applied sciences

MDPI

Article

Improvement of Multimodal Emotion Recognition Based on Temporal-Aware Bi-Direction Multi-Scale Network and Multi-Head Attention Mechanisms

Yuezhou Wu *, Siling Zhang * and Pengfei Li

School of Computer Science, Civil Aviation Flight University of China, Guanghan 618307, China; cafuclpf@163.com

* Correspondence: wuyuezhou@cafuc.edu.cn (Y.W.); cafuczsl@163.com (S.Z.)

Abstract: Emotion recognition is a crucial research area in natural language processing (NLP), aiming to identify emotional states such as happiness, anger, and sadness from various sources like speech, text, and facial expressions. In this paper, we propose an improved MMER (multimodal emotion recognition) method using TIM-Net (Temporal-Aware Bi-Direction Multi-Scale Network) and attention mechanisms. Firstly, we introduce the methods for extracting and fusing the multimodal features. Then, we present the TIM-Net and attention mechanisms, which are utilized to enhance the MMER algorithm. We evaluate our approach on the IEMOCAP and MELD datasets, and compared to existing methods, our approach demonstrates superior performance. The weighted accuracy recall (WAR) on the IEMOCAP dataset is 83.9%, and the weighted accuracy recall rate on the MELD dataset is 62.7%. Finally, the impact of the TIM-Net model and the attention mechanism on the emotion recognition performance is further investigated through ablation experiments.

Keywords: emotion recognition; deep learning; multimodal; TIM-Net; attention mechanism

Citation: Wu, Y.; Zhang, S.; Li, P. Improvement of Multimodal Emotion Recognition Based on Temporal-Aware Bi-Direction Multi-Scale Network and Multi-Head Attention Mechanisms. *Appl. Sci.* **2024**, *14*, 3276. https://doi.org/10.3390/app14083276

Academic Editor: Douglas O'Shaughnessy

Received: 13 March 2024
Revised: 8 April 2024
Accepted: 10 April 2024
Published: 13 April 2024

1. Introduction

Emotion is an important way for humans to express their inner world, and its complexity and diversity play a crucial role in human communication and interaction [1]. The American psychologist Ekman [2] proposed six basic emotions based on research needs. In 1977, the concept of affective computing was introduced [3], and emotion recognition is an important direction in affective computing [4]. In recent years, significant progress has been made in unimodal approaches using text, audio, and video for emotion recognition. Emotion recognition has found applications in various fields, such as traffic safety [5], intelligent interaction [6,7], and healthcare [8–10].

Emotion recognition in speech data is known as speech emotion recognition (SER). SER establishes a mapping relationship between speech feature information and emotional states through different models by extracting the feature information from the speech data [11]. The selection and design of speech emotional features are crucial steps in SER, which directly affect its performance. In 2021, Wang et al. [12] achieved excellent results by using a shared-weight multimodal Transformer [13] to capture the dependencies between modalities. Finally, the success of pre-trained models in speech recognition tasks has garnered increasing attention in speech emotion recognition research. In 2022, Zou et al. [14] utilized the Wav2vec [15] pre-trained model to extract the deep features from speech and combined them with traditional acoustic features for emotion recognition. The ablation experiment proved that adding deep features achieved better performance.

Research on the video modality primarily focuses on facial emotion recognition (FER). AffectNet [16] is a widely recognized dataset for video-based emotion recognition. Bakariya et al. [17] developed a real-time system capable of detecting faces, assessing human emotions, and recommending music to users. Meena et al. [18] proposed a CNN-based facial

image emotion analysis model. The study of emotion in textual data is called emotion analysis (EA). With the rapid increase in text content created by users on the Internet, such as product reviews, social media posts, and blogs, we have access to abundant public opinion information. EA has also been proven to be beneficial in critical events such as the COVID-19 pandemic [19,20].

It is difficult to obtain accurate emotional information through a single modality alone [21,22]. With the continuous development and application of artificial intelligence technology, against a background of multimodal information fusion, emotion recognition technology can comprehensively analyze the information from different modalities, such as text, speech, and facial expressions, to obtain more complementary information [23], thereby improving the accuracy of emotion recognition [24–26]. Multimodal sentiment analysis has a wide range of applications in human–computer interaction, business, and education and holds great social value [27]. The core challenge of multimodal sentiment analysis is representation fusion, which aims to learn representations reflecting the cross-modal interactions between the "individual elements" of different modalities, effectively reducing the number of individual representations [28]. Recent works [24,29,30] have achieved success in the problem of multimodal sentiment analysis. The essence of multimodal sequences is time series, and there are local interactions between modalities at the same time node [31]. For example, the meaning expressed by a word at a particular moment is related to the pronunciation of the word and also the accompanying facial expression. Combining multimodal emotional features with artificial intelligence is an important research direction in the field of emotion recognition. The cartoon animal characters developed by Bates [32] can express set emotions and provide us with a new emotional communication experience. The virtual tiger (Tigrito) developed by Hayes-Roth et al. [33], which can predict and generate emotions, further demonstrates the broad application prospects of multimodal sentiment analysis.

In the process of delving deeper into the field of multimodal emotion recognition, we have discovered that integrating information from different modalities often leads to more precise and comprehensive analysis results. Especially in the current flourishing landscape of deep learning technologies, some advanced model architectures provide powerful tools for processing both speech and text data. Based on this, we propose a multimodal MMER-TAB model focusing on two modalities: speech and text. Building upon the previous research, we have made improvements to the existing MMER model. While MMER has achieved significant results in some respects, it also has some limitations that motivate our work. Firstly, the MMER model faces challenges in handling long-term dependencies and global context. To address this issue, we introduce a new structure consisting of three layers of TAB modules combined with attention mechanisms to enhance the modeling capability of the model. These improvements enable our model to better capture the long-term dependencies and global contextual information in speech signals, thereby improving its performance and robustness. The main contributions of our study are as follows:

- We particularly focus on two Transformer-based feature extraction methods, namely Wav2vec and BERT (Bidirectional Encoder Representation from Transformers) [34]. Additionally, we investigate the TIM-Net [35] model and make corresponding improvements to better adapt it to our application scenario in multi-modal emotion recognition tasks.
- We introduce a multi-head attention mechanism, which accurately captures significant emotional features from both speech and text, thereby enhancing the model's sensitivity and capability to capture emotional information. This innovation gives our model a competitive edge in emotion recognition.
- To validate the effectiveness of our proposed model, we conduct extensive experiments on the IEMOCAP [36] and MELD [37] datasets. The experimental results demonstrate that our MMER-TAB (Multi-Modal Emotion Recognition–Temporal-Aware Block) model exhibits outstanding performance in multi-modal conversation emotion analy-

sis tasks. This not only confirms the effectiveness of the feature extraction methods and model improvements we have adopted but also provides new insights into and methods for research in the field of multi-modal emotion recognition.

The rest of the paper is structured as follows: Section 2 introduces the feature extraction methods for the speech and text modalities, respectively. Section 3 presents the TIM-Net network architecture and attention mechanism. Section 4 elaborates on the proposed model's structure. Section 5 discusses and analyzes the experimental results. Finally, Section 6 provides a summary of the paper.

2. Methodology

2.1. Wav2vec Speech Features

Early approaches to speech features involved manually crafted Low-Level Descriptor (LLD) features, such as prosodic features, acoustic features, and spectral features. Mel-Frequency Cepstral Coefficients (MFCCs) are classic audio features known for their simplicity and efficiency. However, MFCCs only consider frequency information and overlook the temporal correlation in audio data. Additionally, MFCCs often require manual parameter settings, such as for the window size and stride. With the advancement of deep learning, researchers have turned to deep learning methods for extracting speech emotion features. Common deep learning methods for speech feature extraction include Convolutional Neural Networks (CNNs) [38,39], Recurrent Neural Networks (RNNs) [40], Bidirectional Long Short-Term Memory (BiLSTM) [41–43], etc. The Wav2vec2.0 version used in this paper is an end-to-end training approach that can learn representative feature descriptions directly from audio data through self-supervised learning, eliminating the need for manual parameter tuning.

For speech emotion recognition tasks, the Wav2vec2.0 method effectively captures the emotional information in audio. Furthermore, Wav2vec2.0 can enhance the model performance through pre-training, such as Wav2vec2-base-960h, which is pre-trained on diverse audio data for 960 h. This pre-training allows the model to capture more audio patterns and structures, resulting in improved robustness in speech emotion recognition tasks. In practical applications, fine-tuning can be performed based on specific tasks and data to further enhance the model performance. It is important to note that Wav2vec2-base-960h is a relatively complex model, requiring significant computational resources and longer training times.

The Wav2vec2.0 process mainly involves taking an input speech signal X and encoding it using a seven-layer CNN network to obtain the latent variable Z. The latent variable Z is quantized into the quantized variable Q through the Gumbel softmax quantization module. Simultaneously, Z is randomly masked at some positions and put into the Transformer [28] network to obtain the contextual feature vector C. The structure and overall process of Wav2vec2.0 are illustrated in Figure 1.

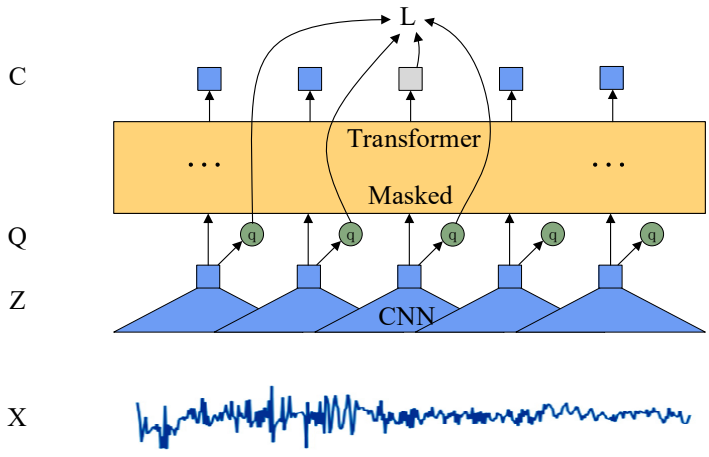


Figure 1. Structure diagram of the Wav2vec2.0 model.

2.2. BERT Text Features

The text information corresponding to speech is the most fundamental and intuitive carrier of emotion, often used to infer the emotional state of the speaker [44]. The process of emotion state recognition based on text is illustrated in Figure 2, primarily encompassing data preprocessing, text feature extraction, model training, and emotion recognition.

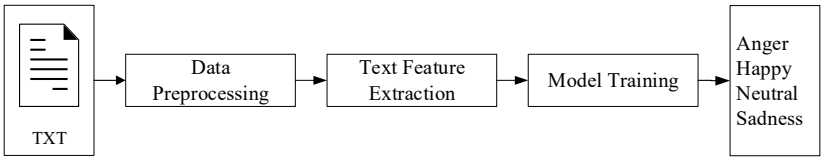


Figure 2. Text emotion recognition process.

Text emotion recognition first requires collecting text data from various sources, such as social media, comments, news articles, etc. The collected text data are then subjected to preprocessing operations, including cleaning, tokenization, the removal of stop words, stemming, and punctuation handling. Text feature extraction involves converting the text into numerical features suitable for machine learning algorithms. Common feature extraction methods include the bag-of-words model, TF-IDF (Term Frequency–Inverse Document Frequency) [45] features, word embedding [46] of syntactic features, etc.

Word embedding is currently the most commonly used method in text feature extraction, aiming to map words from the text data to a continuous vector space. This mapping process is achieved by capturing the contextual relationships between words. Common word embedding techniques include Word2Vec [47], Glove (Global Vectors for Word Representation) [48], Elmo (Embeddings from Language Models) [49], etc. However, it is important to note that using a single vector to represent a word in different contexts may lead to some semantic understanding errors.

Considering these factors, this paper adopts the language representation model BERT for feature extraction. BERT is a pre-trained language model based on the Transformer architecture, utilizing a bidirectional training approach during pre-training, allowing it to consider both the left and right context information simultaneously. BERT's strength lies in its powerful context modeling and multi-task pre-training, enabling the model to learn richer, more universal semantic representations, resulting in outstanding performance across various natural language processing tasks. BERT, proposed by Google in 2018 as an alternative to Word2Vec, is essentially composed of stacked Transformer encoders. It follows a two-phase framework comprising pre-training and fine-tuning on specific tasks. BERT's innovation lies in using two pre-training tasks: a Masked Language Model (MLM), which predicts masked words in a sequence, and Next Sentence Prediction (NSP), which predicts whether the next sentence is related to the current one. This addresses the issue of different semantic expressions for the same word in different contexts. The structure of the BERT model is illustrated in Figure 3.

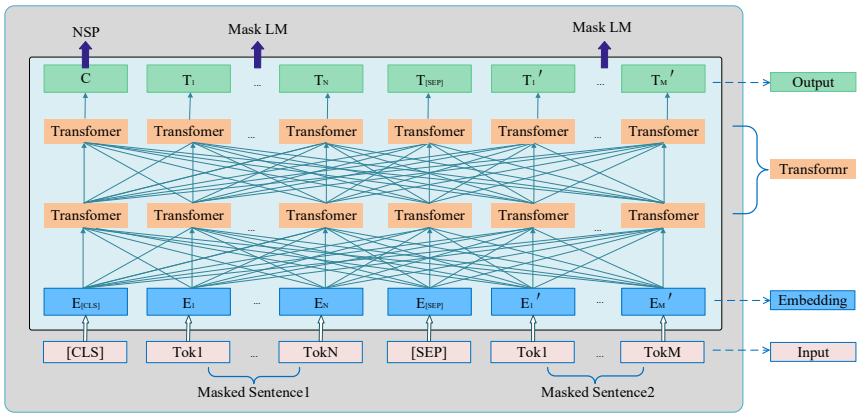


Figure 3. Structure diagram of the BERT model.

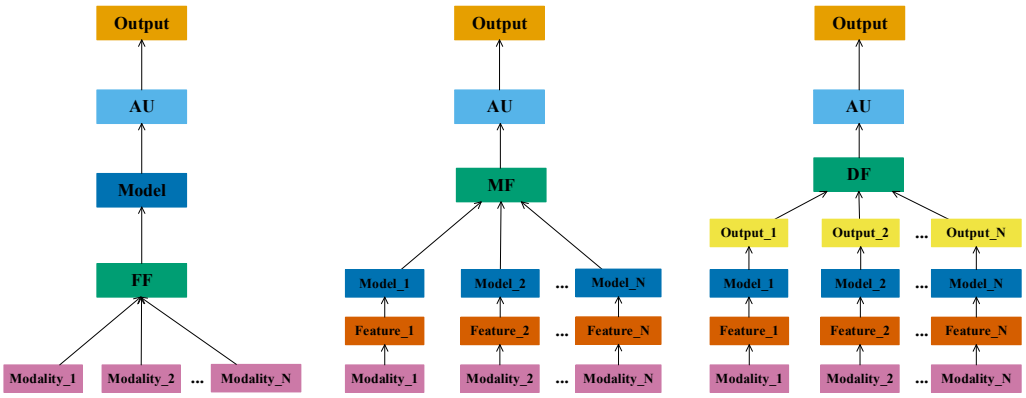
2.3. Multimodal Feature Fusion

Multimodal feature fusion [50] enables the acquisition and interpretation of information from different dimensions, providing a more comprehensive and accurate understanding. The two most commonly used modalities in speech emotion recognition are audio and text data. Audio data contain information such as speech rate, intonation, and volume, which can represent the speaker's emotions but may struggle to convey contextual semantic information. Text data can capture rich semantic information but may suffer from ambiguity and are significantly influenced by the text recognition accuracy. This paper proposes an algorithm that integrates the strengths and mitigates the weaknesses of both audio and text data, achieving complementary multimodal feature information.

The focus of multimodal feature fusion lies in the fusion stage and the fusion method, which will be discussed separately below.

2.3.1. Classification Based on the Fusion Stage

The fusion stage can be categorized into three types, as illustrated in Figure 4: feature-level fusion, model-level fusion, and decision-level fusion.



(a) Feature-level fusion (b) Model-level fusion (c) Decision-level fusion

Figure 4. Multimodal fusion method. AU is an analysis unit.

Feature-level fusion, or early fusion, involves extracting different modality features and concatenating them to form an overall multimodal feature representation. Since various modality information often exhibits high correlation, extracting this correlation after feature-level fusion can be challenging. Therefore, this method may not fully capture the correlation between different modalities, and in the temporal dimension, simple feature fusion may not achieve cross-temporal fusion of multimodal data. As the number of modalities increases, concatenating feature vectors may lead to high-dimensional features, difficulty in training models, and information redundancy.

Model-level fusion involves merging two features at an intermediate stage in the model. Afterward, independent models continue to extract the features, and finally, both types of features are combined for the classification task before the final classification. Taking the example of Multi-Layer LSTM (ML-LSTM), this approach combines multiple layers of neural networks with a traditional LSTM (Long Short-Term Memory) model. The fusion process is as follows: The text features are put into the first LSTM layer (Layer 1), producing hidden layer states for each neuron. Subsequently, the audio features are concatenated with the hidden layer states from Layer 1 and put into the second LSTM layer (Layer 2), generating hidden layer states for each neuron in the second layer. The visual features are then concatenated with the hidden layer states from Layer 2 and put into the third LSTM layer (Layer 3), producing hidden layer states for each neuron in the third layer. Finally, the fused features are input into the fully connected layer (FC) to obtain the prediction result.

Decision-level fusion, also known as late fusion, primarily involves using different network architectures for feature extraction and textual and audio information fusion. Decision-level fusion models each modality separately, treating different modalities as mutually independent. The features are extracted for each modality, and the emotion recognition results for individual modalities are obtained through emotion classifiers. Subsequently, a decision method is applied to recognize the results of each modality, ultimately yielding the final emotion classification result. Designing the decision rules for decision fusion is a challenging task. If the decision rules are too simple, they may not accurately reflect the correlation between different modalities.

2.3.2. Classification Using the Fusion Method

The simplest way to perform multimodal feature fusion is concatenation, such as concatenating using CONCAT or stacking operations. Another approach is to employ attention mechanisms. If a single layer of attention is insufficient, multiple attention operations can be applied. For example, attention operations can be performed from text to audio and vice versa. For instance, the query matrix W_S^Q from speech is computed using the key matrix and value matrix from the text W_T^K, W_T^V, while the key matrix and value matrix from the speech W_S^K, W_S^V are used to compute the query matrix from the text W_T^Q. This type of attention mechanism is also known as a cross-modal attention mechanism.

Based on the above analysis, the chosen fusion stage in this paper is model-level fusion, and the selected fusion method uses cross-modal attention mechanisms. A multimodal emotion recognition framework has been developed to fuse the features from both speech and text, and this fusion framework will be introduced in Section 4.

3. TIM-Net and Attention Mechanisms

3.1. The TIM-Net Emotion Recognition Network Model

TIM-Net is capable of learning contextual representations from different temporal scales. The network structure is illustrated in Figure 5 [35]. Specifically, TIM-Net utilizes a temporal-aware block to learn the temporal emotion representations initially. It then integrates supplementary information from both the past and the future to enrich the contextual representations. Finally, it fuses features from multiple temporal scales with the aim of better adapting to emotional changes. TIM-Net outperforms the other methods in terms of its accuracy on each corpus.

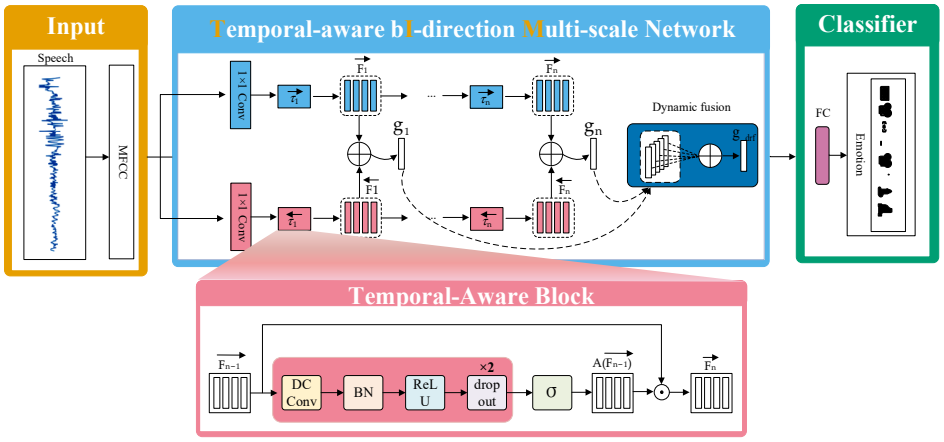


Figure 5. TIM-Net network architecture.

The superior generality and performance exhibited by TIM-Net can be attributed to its core module, called the temporal-aware block (TAB). This core module captures temporal-aware representations. Each TAB consists of three sub-blocks and a sigmoid function for learning the temporal attention map A. The temporal-aware feature F is generated through element-wise multiplication of the input and A. For the same sub-block of the *j-th* (TAB_j),

an expandable dilated causal convolution (DC Conv) [51] with a dilation rate of 2^{j-1} is applied at the beginning of each sub-block. The expandable convolution enlarges and refines the receptive field, while the causal constraint ensures that future information does not leak into the past. Batch normalization, activation functions, and spatial dropout follow the convolution operation. This paper made modifications to the TAB structure, changing it from a 2-layer structure to a 3-layer one. To reduce the complexity of the model, we replaced scalable dilated causal convolution with regular convolution and replaced the spatial pool with a regular pool. The modified structure is illustrated in Figure 6.

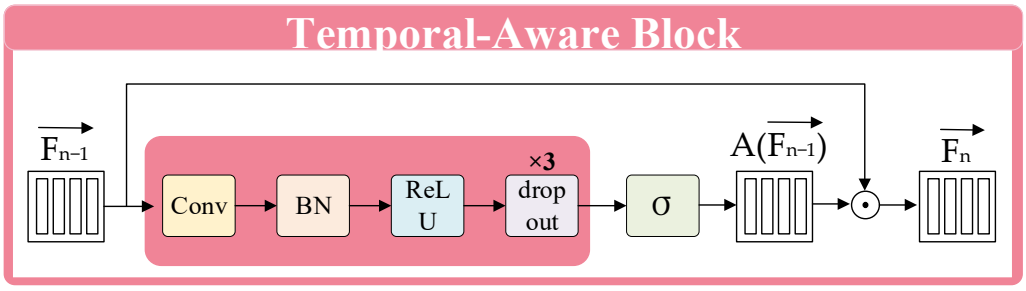


Figure 6. Schematic diagram of the modified TAB structure.

3.2. Attention Mechanism

Attention mechanisms enable neural networks to automatically learn and selectively focus on important information in the input. Multi-head attention is one implementation of attention mechanisms, achieved by running the attention mechanism in parallel multiple times and concatenating the independently computed attention outputs, linearly transforming them into the desired dimensions. Specifically, the multi-head attention mechanism projects the input matrix differently, generating several output matrices that are then concatenated together. Under the multi-head attention mechanism, the input sequence data are divided into multiple heads, each independently computing and producing different outputs. These outputs are then concatenated to form the final output. The output for each head can be expressed as follows, where W_i^Q, W_i^K, W_i^V are the query, key, and value transformation matrices for the i-th head. In summary, the multi-head attention mechanism is an effective implementation of attention that can significantly enhance the model's performance and generalization ability.

We use both the multi-head attention mechanism and the cross-modal attention mechanism. These two attention mechanisms have different positions and functions in the model.

4. Model Design and Interpretation

Considering the two modal features of text and audio in speech emotion recognition, based on the TIM-Net network structure, the TAB design is improved and combined with the use of multi-head attention, and the network framework is constructed, as shown in Figure 7. The features extracted after passing through Wav2Vec2.0 and RoBERT$_a$ have a dimensionality of 768. The multi-head attention mechanism employs 8 heads, with 2 layers in the encoder. By increasing the number of heads, the model can capture more contextual information. The TAB consists of 3 layers, with ReLU used as the activation function within the TAB. The input dimensionality received by the fully connected layer is 768×2 (concatenation of audio and text), with the gelu activation function and AdamW optimizer used and a learning rate of 5×10^{-5}. To prevent overfitting, the dropout is set to 0.1.

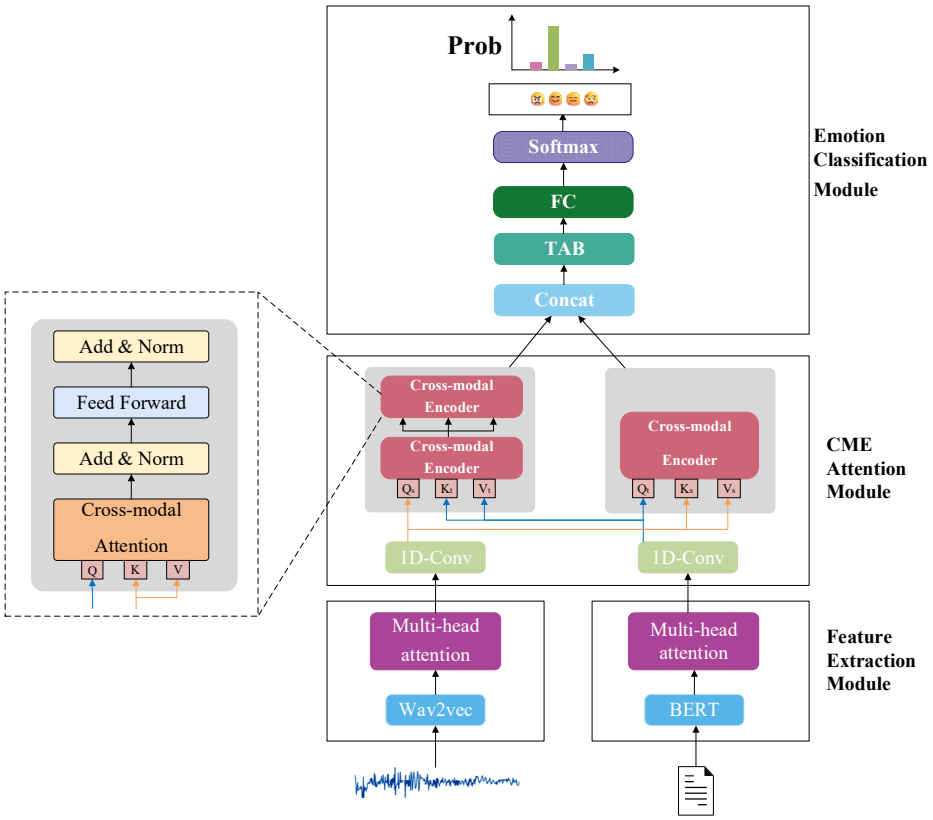


Figure 7. Multi-modal fusion framework.

The first part is the feature extraction module. This section primarily focuses on extracting features from the input data, which include two modalities: audio and text. The process for extracting the features from audio modality data is as follows: first, encode the audio using Wav2vec-2.0, and then introduce a multi-head self-attention mechanism to learn more discriminative speech emotion features. The feature extraction process for text modality data involves using the BERT model, followed by introducing a multi-head self-attention mechanism to focus on significant emotional features within the text sequence.

The second part is the Cross-Modal Encoder (CME) attention module. This section primarily models the cross-modal interactions of the multi-modal features, utilizing a cross-modal attention mechanism to jointly optimize the feature embeddings for audio and text. The cross-modal attention mechanism achieves this by learning two sets of semantic interaction weights separately and readjusting the feature representations of audio and text. This enables capturing interactive information between the audio and text modalities, achieving semantic consistency in the multi-modal context.

The third part is the emotion classification module. In this section, the multi-modal fusion features of audio and text are first concatenated. Then, the TAB module is employed to learn the temporal dimension features with context dependencies. The features learned with context dependencies are then put into an FC layer, utilizing a softmax classifier to obtain a probability matrix. The maximum value in the matrix is taken as the final emotion recognition result.

The following sections will provide separate introductions to the cross-modal attention module and the emotion classification module of this model.

4.1. Cross-Modal Attention Module

This paper employs cross-modal attention to focus on the interaction between different modal data, learning the semantic interaction weights for the speech and text modalities and readjusting the feature representations.

Firstly, both speech and text representations are projected into the same space using 1D-CNN, and the representations are as follows:

$$\overline{H}\{S,T\} = Conv1D\left(H_{\{S,T\}}, k_{\{S,T\}}\right) \in R^d \tag{1}$$

where S represents the speech modality, T represents the text modality, $H_{\{S,T\}}$ represents the final emotional feature representations for speech and text obtained from the feature extraction module, $k_{\{S,T\}}$ represents the convolution kernel size for modality $\{S,T\}$, and d denotes the dimension of the projected features for speech and text. The speech embeddings mapped using 1D-CNN are denoted as $\overline{H}_S$ and $\overline{H}_T$.

We denote the process of transferring information from the speech modality to the text modality as $S \to T$, and correspondingly, $T \to S$ is used to represent the information transfer from the text modality to the speech modality. To learn the relationship between speech and text, linear projection is initially employed to transform each feature sequence into a query matrix Q, a key matrix K, and a value matrix V. The calculation formula is as follows:

$$\begin{aligned} Q_l &= W_l^Q \overline{H}_l \\ K_l &= W_l^K \overline{H}_l \\ V_l &= W_l^V \overline{H}_l \end{aligned} \tag{2}$$

where $Q_l, K_l, \mathrm{V}_l \in R^{d\times d}$ represent the query matrix Q, key matrix K, and value matrix V for the feature sequence of the modality, and $W_l^Q, W_l^K, W_l^V \in R^{d\times d}$ represent the corresponding weight matrices.

Next, the dot product operation is performed on the query matrix and the key matrix for both speech and text. The softmax function is then applied to scale and normalize the results row-wise to obtain the attention weights. Finally, the feature sequences are aggregated using the corresponding weights to obtain the interactive information transferred between the two modalities.

1. Cross-Modal Transfer $S \to T$

The information from the speech modality is transferred to the text modality using a cross-modal attention mechanism with h heads. Unlike the original multi-head self-attention mechanism where $Q = K = V$, with the cross-modal attention mechanism, the query matrix is Q_S, and the key matrix and value matrix are K_T and V_T, respectively. This mechanism facilitates the transfer of speech information to the text modality, enabling learning the text feature representations as guided by the speech information. The similarity is computed by taking the dot product of the query matrix Q_S from the speech and the key matrix K_T from the text. The $Softmax$ function is applied to scale and normalize the results, followed by multiplying them with the value matrix V_T to obtain the attention weights. The specific formula is as follows:

$$Att_{S\to T}\left(\overline{H}_S, \overline{H}_T\right) = \mathrm{softmax}\left(\frac{Q_S K_T^T}{\sqrt{d_k}}\right) V_T \tag{3}$$

Then, the results from h heads are concatenated and mapped. The specific process is as follows:

$$M_{S\to T}\left(\overline{H}_S, \overline{H}_T\right) = Concat(Att_{\mathrm{S}\to\mathrm{T}}(1), \ldots, Att_{S\to T}(i), \ldots, Att_{S\to T}(h))W \tag{4}$$

where $Att_{S\to T}(i)$ represents the i-th (where $i \in [1,h]$) cross-modal attention weight.

Finally, residual connection and layer normalization are applied to the single-modal speech features $\overline{H}_S$ mapped using 1D-CNN and the interactive multi-modal features $M_{S\to T}$. The specific formula is as follows:

$$CM_{S\to T} = LayerNorm\left(\overline{H}_S + M_{S\to T}\left(\overline{H}_S, \overline{H}_T\right)\right) \tag{5}$$

where $CM_{S\to T}$ is the cross-modal output representation from the speech-to-text modality, containing not only complementary information from both modalities but also the original speech emotion features, effectively reducing the information loss.

2. Cross-Modal Transfer $T \to S$

The process of transferring information from the text modality to the speech modality is similar to the process of transferring it from the speech modality to the text modality. All the computation formulas are as follows:

$$\begin{aligned} Att_{T\to S}\left(\overline{H}_S, \overline{H}_T\right) &= \text{softmax}\left(\frac{Q_T K_S^T}{\sqrt{d_k}}\right) V_S \\ M_{T\to S}\left(\overline{H}_S, \overline{H}_T\right) &= Concat(Att_{T\to S}(1), \ldots, Att_{T\to S}(i), \ldots, Att_{T\to S}(h))W \\ CM_{T\to S} &= LayerNorm\left(\overline{H}_S + M_{T\to S}\left(\overline{H}_S, \overline{H}_T\right)\right) \end{aligned} \tag{6}$$

4.2. The Emotion Classification Module

In the emotion classification module, the TAB sub-block is employed to focus on the multi-modal fused feature representation after the cross-modal information interaction. Initially, the two cross-modal fused features are concatenated to obtain the ultimate representation of the multi-modal fused emotion features, denoted as E^{fusion} and expressed as follows:

$$E^{fusion} = [CM_{S\to T}, CM_{T\to S}] \tag{7}$$

Subsequently, the fused features are put into the TAB3 sub-block, which is designed to capture the contextual relationships between the features, thereby capturing temporal-aware representations. The resulting multi-modal features are denoted as P. These multi-modal features P are then fed into a fully connected layer, where linear transformations are applied to learn the correlations between features and map them to the output space. A softmax classifier is utilized to obtain the multi-modal emotion recognition results based on both speech and text.

5. Experimental Verification

5.1. Experimental Simulation Parameters

The experimental environment for this study is shown in Table 1.

Table 1. Experimental environment.

Name	Specific Configuration
Operating System	Windows 11
Processor	NVIDIA GeForce RTX 4060 Ti (NVIDIA, Santa Clara, CA, USA)
Memory	16 GB
OS Bit	64-bit
Programming Language	Python 3.8
IDE	PyCharm 2023.1.2
Dataset	IEMOCAP and MELD
Deep Learning Framework	PyTorch 2.1.0

5.2. Experimental Data

5.2.1. Interactive Emotional Dyadic Motion Capture (IEMOCAP) [36]

This study was primarily conducted on the publicly available dataset IEMOCAP. It was created by the SAIL (Signal Analysis and Interpretation Laboratory) at the University of Southern California and is a multimodal database widely used in emotion recognition. The dataset comprises approximately 12 h of audiovisual data, including videos, speech, facial motion capture, and text. It consists of five sessions recorded with ten different actors, with each session featuring recordings of two speakers, one male and one female.

The dataset is labeled into four main emotion categories (Figure 8): anger (1102), sadness (1083), neutral (1708), and happiness (1636). To ensure fair comparisons when evaluating our model on the IEMOCAP dataset, we employed a five-fold cross-validation method, where one session was held out as the test set for each training iteration. The training–test split for each iteration is illustrated in Figure 9. It can be observed that the data distribution for each cross-validation is relatively uniform.

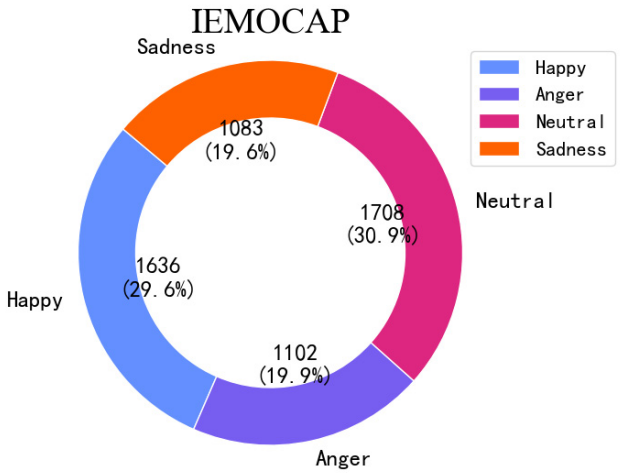


Figure 8. IEMOCAP data distribution.

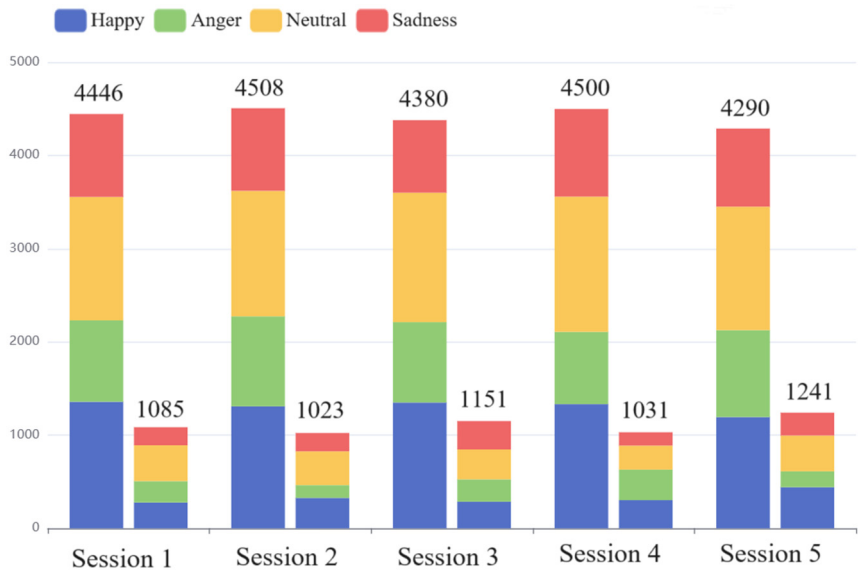


Figure 9. Each training test's data. The *x*-axis represents the distribution of training and testing data when session *i* is used as the testing set, where the left column represents the training data and the right column represents the testing data. The *y*-axis represents the quantity of data.

5.2.2. Multimodal EmotionLines Dataset (MELD) [37]

MELD evolved from the EmotionLines dataset and is a multimodal emotional dialogue dataset. It consists of organized dialogue from the popular American TV series *Friends*, comprising 1433 instances of dialogue with a total of 13,708 utterances. Each utterance in the dialogue is annotated with one of seven emotion labels: anger, disgust, fear, joy, surprise, sadness, or neutral. Additionally, each utterance in MELD is annotated as positive, negative, or neutral. The distribution of the training/test and validation data samples in MELD is shown in Figure 10.

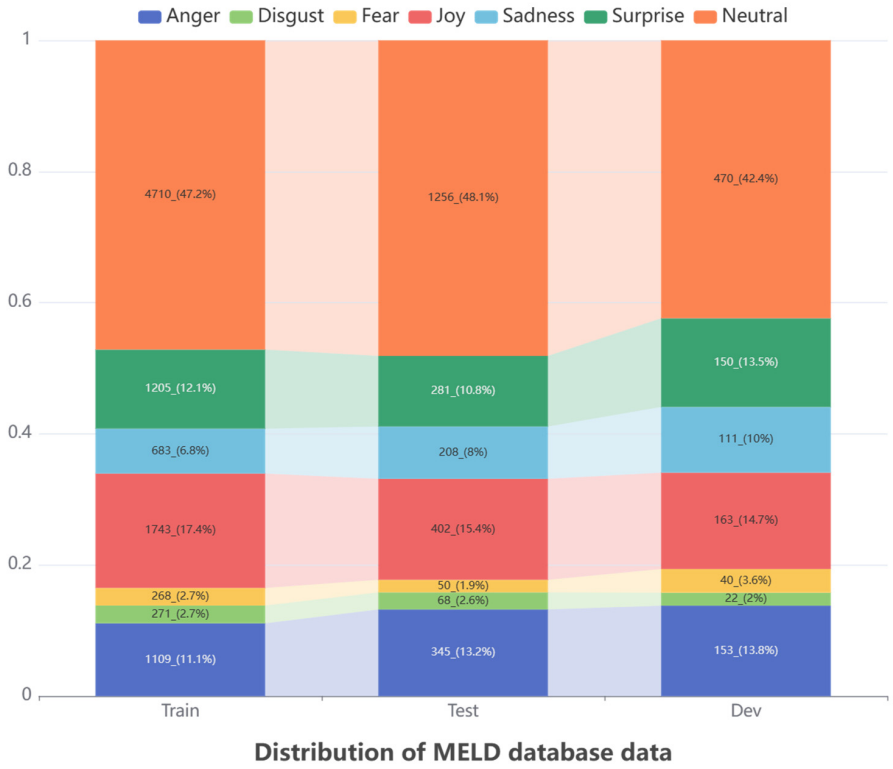


Figure 10. Training/testing and validation data samples for MELD.

5.3. Experimental Analysis

5.3.1. Evaluation Metrics

To evaluate the performance of the model, we use weighted average recall (WAR) and unweighted average recall (UAR) as the evaluation metrics. UAR averages the recall for each class without considering the number of samples per class. WAR, on the other hand, considers the number of samples for each class and calculates a weighted average recall. The difference between UAR and WAR lies in whether they consider the weights of the class sample sizes. UAR treats each class equally, while WAR assigns different weights based on the sample sizes of classes. The formulas for calculating UAR and WAR are as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

$$UAR = \frac{1}{N}\sum_{i=1}^{N} Recall_i \tag{9}$$

$$WAR = \sum_{i=1}^{N}\left(\frac{Recall_i \times Class\ Size_i}{Total\ Sample\ Size}\right) \tag{10}$$

where TP represents True Positives; TN represents True Negatives; FP represents False Positives; FN represents False Negatives; N is the number of classes; $Class\ Size_i$ is the sample size of the *i-th* class; and Total Sample Size is the total sample size across all classes.

5.3.2. Mul-TAB Result Analysis

This section analyzes the accuracy, robustness, hyperparameters, TAB3, and multi-head attention mechanism of Mul-TAB.

1. Accuracy Analysis

In this study, we proposed a speech emotion recognition model based on the improved TIM-Net network and multimodal fusion and conducted experiments on the IEMOCAP and MELD datasets. The test accuracy of our model is shown in the comparative line charts in Figures 11 and 12. On the IEMOCAP dataset, our model achieved a testing accuracy, as shown in Figure 11, reflected in the best WAR reaching 83.9% and the best UAR reaching 82.0%. For MELD, the best testing accuracy was 63.6%, and the training loss was 0.359, as illustrated in Figure 12. The corresponding confusion matrices for the experiments are shown in Figure 13a on the IEMOCAP dataset, our model performed best in recognizing the "anger" emotion category, with an accuracy of 90.5%, while the recognition accuracy of the "neutral" emotion category was the lowest at 73.5%. Figure 13b for MELD, our model performed best in recognizing the "neutral" emotion category, with an accuracy of 89.9%, while the recognition accuracy of the "disgust" emotion category was the lowest at 8.82%.

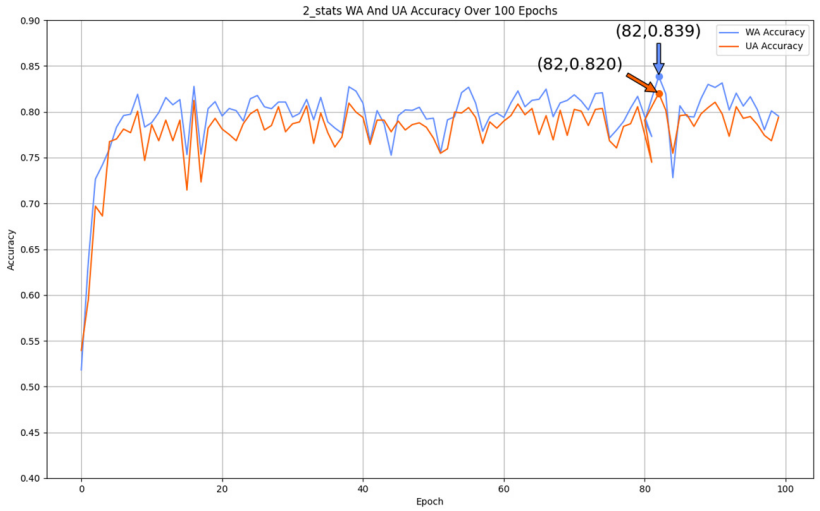


Figure 11. Test accuracy comparison when using Session 2 as the test set. (Experiments conducted on the IEMOCAP dataset).

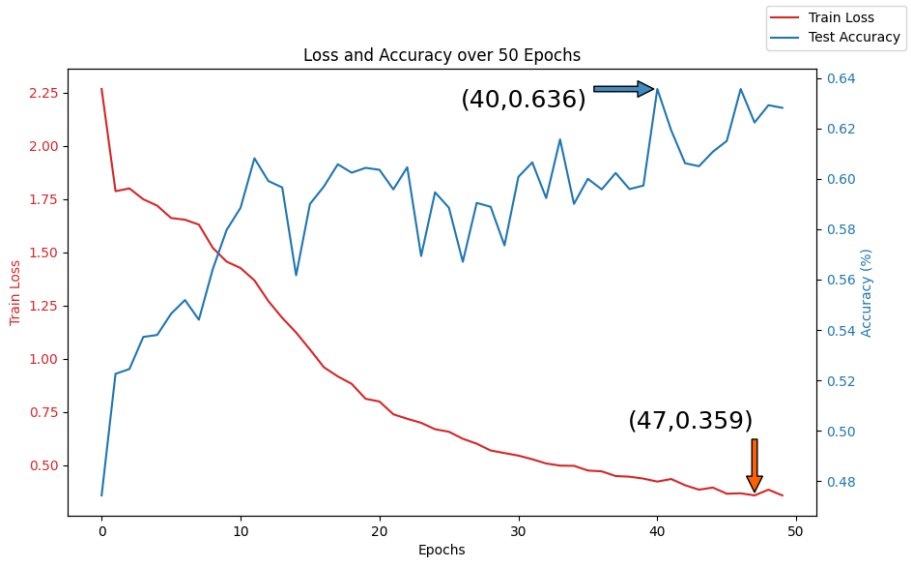


Figure 12. Training loss and testing accuracy on MELD.

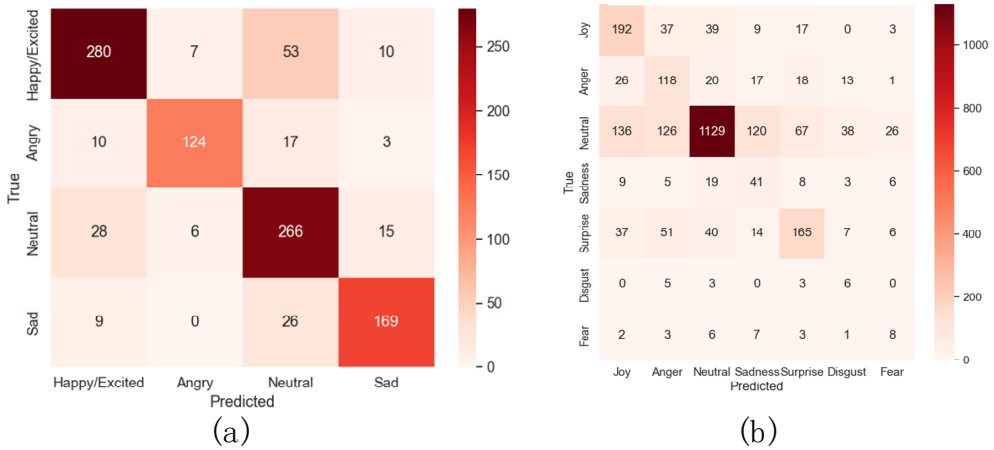


Figure 13. Confusion matrices for the (**a**) IEMOCAP and (**b**) MELD datasets.

2. Robustness Analysis

During training, we used 80% of the dataset as the training set and 20% as the test set. As the IEMOCAP dataset is divided into five sessions, we performed five-fold cross-validation by leaving one session out as the test set in each training iteration. Through multiple experiments, it can be observed that our model exhibits strong robustness. A comparative line chart of the model's test accuracy is depicted in Figure 14.

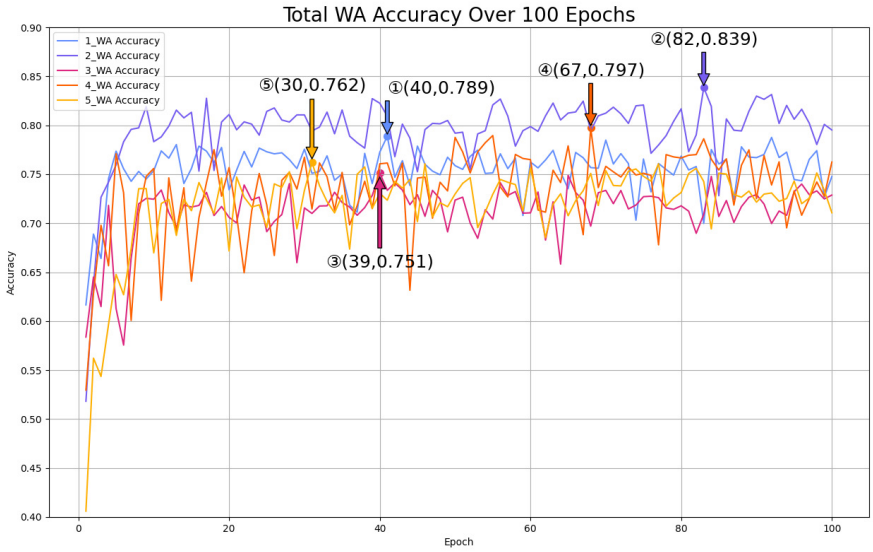


Figure 14. Comparative line chart of cross-validation accuracy.

3. Hyperparameter Analysis

Given that we used RoBERT$_a$ and Wav2Vec2.0 for the feature extraction, the feature dimension was 768. After multiple rounds of training and testing, it was observed that the optimal configuration is a batch size of 2, 100 epochs, a learning rate of 5×10^{-5}, three layers in the TAB sub-block, and two layers in the multi-head attention mechanism. Using a single GPU, each training and inference iteration takes approximately 4–5 min. The entire training process, with five iterations and 100 epochs, requires approximately $\frac{5 \times 100 \times 5}{60 \times 24} \approx 1.7$ days. The Mul-TAB model has a total of 143 million parameters.

4. TAB3 Analysis

The emotion classification module of the model was modified by removing the TAB sub-block, and the resulting emotion classification results are shown in Figure 15. The best achieved WAR is 82.9%. This indicates that the TAB sub-block, which is capable

of capturing the contextual relationships in the features and obtaining temporal-aware representations, contributes to the improvement of the model's performance.

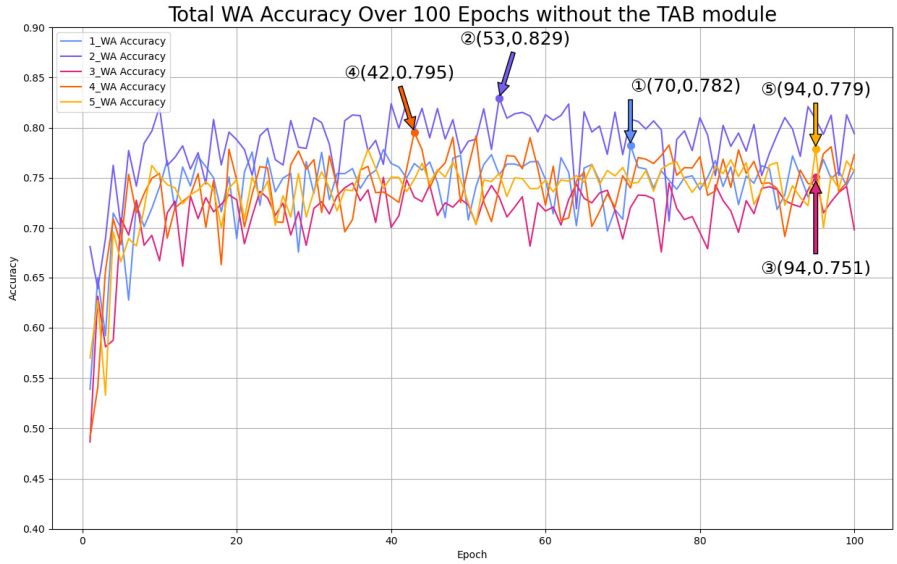


Figure 15. Cross-validation accuracy comparison without the TAB module.

5. Multi-Head Attention Mechanism Analysis

The feature extraction module of the model was modified by removing the multi-head attention mechanism. In this setup, only Wav2Vec2.0 feature extraction is performed for speech data, and only BERT feature extraction is performed for text data, while the subsequent steps remain unchanged. The resulting emotion classification results are shown in Figure 16, with the best achieved WAR being 82.3%. This indicates that the multi-head attention mechanism helps in learning more discriminative features for both speech and text, thereby enhancing the model's performance.

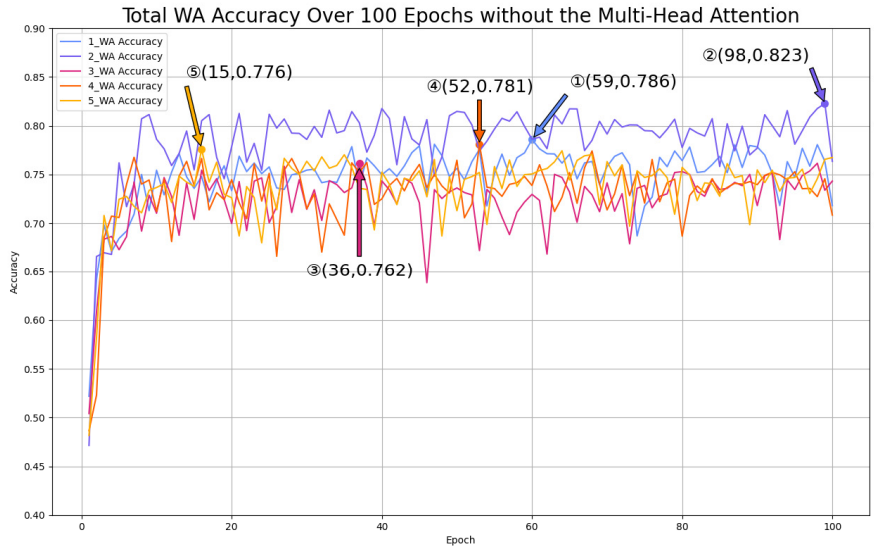


Figure 16. Cross-validation accuracy comparison without multi-head attention.

When comparing the algorithms with and without the TAB module, which lacks the multi-head attention mechanism, with the final algorithm, we obtain the results shown in Table 2. Analysis of Figure 17 reveals that individual use of the TAB module and the multi-head attention mechanism is not as effective as their combined usage.

Table 2. Results of ablation experiments. The table has been converted into a bar chart, as shown in Figure 17.

Model	IEMOCAP					MELD
	1_WA (%)	2_WA (%)	3_WA (%)	4_WA (%)	5_WA (%)	UA (%)
Mul-TAB	78.9	83.9	75.1	79.7	76.2	63.9
Without TAB	78.2	82.9	75.1	79.5	77.9	60.2
Without multi-head attention	78.6	82.3	76.2	78.1	77.6	59.4

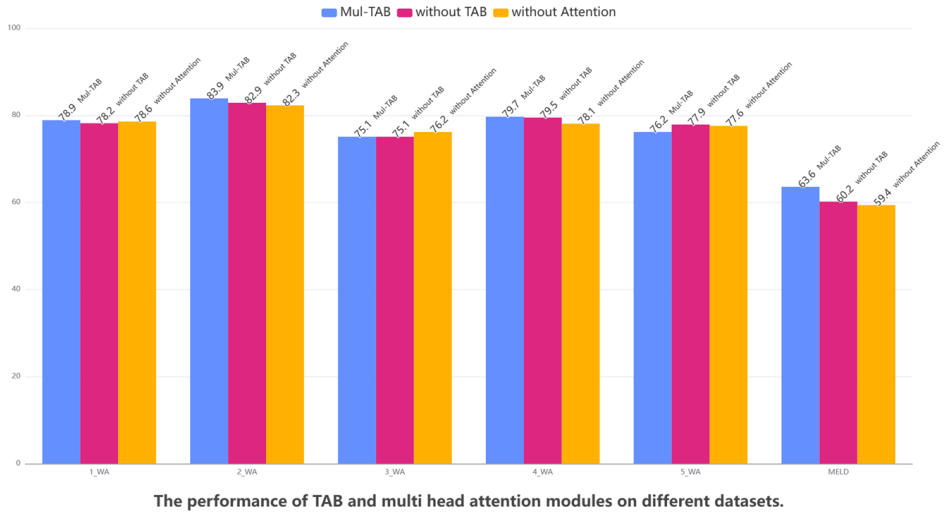


Figure 17. The performance of TAB and multi-head attention modules on different datasets. The first five columns are the experimental results on IEMOCAP, and the last column is the experimental results on MELD.

5.3.3. Comparison and Analysis with Other Experiments

In this paper, we proposed a speech emotion recognition model based on the improved TIM-Net network and multimodal fusion, which was experimentally validated on the IEMOCAP and MELD datasets. To evaluate our approach, we compared it with other models. The results show that our method performs well regarding WAR and UAR. The comparative results are presented in Tables 3 and 4, where "CV" stands for cross-validation, "5-fold" indicates five-fold cross-validation, and "10-fold" indicates ten-fold cross-validation. (Note: The UAR and WAR for MMER in the table are results reproduced using only the pure IEMOCAP dataset after removing the enhanced speech and text data added by the authors of the MMER paper).

Table 3. Comparative results with other models on IEMOCAP dataset. In the modal column, *S* represents speech, *T* represents text, and *V* represents vision.

Models	CV Type	Modality	Metrics	
			UAR (%)	WAR (%)
MHA + DRN [52]	-	{*S*}	67.40	-
CNN + Bi-GRU [53]	-	{*S*}	71.72	70.39
MSCNN-SPU [54]	10-fold	{*S*, *T*}	78.20	77.40
LightSER [55]	10-fold	{*S*}	70.76	70.23
Article [56]	5-fold	{*S*, *T*}	-	76.31
TIM-Net [35]	10-fold	{*S*}	72.50	71.65
MMER [57]	5-fold	{*S*, *T*}	78.69	80.18
Mul-TAB (ours)	5-fold	{*S*, *T*}	81.92	83.85

Table 4. Comparative results with other models on MELD.

Models	Modality	UAR (%)
MTAF [58]	{*S*, *T*}	48.12
BLSTM + IA-MMTF [59]	{*S*, *T*}	54.79
M2R2 [60]	{*S*, *T*, *V*}	55.83
MMGCN [61]	{*S*, *T*, *V*}	58.65
DAG-ERC [62]	{*S*}	61.04
CTNet [63]	{*S*, *T*}	62.0
Mul-TAB (ours)	{*S*, *T*}	63.6

6. Conclusions

This experiment investigated a deep-learning-based multimodal emotion recognition approach. The proposed multimodal fusion method based on TIM-Net and a multi-head attention mechanism demonstrates significant advantages in speech emotion recognition tasks, effectively improving the accuracy of emotion classification. This paper extensively discussed the methods for fusing multimodal features, elucidates and analyzes the TIM-Net model, and proposes enhancements. Finally, it combined the multimodal features to accomplish emotion recognition. Through analyzing and discussing the experimental results, we gain further insights into the contributions of different modal features. These research findings have important guiding significance and application value for future studies and applications in speech emotion recognition.

The model we propose also has shortcomings, such as identifying emotions by capturing the context of the conversation and applying common sense reasoning to understand the emotional changes in the conversation between the listener and the speaker. Future research directions include further optimizing the multimodal feature extraction and fusion methods to enhance the collaboration between different feature modalities. Additionally, an in-depth exploration of the design and implementation details of the TIM-Net model is necessary to optimize the model's structure and parameters further. Exploring more effective training methods and optimization strategies to improve the model's generalization ability, studying cross-domain and cross-language speech emotion recognition issues, and combining common sense reasoning are also required.

Author Contributions: Conceptualization, Y.W.; methodology, S.Z.; software, S.Z.; validation, Y.W., S.Z. and P.L.; formal analysis, Y.W.; investigation, S.Z.; writing—original draft preparation, Y.W. and S.Z.; writing—review and editing, Y.W., S.Z. and P.L.; project administration, Y.W.; funding acquisition, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Key R&D Program of China (program no. 2021YFF0603904) and in part by the Fundamental Research Funds for the Central Universities (program no. ZJ2022-004).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural language processing
TIM-Net	Temporal-Aware Bi-direction Multi-Scale Network
TAB	Temporal-aware block
SER	Speech emotion recognition
FER	Facial emotion recognition

LLD	Low-Level Descriptor
MFCCs	Mel-Frequency Cepstral Coefficients
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
BiLSTM	Bidirectional Long Short-Term Memory
TF-IDF	Term Frequency–Inverse Document Frequency
Glove	Global Vectors for Word Representation
Elmo	Embeddings from Language Models
BERT	Bidirectional Encoder Representation from Transformers
MLM	Masked Language Model
NSP	Next Sentence Prediction
LSTM	Long Short-Term Memory
ML-LSTM	Multi-Layer LSTM
FC	Fully connected
DC Conv	Dilated causal convolution
CME	Cross-Modal Encoder

References

1. Ekman, P. An argument for basic emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [CrossRef]
2. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124–129. [CrossRef]
3. Picard, R.W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 2000.
4. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5200–5204.
5. Zepf, S.; Hernandez, J.; Schmitt, A.; Minker, W.; Picard, R.W. Driver Emotion Recognition for Intelligent Vehicles: A Survey. *ACM Comput. Surv.* **2020**, *53*, 1–30. [CrossRef]
6. Franzoni, V.; Milani, A.; Nardi, D.; Vallverdú, J. Emotional machines: The next revolution. *Web Intell.* **2019**, *17*, 1–7. [CrossRef]
7. Rheu, M.; Shin, J.Y.; Peng, W.; Huh-Yoo, J. Systematic Review: Trust-Building Factors and Implications for Conversational Agent Design. *Int. J. Hum. Comput. Interact.* **2021**, *37*, 81–96. [CrossRef]
8. Suryadevara, N.K.; Mukhopadhyay, S.C. Determining wellness through an ambient assisted living environment. *IEEE Intell. Syst.* **2014**, *29*, 30–37. [CrossRef]
9. Suryadevara, N.K.; Chen, C.-P.; Mukhopadhyay, S.C.; Rayudu, R.K. Ambient assisted living framework for elderly wellness determination through wireless sensor scalar data. In Proceedings of the Seventh International Conference on Sensing Technology, Wellington, New Zealand, 3–5 December 2013; pp. 632–639.
10. Ghayvat, H.; Awais, M.; Pandya, S.; Ren, H.; Akbarzadeh, S.; Chandra Mukhopadhyay, S.; Chen, C.; Gope, P.; Chouhan, A.; Chen, W. Smart aging system: Uncovering the hidden wellness parameter for well-being monitoring and anomaly detection. *Sensors* **2019**, *19*, 766. [CrossRef]
11. Poorna, S.S.; Nair, G.J. Multistage classification scheme to enhance speech emotion recognition. *Int. J. Speech Technol.* **2019**, *22*, 327–340. [CrossRef]
12. Wang, Y.; Shen, G.; Xu, Y.; Li, J.; Zhao, Z. Learning Mutual Correlation in Multimodal Transformer for Speech Emotion Recognition. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Brno, Czech Republic, 30 August–3 September 2021; pp. 4518–4522.
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
14. Zou, H.; Si, Y.; Chen, C.; Rajan, D.; Chng, E.S. Speech Emotion Recognition with Co-Attention based Multi-level Acoustic Information. In Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022.
15. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460.
16. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **2019**, *10*, 18–31. [CrossRef]
17. Bakariya, B.; Singh, A.; Singh, H.; Raju, P.; Rajpoot, R.; Mohbey, K.K. Facial emotion recognition and music recommendation system using cnn-based deep learning techniques. *Evol. Syst.* **2024**, *15*, 641–658. [CrossRef]
18. Meena, G.; Mohbey, K.K.; Indian, A.; Khan, M.Z.; Kumar, S. Identifying emotions from facial expressions using a deep convolutional neural network-based approach. *Multimed. Tools Appl.* **2024**, *83*, 15711–15732. [CrossRef]
19. Lisitsa, E.; Benjamin, K.S.; Chun, S.K.; Skalisky, J.; Hammond, L.E.; Mezulis, A.H. Loneliness among Young Adults during COVID-19 Pandemic: The Mediational Roles of Social Media Use and Social Support Seeking. *J. Soc. Clin. Psychol.* **2020**, *39*, 708–726. [CrossRef]

20. Mohbey, K.K.; Meena, G.; Kumar, S.; Lokesh, K. A CNN-LSTM-Based Hybrid Deep Learning Approach for Sentiment Analysis on Monkeypox Tweets. *New Gener. Comput.* **2023**, 1–19. [CrossRef]
21. Nguyen, D.; Nguyen, K.; Sridharan, S.; Ghasemi, A.; Dean, D.; Fookes, C. Deep spatio-temporal features for multimodal emotion recognition. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 1215–1223.
22. Guanghui, C.; Xiaoping, Z. Multi-modal emotion recognition by fusing correlation features of speech-visual. *IEEE Signal Process. Lett.* **2021**, *28*, 533–537. [CrossRef]
23. Wang, Y.; Shen, Y.; Liu, Z.; Liang, P.P.; Zadeh, A.; Morency, L.P. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 7216–7223.
24. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. *arXiv* **2017**, arXiv:1707.07250.
25. Zhu, L.; Zhu, Z.; Zhang, C.; Xu, Y.; Kong, X. Multimodal sentiment analysis based on fusion methods: A survey. *Inf. Fusion* **2023**, *95*, 306–325. [CrossRef]
26. Gandhi, A.; Adhvaryu, K.; Poria, S.; Cambria, E.; Hussain, A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Inf. Fusion* **2023**, *91*, 424–444. [CrossRef]
27. Zhang, Y.; Song, D.; Zhang, P.; Wang, P.; Li, J.; Li, X.; Wang, B. A quantum-inspired multimodal sentiment analysis framework. *Theor. Comput. Sci.* **2018**, *752*, 21–40. [CrossRef]
28. Liang, P.P.; Zadeh, A.; Morency, L.P. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv* **2022**, arXiv:2209.03430.
29. Hazarika, D.; Zimmermann, R.; Poria, S. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1122–1131.
30. Sun, H.; Wang, H.; Liu, J.; Chen, Y.W.; Lin, L. Cubemlp: An MLP-based model for multimodal sentiment analysis and depression estimation. In Proceedings of the 30th ACM International Conference on Multimedia, New York, NY, USA, 10–14 October 2022; pp. 3722–3729.
31. Chen, M.; Wang, S.; Liang, P.P.; Baltrušaitis, T.; Zadeh, A.; Morency, L.P. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017; pp. 163–171.
32. Bates, J. The role of emotion in believable agents. *Commun. ACM* **1994**, *37*, 122–125. [CrossRef]
33. Hayes-Roth, B.; Doyle, P. Animate characters. *Auton. Agents Multi-Agent Syst.* **1998**, *1*, 195–230. [CrossRef]
34. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
35. Ye, J.; Wen, X.C.; Wei, Y.; Xu, Y.; Liu, K.; Shan, H. Temporal Modeling Matters: A Novel Temporal Emotional Modeling Approach for Speech Emotion Recognition. In Proceedings of the CASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
36. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
37. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv* **2018**, arXiv:1810.02508.
38. Lee, S.; Han, D.K.; Ko, H. Fusion-ConvBERT: Parallel Convolution and BERT Fusion for Speech Emotion Recognition. *Sensors* **2020**, *20*, 6688. [CrossRef]
39. Dai, W.; Cahyawijaya, S.; Liu, Z.; Fung, P. Multimodal end-to-end sparse model for emotion recognition. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, Online, 6–11 June 2021; pp. 5305–5316.
40. Huddar, M.G.; Sannakki, S.S.; Rajpurohit, V.S. Attention-based Multi-modal Sentiment Analysis and Emotion Detection in Conversation using RNN. *Int. J. Interact. Multimed. Artif. Intell.* **2021**, *6*, 112–121. [CrossRef]
41. Graves, A.; Fernández, S.; Schmidhuber, J. Bidirectional LSTM networks for improved phoneme classification and recognition. In Proceedings of the Artificial Neural Networks: Formal Models and Their Applications, Warsaw, Poland, 11–15 September 2005; pp. 799–804.
42. Eyben, F.; Wöllmer, M.; Graves, A.; Schuller, B.; Douglas-Cowie, E.; Cowie, R. On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. *J. Multimodal User Interfaces* **2010**, *3*, 7–19. [CrossRef]
43. Wu, Y.; Li, G.; Fu, Q. Non-Intrusive Air Traffic Control Speech Quality Assessment with ResNet-BiLSTM. *Appl. Sci.* **2023**, *13*, 10834. [CrossRef]
44. Chatterjee, A.; Narahari, K.N.; Joshi, M.; Agrawal, P. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 39–48.
45. Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523. [CrossRef]
46. Deng, J.; Ren, F. A survey of textual emotion recognition and its challenges. *IEEE Trans. Affect. Comput.* **2021**, *14*, 49–67. [CrossRef]
47. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 3111–3119.

48. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
49. Ilić, S.; Marrese-Taylor, E.; Balazs, J.A.; Matsuo, Y. Deep contextualized word representations for detecting sarcasm and irony. *arXiv* **2018**, arXiv:1809.09795.
50. D'mello, S.K.; Kory, J. A Review and Meta-Analysis of Multimodal Affect Detection Systems. *ACM Comput. Surv.* **2015**, *47*, 1–36. [CrossRef]
51. Zhang, L.; Na, J.; Zhu, J.; Shi, Z.; Zou, C.; Yang, L. Spatiotemporal causal convolutional network for forecasting hourly PM2.5 concentrations in Beijing, China. *Comput. Geosci.* **2021**, *155*, 104869. [CrossRef]
52. Li, R.; Wu, Z.; Jia, J.; Zhao, S.; Meng, H. Dilated residual network with multi-head self-attention for speech emotion recognition. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6675–6679.
53. Zhong, Y.; Hu, Y.; Huang, H.; Silamu, W. A Lightweight Model Based on Separable Convolution for Speech Emotion Recognition. In Proceedings of the INTERSPEECH, Shanghai, China, 25–29 November 2020; pp. 3331–3335.
54. Peng, Z.; Lu, Y.; Pan, S.; Liu, Y. Efficient Speech Emotion Recognition Using Multi-Scale CNN and Attention. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 3020–3024.
55. Aftab, A.; Morsali, A.; Ghaemmaghami, S.; Champagne, B. LIGHT-SERNET: A lightweight fully convolutional neural network for speech emotion recognition. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 6912–6916.
56. Zhao, Z.; Wang, Y.; Wang, Y. Multi-level fusion of wav2vec 2.0 and BERT for multimodal emotion recognition. *arXiv* **2022**, arXiv:2207.04697.
57. Ghosh, S.; Tyagi, U.; Ramaneswaran, S.; Srivastava, H.; Manocha, D. MMER: Multimodal Multi-task Learning for Speech Emotion Recognition. *arXiv* **2022**, arXiv:2203.16794.
58. Wang, Y.; Gu, Y.; Yin, Y.; Han, Y.; Zhang, H.; Wang, S.; Li, C.; Quan, D. Multimodal transformer augmented fusion for speech motion recognition. *Front. Neurorobot.* **2023**, *17*, 1181598. [CrossRef]
59. Guo, L.; Wang, L.; Dang, J.; Fu, Y.; Liu, J.; Ding, S. Emotion Recognition with Multimodal Transformer Fusion Framework Based on Acoustic and Lexical Information. *IEEE MultiMedia* **2022**, *29*, 94–103. [CrossRef]
60. Wang, N.; Cao, H.; Zhao, J.; Chen, R.; Yan, D.; Zhang, J. M2R2: Missing-Modality Robust emotion Recognition framework with iterative data augmentation. *IEEE Trans. Artif. Intell.* **2022**, *4*, 1305–1316. [CrossRef]
61. Hu, J.; Liu, Y.; Zhao, J.; Jin, Q. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. *arXiv* **2021**, arXiv:2107.06779.
62. Shen, W.; Wu, S.; Yang, Y.; Quan, X. Directed acyclic graph network for conversational emotion recognition. *arXiv* **2021**, arXiv:2105.12907.
63. Lian, Z.; Liu, B.; Tao, J. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 985–1000. [CrossRef]