. 206–210.
- [10] Y. Qian, Z. Chen, and S. Wang, "Audio-visual deep neural network for robust person verification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, no. 2, pp. 1079–1092, Feb. 2021.
- [11] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 2822–2826.
- [12] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 3507–3511.
- [13] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proc. Conf. Assoc. Comput. Linguistics Meeting*, 2018, pp. 2225–2235.
- [14] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," in *Proc. Interspeech*, 2018, pp. 247–251.
- [15] M. Chen, S. Wang, P. P. Liang, K. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, 2017, pp. 163–171.
- [16] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.
- [17] S. Mirsamadi *et al.*, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 2227–2231.
- [18] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6675–6679.
- [19] S.-L. Yeh *et al.*, "An interaction-aware attention network for speech emotion recognition in spoken dialogs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6685–6689.
- [20] E. Cambria, Y. Li, F. Z. Xing, S. Poria, and K. Kwok, "SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, 2020, pp. 105–114.
- [21] M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, and U. R. Acharya, "ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis," *Future Gener. Comput. Syst.*, vol. 115, pp. 279–294, 2021.
- [22] M. S. Akhtar, A. Ekbal, and E. Cambria, "How intense are you? predicting intensities of emotions and sentiments using stacked ensemble," *IEEE Comput. Intell. Mag.*, vol. 15, no. 1, pp. 64–75, Feb. 2020.
- [23] C. Song, X.-K. Wang, P.-f. Cheng, J.-Q. Wang, and L. Li, "SACPC: A framework based on probabilistic linguistic terms for short text sentiment analysis," *Knowl.-Based Syst.*, vol. 194, 2020, Art. no. 05572.
- [24] S. Zhou, J. Jia, Q. Wang, Y. Dong, Y. Yin, and K. Lei, "Inferring emotion from conversational voice data: A semi-supervised multi-path generative neural network approach," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 579–587.
- [25] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 112–118.
- [26] J.-L. Li and C.-C. Lee, "Attentive to individual: A multimodal emotion recognition network with personalized attention profile," in *Proc. Interspeech*, 2019, pp. 211–215.
- [27] E. Kim and J. W. Shin, "Dnn-based emotion recognition based on bottleneck acoustic features and lexical features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6720–6724.
- [28] W. Y. Choi, K. Y. Song, and C. W. Lee, "Convolutional attention networks for multimodal emotion recognition from speech and text data," in *Proc. Grand Challenge Workshop Hum. Multimodal Lang.*, 2018, pp. 28–34.
- [29] G. Aguilar, V. Rozgic, W. Wang, and C. Wang, "Multimodal and multi-view models for emotion recognition," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 991–1002.
- [30] S. Poria, H. Peng, A. Hussain, N. Howard, and E. Cambria, "Ensemble application of convolutional neural networks and multiple Kernel learning for multimodal sentiment analysis," *Neurocomputing*, vol. 261, pp. 217–230, 2017.
- [31] I. Chaturvedi, R. Satapathy, S. Cavallari, and E. Cambria, "Fuzzy commonsense reasoning for multimodal sentiment analysis," *Pattern Recognit. Lett.*, vol. 125, pp. 264–270, 2019.
- [32] K. Dashtipour, M. Gogate, E. Cambria, and A. Hussain, "A novel context-aware multimodal framework for persian sentiment analysis," *Neurocomputing*, vol. 457, pp. 377–388, 2021.
- [33] L. Stappen, A. Baird, E. Cambria, and B. W. Schuller, "Sentiment analysis and topic recognition in video transcriptions," *IEEE Intell. Syst.*, vol. 36, no. 2, pp. 88–95, Mar./Apr. 2021.
- [34] T. N. Sainath, B. Kingsbury, A.-R. Mohamed, and B. Ramabhadran, "Learning filter banks within a deep neural network framework," in *Proc. IEEE Workshop Autom. Speech, Recognit., Understanding*, 2013, pp. 297–302.
- [35] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.
- [36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
- [37] V. Heusser, N. Freymuth, S. Constantin, and A. Waibel, "Bimodal speech emotion recognition using pre-trained language models," 2019, *arXiv:1912.02610*.
- [38] G. Xu, W. Li, and J. Liu, "A social emotion classification approach using multi-model fusion," *Future Gener. Comput. Syst.*, vol. 102, pp. 347–356, 2020.
- [39] L. Tarantino *et al.*, "Self-attention for speech emotion recognition," in *Proc. Interspeech*, 2019, pp. 2578–2582.
- [40] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. Meng, "Learning discriminative features from spectrograms using center loss for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 7405–7409.
- [41] J. Liu, Z. Liu, L. Wang, L. Guo, and J. Dang, "Speech emotion recognition with local-global aware deep representation learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7174–7178.
- [42] S. Sahu, V. Mitra, N. Seneviratne, and C. Y. Espy-Wilson, "Multi-modal learning for speech emotion recognition: An analysis and comparison of ASR outputs with ground truth transcription," in *Proc. Interspeech*, 2019, pp. 3302–3306.
- [43] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, "Fusion approaches for emotion recognition from speech using acoustic and text-based features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6484–6488.

- [44] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 873–883.
- [45] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," *Proc. Interspeech*, 2019, pp. 3569–3573.
- [46] A. S. Das, M. Datar, A. Garg, and S. Rajaram, "Google news personalization: Scalable online collaborative filtering," in *Proc. 16th Int. Conf. World Wide Web*, 2007, pp. 271–280.
- [47] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext: Zip: Compressing text classification models," 2016, *arXiv:1612.03651*.
- [48] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.



**Bingzhi Chen** received the B.S. degree in software engineering from South China Normal University, Guangzhou, China, in 2017. He is currently working toward the Ph.D. degree with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China. His current research interests include computerized medical diagnosis, multimodal learning, pattern recognition, deep learning, and machine learning.



**Qi Cao** received the B.S. degree from the Hebei University of Science and Technology, Shijiazhuang, China, in 2017. He is currently working toward the master's degree with the Harbin Institute of Technology (Shenzhen), Shenzhen, China. His current research interests include deep learning, speech recognition, and relevant applications.



**Mixiao Hou** received the B.S. and M.S. degrees in computer science and technology from Qufu Normal University, Jining, China, in 2016 and 2019, respectively. She is currently working toward the Ph.D. degree with the Harbin Institute of Technology (Shenzhen), Shenzhen, China. Her current research interests include speech emotion recognition, pattern recognition, and deep learning.



**Zheng Zhang** (Senior Member, IEEE) received the M.S. and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 2014 and 2018, respectively. He was a Postdoctoral Research Fellow with The University of Queensland, Brisbane, QLD, Australia. He is currently with the Harbin Institute of Technology (Shenzhen), Shenzhen, China. He has authored or coauthored more than 100 technical papers at prestigious journals and conferences. His research interests include machine learning, computer vision, and multimedia analytics. He is an Editorial Board Member of the *Information Processing & Management Journal*, and also serves/served as the AC/SPC/PC member of several top conferences.



**Guangming Lu** (Member, IEEE) received the B.S. degree in electrical engineering, the M.S. degree in control theory and control engineering, and the Ph.D. degree in computer science and engineering from the Harbin Institute of Technology, Harbin, China, in 1998, 2000, and 2005, respectively. From 2005 to 2007, he was a Postdoctoral Fellow with Tsinghua University, Beijing, China. He is currently a Professor with the Bio-Computing Research Center, Harbin Institute of Technology (Shenzhen), Shenzhen, China. He has authored or coauthored more than 120 technical papers at prestigious international journals and conferences, including TIP, TNNLS, TCYB, TCSVT, CVPR, NeurIPS, AAAI, ACMM, and IJCAI. His current research interests include pattern recognition, image processing, and automated biometric technologies and applications.



**David Zhang** (Life Fellow, IEEE) graduated in computer science from Peking University, Beijing, China. He received the M.Sc. and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1982 and 1985, respectively, and the second Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 1994. From 1986 to 1988, he was a Postdoctoral Fellow with Tsinghua University, Beijing, China, and then an Associate Professor with Academia Sinica, Beijing, China. Since 2005, he has been the Chair Professor with Hong Kong Polytechnic University, Hong Kong, where he is the Founding Director of Biometrics Research Centre (UGC/CRC) supported by Hong Kong SAR Government. Prof. Zhang is a Croucher Senior Research Fellow, Distinguished Speaker of the IEEE Computer Society, and a Fellow of the Royal Society of Canada and Canadian Academy of Engineering, and also IAPR Fellow.