electronics

MDPI

Article

Speech Emotion Recognition Using Dual-Stream Representation and Cross-Attention Fusion

Shaode Yu [1], Jiajian Meng [1], Wenqing Fan [1], Ye Chen [1], Bing Zhu [1], Hang Yu [2], Yaoqin Xie [3] and Qiurui Sun [4,*]

[1] School of Information and Communication Engineering, Communication University of China, Beijing 100024, China; yushaodemia@163.com (S.Y.); mengjiajian@cuc.edu.cn (J.M.); fanwenqing@cuc.edu.cn (W.F.); 2020211023044@cuc.edu.cn (Y.C.); zhubing1218@163.com (B.Z.)
[2] School of Aerospace Science and Technology, Xidian University, Xi'an 710126, China; yuhang9551@163.com
[3] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; yq.xie@siat.ac.cn
[4] Center of Information & Network Technology, Beijing Normal University, Beijing 100875, China
* Correspondence: qiuruisun@bnu.edu.cn

Abstract: Speech emotion recognition (SER) aims to recognize human emotions through in-depth analysis of audio signals. However, it remains challenging to encode emotional cues and to fuse the encoded cues effectively. In this study, dual-stream representation is developed, and both full training and fine-tuning of different deep networks are employed for encoding emotion patterns. Specifically, a cross-attention fusion (CAF) module is designed to integrate the dual-stream output for emotion recognition. Using different dual-stream encoders (fully training a text processing network and fine-tuning a pre-trained large language network), the CAF module is compared to other three fusion modules on three databases. The SER performance is quantified with weighted accuracy (WA), unweighted accuracy (UA), and F1-score (F1S). The experimental results suggest that the CAF outperforms the other three modules and leads to promising performance on the databases (EmoDB: WA, 97.20%; UA, 97.21%; F1S, 0.8804; IEMOCAP: WA, 69.65%; UA, 70.88%; F1S, 0.7084; RAVDESS: WA, 81.86%; UA, 82.75.21%; F1S, 0.8284). It is also found that fine-tuning a pre-trained large language network achieves superior representation than fully training a text processing network. In a future study, improved SER performance could be achieved through the development of a multi-stream representation of emotional cues and the incorporation of a multi-branch fusion mechanism for emotion recognition.

Keywords: speech emotion recognition; dual-stream representation; cross-attention fusion; large language model; text processing network

Citation: Yu, S.; Meng, J.; Fan, W.; Chen, Y.; Zhu, B.; Yu, H.; Xie, Y.; Sun, Q. Speech Emotion Recognition Using Dual-Stream Representation and Cross-Attention Fusion. *Electronics* **2024**, *13*, 2191. https://doi.org/10.3390/electronics13112191

Academic Editor: Eng-Siong Chng

Received: 2 May 2024
Revised: 23 May 2024
Accepted: 30 May 2024
Published: 4 June 2024

1. Introduction

Emotion recognition is important in human speech communication [1], brain–computer interface [2], human–robot interaction [3], intelligent transportation systems [4], and affective computing [5]. It is also crucial in physiological signal analysis for certain mental state prediction in healthcare management [6] and emotional talking-face generation in digital humanities [7]. In general, accurate recognition of subjective human emotions benefits a wide range of real-life applications [8].

Human emotion can be conveyed in various ways. It encompasses verbal and non-verbal modalities in daily communication, such as verbal communication, facial expression, tone of voice, hand movement, eye contact, and physiological response [8]. Technically speaking, emotion signals can be captured through audio, video, text, electroencephalogram (EEG), and other types of routine and clinical recordings [9].

Many multi-modal frameworks [10] have been developed for automated emotion recognition, since multi-modal representation offers the potential for a thorough and nuanced comprehension of emotional states. Liu et al. explore peripheral physiological

signals, EEG, and facial videos, and emotion dictionary learning with modality attention is proposed for mixed emotion recognition [11]. Tan et al. consider EEG and facial expressions, and a Monte Carlo simulation is used to merge the prediction probabilities [12]. Liu et al. incorporate EEG, peripheral physiological signals, electrocardiogram, and movement features, and deep canonical correlation analysis and bimodal auto-encoder are extended for multi-modal signal fusion [13]. Emotion recognition has also been devoted to fusing acoustic, linguistic, and textual information [14], such as speech–text dual-modal frameworks [15,16], audio–visual–text tri-modal frameworks [17], and multi-modal frameworks using audio, text, facial expressions, and hand movements [18].

No surprises arise from the fact that multi-modal frameworks have shown promising performance in emotion analysis [8]. However, in real-world scenarios, most of the aforementioned signals are either unavailable or impractical. For instance, the acquisition of physiological signals is very complicated, time-consuming, and expensive, and professional personnel are required in the procedure. Notably, in diverse scenarios, such as psychological consultations, medical dialogues, and routine communication, voice recordings often constitute the primary data available. Therefore, accurate emotion recognition from speech audio signals is paramount in these contexts [19].

This study focuses on speech emotion recognition (SER). Within the domain of emotion recognition, audio signals offer numerous advantages. Firstly, audio signals are the easiest data source among multimedia and physiological signals. These signals encapsulate not only linguistic content but also emotional cues, including pitch, tone, intensity, tempo, rhythm, and spectral content. Secondly, emotions expressed through audio signals tend to be more universal and less language-dependent than those conveyed through text or facial expressions. This universality makes audio signals valuable for emotion recognition across diverse cultures and languages. In other words, audio signals hold the potential for accurate emotion recognition, non-contact medical diagnosis, and human–computer interaction, providing diverse benefits across healthcare and technology applications [8].

A large number of SER algorithms have been proposed [1]. In terms of speech signal representation techniques, the algorithms can be categorized into single-, dual-, and multi-stream groups. Most existing approaches exploit single-stream representation to unearth emotional cues. (a) Traditional methods extract acoustic features, such as low-level description features, high-level statistical features, and spectrum features [20], followed by feature selection [21] and machine learning [22]. Spectrum features, including Mel-frequency cepstral coefficients (MFCCs) and Mel spectrograms, represent time–frequency changes in the frequency domain [20]. However, these features are prone to fit emotion-independent noise during model training [23]. Based on MFCC features, Sha et al. [24] introduce a time–frequency domain convolution module that uses hybrid dilated convolution to expand the receptive field of neurons and to enhance the feature richness and diversity. In addition, several toolboxes are available for crafting these commonly used audio signal features [25,26]. (b) Deep learning methods recast SER as an end-to-end optimization procedure for learning spatial, temporal, or joint feature representation [27]. Some works hand-craft the short-timeframe-level acoustic features and aggregate these features into an utterance-level representation using recurrent neural networks (RNNs) and local attention modules [28]. Liu et al. [29] design a one-dimensional convolutional neural network (CNN) to explore multi-scale and multichannel feature extraction for emotion classification. Some works quantify the audio as spectrogram images, and SER is treated as a deep-learning-based image classification problem [23,30]. Some works are driven by textual features. To capture the audio changes in time and frequency domains, these works transcribe speech into text, and language models are used to predict the emotion from the text [31]. However, without text carriers, the procedure of transcription might cause information loss, thereby reducing the SER accuracy [32]. (c) A hot topic is to explore lightweight large language models (LLMs), such as Bidirectional Encoder Representations from Transformers (BERT) [33], hidden-unit BERT (HuBERT) [34], wave to vector (Wav2Vec) [35], and Wav2Vec 2.0 (Wav2Vec2) [36]. Xia et al. [37] exploit learned features

from pre-trained Wav2Vec, which outperforms several traditional methods. Chen et al. [38] learn the contextualized emotional representation by fine-tuning Wav2Vec2. Sun et al. [39] implement a feed-forward network with skip connections to fine-tune Wav2Vec2 for emotion embeddings, and a contrastive-learning-based function is designed for discriminative feature representation. Pepino et al. [40] develop a transfer learning method in which features are extracted from pre-trained Wav2Vec2, and the framework transcribes speeches into text for final emotion prediction. Ma et al. [41] select data2vec [42] as a backbone and use generated congruent text and speech from a generative pre-trained Transformer (GPT) model [43] and an emotional text-to-speech model for data generation and augmentation to boost the SER performance. Feng and Narayanan [44] use multiple foundation models to assist speech emotion classification by curation covering transcribing, emotion annotation, and data augmentation.

To fully utilize the single modality of audio signals, dual- or multi-stream frameworks are implemented by using different perspectives of audio representation. This might come from the multi-timescale audio segment analysis for cepstral features, spectrogram images, and text transcription. Chen et al. use a log-Mel spectrogram and transcribe text for speech representation, and a multi-scale strategy of feature fusion and ensemble learning is applied to improve the performance [45]. Hu et al. propose a joint network combining pre-trained and spectrum-based networks, and an interactive attention module is designed for effective fusion of dual-stream output [19]. Tellai et al. introduce a dual-stream CNN-Transformer fusion network which captures both spatial and temporal information by disentangling long-distance MFCCs and Mel spectrograms [46]. Zou et al. extract three levels of acoustic information (MFCCs, spectrogram images, and audio waves) using different deep encoders, followed by concatenation for feature fusion and emotion recognition [47].

Even though remarkable SER performance has been achieved, how to represent speech signals comprehensively and how to integrate encoded features effectively remains a challenge [14,48]. To address these issues, this study employs a dual-stream representation to unearth emotional cues. One stream uses full training of a text processing network for encoding the log-Mel spectrograms, and the other employs fine-tuning of a pre-trained LLM for representing audio signals. Furthermore, a cross-attention fusion (CAF) module is designed for effective integration of the dual-stream outputs. On three databases, extensive experiments are conducted to validate the effectiveness of the dual-stream encoder, the CAF module, and the proposed SER framework. The contributions of this study can be summarized as follows:

1. Dual-stream representation is designed. It fully trains a text processing network and fine-tunes a language processing network for encoding emotional cues in audio.
2. A novel feature fusion module CAF is developed. It enables feature dimensionality alignment and generates a more informative representation for emotion recognition.
3. The proposed SER framework is validated on three public databases and achieves promising performance.

The remainder of this paper is organized as follows. Section 2 presents the relevant techniques of speech emotion representation and dual-stream feature fusion. Specifically, three dual-stream fusion modules are introduced, including summarization ("SUM"), concatenation ("CON"), and feature-wise linear modulation (FiLM) [49]. Section 3 describes the proposed dual-stream representation and CAF module. Data collection, experiment design, performance evaluation, and implementation details are presented in Section 4. The SER performance on the databases is reported from an ablation study and current achievements in Section 5. After that, the results and limitations are discussed in Section 6, and the study is concluded in Section 7.

2. Related Techniques

For dual- and multi-stream algorithms, feature fusion is important, since valid integration of emotional cues helps improve the SER performance. This section introduces the

related techniques on how to encode emotional patterns in audio signals and on how to integrate dual- or multi-stream output features.

2.1. Speech Emotion Representation

Diverse approaches have been proposed for encoding emotional cues in audio signals, including audio feature extraction methods [25,26] and deep-learning-based spatial, temporal, and joint feature representation methods [27].

Fine-tuning a pre-trained deep network is preferable for encoding emotion patterns. One representative is BERT [33], which has been pre-trained by using a large-scale corpus for learning general expression and strong feature representation. To standardize the segmentation of an audio utterance, HuBERT [34] explores audio clustering to generate discrete labels by adapting the data and labels into the loss calculation. Moreover, a masked loss model is used to restore the randomly masked parts of encoded features and the learned acoustic information in continuous data. Another representative is Wav2Vec [35] which borrows a masked CNN and a contrastive learning objective. It employs a context window and masked training objective to predict contextually relevant representations, and quantization is performed using vector quantization. By incorporating Transformers, Wav2Vec2 [36] has demonstrated improvement in learning more robust and useful speech representations. It designs the source of negative samples and the temperature annealing schedule followed by the quantification of the encoder output, and meanwhile, both contrast loss and diversity loss are used in the training phase.

On the other hand, text processing networks have rarely been used in audio signal representation [5]. In this study, two networks are investigated. One is the deep pyramid CNN (DPCNN) [50]. This consists of a stack of convolutional layers and shortcut connections, and the shortcut connections are used to mitigate the vanishing gradient problem. DPCNN has shown effectiveness in capturing both local and global contextual information for sentiment analysis [51]. The other text processing network is the text region CNN (TextRCNN) [52], which combines bidirectional RNN layers and convolutional layers. It uses the context information from both left and right directions to capture semantic information. TextRCNN has shown good performance in understanding of sequential dependencies and contextual information for sentiment analysis.

To the best of our knowledge, the utilization of text processing networks (DPCNN [51] and TextRCNN [52]) to encode emotional cues in speech audio signals represents a pioneering application, while the pre-trained networks of HuBERT [34] and Wav2Vec2 [36] have been widely used in speech emotion representation [5].

2.2. Dual-Stream Feature Fusion

Dual-stream representation of one single modality is described, and any representation method is denoted as a transfer function f with a specific name. Assume B_s training samples $\{(x_i, y_i)\}_{i=1}^{B_s}$, x_i is a sample and y_i is its label. Using different representation methods (F_1 and F_2), we obtain $x_\gamma \in \mathbb{R}^{D_\gamma}$ and $x_\beta \in \mathbb{R}^{D_\beta}$ to the input sample x_i, as shown in Equation (1). The outputs x_γ and x_β have different dimensions, D_γ and D_β.

$$\begin{aligned} x_\gamma &= f_{F_1}(x_i) \\ x_\beta &= f_{F_2}(x_i) \end{aligned} \tag{1}$$

After that, fully connected models (FC_1 and FC_2) are designed for aligning the feature dimensionality, as shown in Equation (2):

$$\begin{aligned} x'_\gamma &= f_{FC_1}(x_\gamma) \\ x'_\beta &= f_{FC_2}(x_\beta) \end{aligned} \tag{2}$$

where $x'_\gamma \in \mathbb{R}^{D_m}$ and $x'_\beta \in \mathbb{R}^{D_m}$ have the same feature dimensionality as D_m.

Finally, the features (x'_γ and x'_β) are fused. One widely used fusion paradigm is the SUM operation (Equation (3)),

$$O = x'_\gamma + x'_\beta \in \mathbb{R}^{D_m} \tag{3}$$

or the CON operation (Equation (4)),

$$O = [x'_\gamma; x'_\beta] \in \mathbb{R}^{(2D_m)} \tag{4}$$

of the output features. It should be noted that the fusion paradigm of concatenation could be used to fuse features of different dimensionalities.

The fusion paradigm of SUM and CON is simple and intuitive, while it might cause a bit of controversy. Even though the intermediate features are from the same network architecture, they are learned from different modalities or representations. Rough SUM or CON is prone to information conflicts, and the fused features become meaningless, potentially impeding model learning.

To address this issue, a dual-stream feature fusion strategy FiLM is developed [49]. The two features x'_γ and x'_β are split into two equal-sized parts as

$$\begin{aligned} x'_\gamma &= [v_{x'_\gamma}; \omega_{x'_\gamma}] \\ x'_\beta &= [v_{x'_\beta}; \omega_{x'_\beta}] \end{aligned} \tag{5}$$

where $v_{x'_\gamma}$, $\omega_{x'_\gamma}$, $v_{x'_\beta}$, and $\omega_{x'_\beta}$ are $\frac{D_m}{2}$-dimensional features. After that, new feature vectors are constructed as below:

$$\begin{aligned} x''_\gamma &= (v_{x'_\gamma} \odot \omega_{x'_\beta}) \oplus \omega_{x'_\gamma} \\ x''_\beta &= (v_{x'_\beta} \odot \omega_{x'_\gamma}) \oplus \omega_{x'_\beta} \end{aligned} \tag{6}$$

where $x''_\gamma \in \mathbb{R}^{\frac{D_m}{2}}$, $x''_\beta \in \mathbb{R}^{\frac{D_m}{2}}$, $\odot$ is the Hadamard product, and $\oplus$ denotes matrix addition. Finally, the output feature is formed by splicing x''_γ and x''_β using a simple CON fusion as

$$O_{ut} = [x''_\gamma; x''_\beta], \tag{7}$$

where $O_{ut} \in \mathbb{R}^{D_m}$.

In a word, the procedure of FiLM-based feature fusion consists of feature splitting (Equation (5)), re-construction (Equation (6)), and simple concatenation (Equation (7)).

3. The Proposed Method

The proposed framework is shown in Figure 1. It consists of dual-stream representation of audio signals and the CAF module of encoded cues. This section introduces the details of dual-stream representation using full training and fine-tuning of deep networks as well as the computation behind the CAF module. Any representation network in this section is treated as a mapping function f with a specific name.

3.1. Dual-Stream Representation of Audio Signals

Dual-stream representation consists of two subdivisions of feature encoding, as shown in Figure 1. Encoder A explores the temporal nature of speech, and DPCNN or TextRCNN is used as the encoder after the log-Mel spectrogram images are computed. Encoder B utilizes a pre-trained language model, HuBERT or Wav2Vec2, to encode the emotional cues buried in the original audio signals.

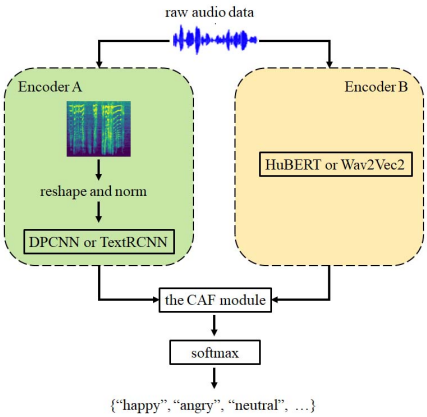

Figure 1. The structure of the proposed SER framework that uses dual-stream representation of speech audio signals and cross-attention fusion of encoded emotional cues for emotion classification.

Assuming original speech audio signals $x_w = \{x_i\}_{i=1}^{B_s}$, encoder A is made up of three steps. The first step uses f_{logMel} to compute the log-Mel spectrogram images,

$$M_{s\_w} = f_{\text{logMel}}(x_w), \tag{8}$$

and the spectrogram images are reshaped (f_{reshape}) and normalized (f_{norm}) as the input of the text processing networks:

$$M_{s\_w}^{n} = f_{\text{norm}}(f_{\text{reshape}}(M_{s\_w})). \tag{9}$$

For encoding emotion patterns buried in the speech signals, the network could be DPCNN [50]. It (f_{DPCNN}) takes a batch of pre-processed two-dimensional log-Mel spectrogram images ($[b, c, h, w]$) as the input:

$$x_{\gamma\_w} = f_{\text{DPCNN}}(M_{s\_w}^{n}). \tag{10}$$

An alternative is TextRCNN [52]. It (f_{TextRCNN}) uses the squeeze images as its input (Equation (11)). The image patch is adjusted from $[b, c, h, w]$ to ($[b, h, w]$), since the log-Mel spectrogram images are grayscale and the channel number is $c = 1$.

$$x_{\gamma\_w} = f_{\text{TextRCNN}}(M_{s\_w}^{n}). \tag{11}$$

In Equations (10) and (11), b denotes the batch size, c is the channel number of an image, and $[h, w]$, respectively, indicate the image height and width.

As for encoder B, since LLMs (HuBERT or Wav2Vec2) have been pre-trained on related large databases, the models need to be fed with the training samples x_w for fine-tuning and emotion encoding using HuBERT [34],

$$x_{\beta\_w} = f_{\text{HuBERT}}(x_w), \tag{12}$$

or using Wav2Vec2 [36],

$$x_{\beta\_w} = f_{\text{Wav2Vec2}}(x_w). \tag{13}$$

In the end, x_{γ_w} and x_{β_w} perform as the input of the CAF module. After the probability normalization of fusion results, the speech emotion is predicted. The procedure can be generally described as

$$E_{mo} = f_{\text{softmax}}(f_{\text{CAF}}(x_{\gamma\_w}, x_{\beta\_w})). \tag{14}$$

In summary, encoder A consists of computing the log-Mel spectrograms (Equation (8)), reshaping and normalization of spectrograms (Equation (9)), and emotion encoding using text processing models (Equation (10) or Equation (11)); encoder B utilizes the fine-tuning

of pre-trained models (Equation (12) or Equation (13)), and the dual-stream outputs are fused and the framework is trained towards speech emotion prediction (Equation (14)). How to define the output of the CAF module is described in Equation (20).

3.2. Cross-Attention Fusion of Encoded Cues

The CAF module is designed for integrating the dual-stream outputs towards emotion classification. It enables feature dimensionality alignment and yields a new feature vector. Figure 2 shows the feature fusion workflow for dual-stream representation.

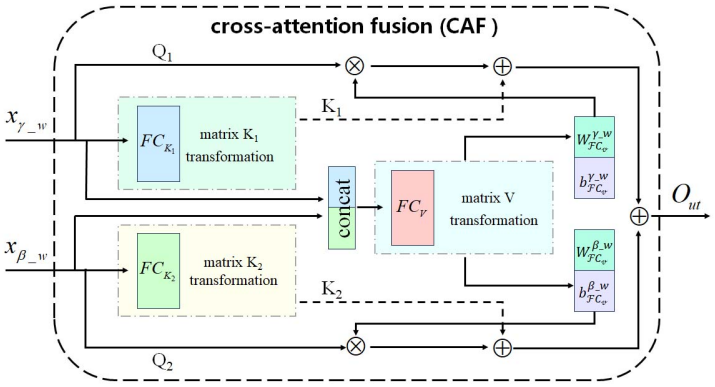

Figure 2. The structure of the proposed CAF module for dual-stream output integration.

The widely used "scaled dot-product attention" [53] practically requires transforming the three matrices of queries (Qs), keys (Ks), and values (Vs) into an output vector. In general, it computes the dot product of the query Q with all keys Ks, which is then divided by $\sqrt{d_k}$ (d_k, the dimension of queries), and finally, a softmax layer (f_{softmax}) is applied to tune the weights on the values V. The attention can be computed as shown in Equation (15).

$$Attention(Q, K, V) = f_{\text{softmax}}(\frac{QK^T}{\sqrt{d_k}})V. \tag{15}$$

The proposed CAF is different from the traditional attention mechanism. It treats the output of the dual-stream network, Q_1 ($Q_1 = x_{\gamma\_w} \in \mathbb{R}^{B_s \times D_1}$) and Q_2 ($Q_2 = x_{\beta\_w} \in \mathbb{R}^{B_s \times D_2}$), as the module input. D_1 and D_2 indicate the feature length or dimensionality of one encoded audio sample in Q_1 and Q_2, respectively. Using full connection layers (FC_{k_1} and FC_{k_2}), x_{γ_w} and x_{β_w} are mapped to obtain the matrices K_1 and K_2 as

$$\begin{aligned} K_1 &= f_{FC_{k_1}}(x_{\gamma\_w}) \\ K_2 &= f_{FC_{k_2}}(x_{\beta\_w}) \end{aligned} \tag{16}$$

where $K_1 \in \mathbb{R}^{B_s \times D_3}$ and $K_2 \in \mathbb{R}^{B_s \times D_3}$, and thereby, the features are embedded and aligned with the same dimensionality D_3.

Then, x_γ and x_β are spliced. The matrix V is obtained after feature transformation of the splicing vector using a full-connection layer FC_V,

$$V = f_{FC_v}([x_{\gamma\_w}; x_{\beta\_w}]), \tag{17}$$

where $V \in \mathbb{R}^{B_s \times D_3}$.

After that, we extract the weights $W_{\mathcal{FC}_v}$ ($W_{\mathcal{FC}_v} \in \mathbb{R}^{D_3 \times (D_1+D_2)}$) and the bias $b_{\mathcal{FC}_v}$ ($b_{\mathcal{FC}_v} \in \mathbb{R}^{D_3}$) of the feature transformation network FC_v, and split them in alignment with the dimensions of x_{γ_w} and x_{β_w}:

$$\begin{aligned} W_{\mathcal{FC}_v} &= [W^{\gamma\_w}_{\mathcal{FC}_v}; W^{\beta\_w}_{\mathcal{FC}_v}] \\ b_{\mathcal{FC}_v} &= [b^{\gamma\_w}_{\mathcal{FC}_v}; b^{\beta\_w}_{\mathcal{FC}_v}] \end{aligned} \tag{18}$$

where $W^{\gamma\_w}_{\mathcal{FC}_v} \in \mathbb{R}^{D_3 \times D_1}$, $W^{\beta\_w}_{\mathcal{FC}_v} \in \mathbb{R}^{D_3 \times D_2}$, $b^{\gamma\_w}_{\mathcal{FC}_v} \in \mathbb{R}^{D_3}$, and $b^{\beta\_w}_{\mathcal{FC}_v} \in \mathbb{R}^{D_3}$.

Next, we calculate the cross-attention score by using a new attention function, which is defined as

$$f_{C_{Attention}}(Q_i, K_i, W^j_{\mathcal{FC}_v}, b^j_{\mathcal{FC}_v}) = ((Q_i \otimes (W^j_{\mathcal{FC}_v})^T) \oplus b^j_{\mathcal{FC}_v}) \oplus K_i, \tag{19}$$

where $\otimes$ denotes matrix multiplication. The final fusion feature is formulated as

$$O_{ut} = \sum_{(i,j)}^{\{(1,\gamma\_w),(2,\beta\_w)\}} f_{C_{Attention}}(Q_i, K_i, W^j_{\mathcal{FC}_v}, b^j_{\mathcal{FC}_v}), \tag{20}$$

where $O_{ut} \in \mathbb{R}^{B_s \times D_3}$. Consequently, the outputs from the top $(1, \gamma\_w)$ and bottom $(2, \beta\_w)$ branches are summed to generate the new features.

Generally, the workflow includes feature mapping (Equation (16)), feature splicing (Equation (17)), splitting for dimension alignment (Equation (18)), cross-attention score computing (Equation (19)), and feature fusion (Equation (20)). Finally, the procedure of the CAF module generates new features as the intermediate output ($f_{CAF}(x_{\gamma\_w}, x_{\beta\_w}) = O_{ut}$) in Equation (14) for final emotion prediction.

4. Materials and Methods

This section introduces the three databases, experiment design and dataset division, metrics for performance evaluation and comparison, and the implementation details of the proposed SER framework.

4.1. Databases

The first database, the Berlin Database of Emotional Speech (EmoDB) [54], is a German emotional speech database. It was created by the Institute of Communication Sciences of the Technical University of Berlin in Germany. The database has been available since 2005. It involves 10 professional speakers (5 females and 5 males) in speech recording. Each speaker produced five short and five long sentences in an anechoic chamber. The sentences or sessions are used in daily communication. High-quality equipment was used to record the utterances in addition to the sound electro-glottograms. The database contains around 800 German utterances in 7 categories of emotions ("happy", "angry", "anxious", "fearful", "bored", "disgusted", and "neutral").

The second database is the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [55]. It was created by the Speech Analysis and Interpretation Laboratory at the University of South California. The database has been accessible since 2008. It was recorded by 10 actors (5 males and 5 females) during five dyadic sessions. The database contains $\approx$ twelve hours of recordings, and a segment labeled by three to four human evaluators lasts for three to fifteen seconds. Notably, each dialog in the database contains audio, transcriptions, video, and motion-capture recordings. The recorded dialogues have been segmented into utterances and labeled as ten categories. In this study, the 5531 audio files of 4 categories of emotions ("happy + excited", "sad", "neutral", and "angry") are used.

The last database, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [56], has been available online since 2018. The database involves 24 professional actors (12 females and 12 males) speaking two lexically matched sentences in a neutral

North American accent. The sentences or the two sessions are "Kids are talking by the door" and "Dogs are sitting by the door". It consists of more than seven thousand audio and video clips. The database is provided in three modalities of audio-only, audio–video, and video-only (no sound) in two types (speech and song). The audio files are in lossless wave format, and each file lasts for about three seconds. This study was conducted on the speech-type data (2452 files) with 8 categories of emotions ("neutral", "calm", "happiness", "sadness", "anger", "fear", "disgust", and "surprise").

4.2. Experiment Design

Extensive experiments were conducted by using single- and dual-stream representation with different feature fusion modules. In detail, we trained the framework using single-stream representation of log-Mel spectrograms or audio signals with different networks (full training of TextRCNN [52] or DPCNN [50], or fine-tuning of HuBERT [34] or Wav2Vec2 [36]). Then, we carried out experiments with dual-stream representations of log-Mel spectrograms and speech audio signals in combination with different fusion modules. Finally, literature screening was used for summarizing recent unimodal SER achievement on these databases.

4.3. Performance Evaluation

Three metrics were employed for performance evaluation. Weighted accuracy (WA) is the overall accuracy of all samples ($WA = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} N_i}$), unweighted accuracy (UA) is the average accuracy of each emotion category ($UA = \frac{\sum_{i=1}^{k} n_i/N_i}{k}$), and F1-score ($F1S = \frac{2PR}{P+R}$) is the harmonic mean of precision (P) and recall (R) that gives more weight to the lower value. N_i denotes the number of utterances in the i-th class, n_i is the number of correctly classified utterances in i-th class, and k means the number of emotion classes. The metric values of WA, UA, and F1S range from 0 to 1, and a higher value indicates better recognition performance.

4.4. Implementation Details

Speech audio signals are sampled at 16 kHz at the utterance level without multi-timescale segment analysis. For the log-Mel spectrogram images, a Hanning window with a hop length of 512 ms and a discrete Fourier transform of length 2048 are applied. A log-Mel spectrogram image is presented with an $[h, w]$ matrix for each signal, and $h = w = 224$.

For encoder A, the size of the input matrix is $[b, c, h, w]$, $c = 1$, and the batch size is $b = 64$. The convolution kernel size of DPCNN is modified to 3×224, and its other parameters remain unchanged. As for TextRCNN, for model computing compatibility, the input matrix size is squeezed to $[b, h, w]$. In addition, the input dimension of the Bi-LSTM is set to 768, the number of hidden units is 384, and the rest of the parameters are set to default values. The other parameters are kept the same.

For encoder B, HuBERT and Wav2Vec2 have been pre-trained on LibriSpeech [57], and the models were downloaded from HuggingFace [58]. The codes are implemented with PyTorch (version 2.0.0). The optimizer is AdamW [59], and all the hyperparameters are fine-tuned with the learning rate 10^{-4}.

Since different numbers of actors are recorded in the databases (10 speakers in EmoDB, 10 speakers in IEMOCAP, and 24 speakers in RAVDESS), we use the leave-one-speaker-out cross-validation technique for verifying the effectiveness of the proposed SER framework. The technique offers a rigorous evaluation of SER performance and ensures a fair assessment of the model's generalization ability to new speakers.

The proposed framework is implemented with PyTorch (version 2.0.0). The codes are run on a Kaggle Notebook with a 16 GB GPU card (NVIDIA Tesla P100). Our codes are available on GitHub (https://github.com/NicoYuCN/SERcaf, accessed on 29 May 2024).

5. Results

This section reports the results of an ablation study when using different single- and dual-stream representation encoders as well as the current achievements on the three databases (EmoDB, IEMOCAP, and RAVDESS). It should be noted that in the ablation studies, the highest metric values of the SER frameworks are presented in boldface, while the metric values of the top-tier works are underlined to highlight recent achievements in the databases.

5.1. On the EmoDB Database

Table 1 shows the ablation study on the EmoDB database. It is found that fine-tuned HuBERT leads to the best prediction (WA, 0.8692; UA, 0.8806; F1S, 0.8804), followed by Wav2Vec2, when using the single-stream representation. Specifically, HuBERT outperforms TextRCNN by a large margin (WA, $\approx$ 0.28 $\uparrow$; UA, $\approx$ 0.34 $\uparrow$; F1S, $\approx$ 0.32 $\uparrow$).

Table 1. Ablation study on the EmoDB database.

		WA	UA	F1S
TextRCNN (full training)		0.5888	0.5401	0.5640
DPCNN (full training)		0.6449	0.6477	0.6610
HuBERT (fine-tuning)		**0.8692**	**0.8806**	**0.8804**
Wav2Vec2 (fine-tuning)		0.8598	0.8634	0.8687
HuBERT + DPCNN	CON	0.9252	0.9177	0.9198
	SUM	0.9439	0.9333	0.9446
	FiLM	0.8785	0.8890	0.8823
	CAF	**0.9720**	**0.9721**	**0.9764**
HuBERT + TextRCNN	CON	0.9065	0.9037	0.9115
	SUM	0.9159	0.9151	0.9205
	FiLM	0.8037	0.8184	0.8307
	CAF	**0.9259**	**0.9253**	**0.9264**
Wav2Vec2 + DPCNN	CON	0.9065	0.9047	0.9184
	SUM	0.9252	0.9136	0.9250
	FiLM	0.8224	0.8206	0.8363
	CAF	**0.9346**	**0.9317**	**0.9385**
Wav2Vec2 + TextRCNN	CON	0.8879	0.8597	0.8713
	SUM	0.8318	0.8161	0.8289
	FiLM	0.8598	0.8424	0.8536
	CAF	**0.9159**	**0.8947**	**0.8971**

When using dual-stream representation, CAF obtains higher metric values than the other three modules regardless of the encoder. This leads to the best result when HuBERT + DPCNN performs the emotion encoder (WA, 0.9720; UA, 0.9721; F1S, 0.9764). Using FiLM leads to the lowest values, while its result is superior to the best single-stream model.

Compared to the best single-stream representation framework (HuBERT), dual-stream representation (HuBERT + DPCNN) enriches emotion encoding and leads to obvious improvement when CAF is used for feature fusion. The metric values are increased from 0.8692 to 0.9720 (WA, $\approx$ 0.10 $\uparrow$), from 0.8806 to 0.9721 (UA, $\approx$ 0.09 $\uparrow$), and from 0.8804 to 0.9764 (F1S, $\approx$ 0.10 $\uparrow$).

Table 2 lists representative works on the EmoDB database. The proposed framework obtains the state-of-the-art performance, comparable to [46], which is followed by the work in [60–62]. Specifically, Ref. [62] presents speech signals with MFCC features, learns multi-scale contextual affective representation, and uses dynamic feature fusion for emotion

classification. Ref. [61] extracts various kinds of features, such as MFCCs, chromagrams, Mel-scale spectrograms, and spectral contrast features, and uses 1D CNN for emotion identification. Ref. [46] proposes dual-stream CNN-Transformer fusion and explores both spatial and temporal cues from MFCCs and Mel spectrogram features. Ref. [60] designs dual-stream representation from different kinds of multi-level wavelet coefficients, and discriminative features are selected using iterative neighborhood component analysis for emotion prediction.

Table 2. Current achievements on the EmoDB database.

	WA (%)	UA (%)	Year
[63]	84.49	83.31	2020
[64]		86.10	2020
[65]		69.72	2020
[66]		88.78	2021
[60]	90.09	89.47	2021
[61]		92.51	2022
[62]	95.70	95.17	2023
[46]	97.6	97.5	2023
[67]	87.89	87.14	2024
(HuBERT + DPCNN) + CAF	97.20	97.21	2024

5.2. On the IEMOCAP Database

Table 3 shows the ablation study on the IEMOCAP database. Using a single-stream representation, fine-tuned Wav2Vec2 obtains the best result (WA, 0.6269; UA, 0.6561; F1S, 0.6414), followed by fine-tuned HuBERT. Specifically, fine-tuned Wav2Vec2 leads to better results than fully trained TextRCNN (WA, $\approx$ 0.09 $\uparrow$; UA, $\approx$ 0.10 $\uparrow$; F1S, $\approx$ 0.09 $\uparrow$).

Table 3. Ablation study on the IEMOCAP database.

		WA	UA	F1S
TextRCNN (full training)		0.5375	0.5502	0.5504
DPCNN (full training)		0.5537	0.5718	0.5635
HuBERT (fine-tuning)		0.6161	0.6283	0.6254
Wav2Vec2 (fine-tuning)		**0.6269**	**0.6561**	**0.6414**
HuBERT + DPCNN	CON	0.6061	0.6318	0.6194
	SUM	0.5935	0.6098	0.6133
	FiLM	0.5944	0.5950	0.6092
	CAF	**0.6242**	**0.6442**	**0.6512**
HuBERT + TextRCNN	CON	0.6161	0.6268	0.6257
	SUM	0.6378	0.6490	0.6544
	FiLM	0.5673	0.5640	0.5797
	CAF	**0.6965**	**0.7088**	**0.7084**
Wav2Vec2 + DPCNN	CON	0.6070	0.6123	0.6225
	SUM	0.6025	0.6102	0.6085
	FiLM	0.5772	0.5865	0.5886
	CAF	**0.6242**	**0.6348**	**0.6377**
Wav2Vec2 + TextRCNN	CON	0.5891	0.6235	0.6234
	SUM	0.6079	0.6248	0.6211
	FiLM	0.5763	0.6078	0.5964
	CAF	**0.6405**	**0.6523**	**0.6540**

When using dual-stream encoders, CAF achieves better performance than the other modules. When HuBERT + TextRCNN performs the emotion representation, CAF leads to the best result (WA, 0.9720; UA, 0.9721; F1S, 0.9764).

In comparison to the best single-stream model (Wav2Vec2), dual-stream representation (HuBERT + TextRCNN) enriches emotion encoding, and good improvement is obtained with the CAF for emotion recognition. The values are increased by up to 11.10%, 8.03%, and 10.45% on WA, UA, and F1S, respectively.

Table 4 shows the achievements on the IEMOCAP database. The proposed framework achieves competitive performance compared to several studies, such as [19,47,62,68], while their results are obviously inferior to [69]. Among the works, Ref. [69] uses multi-scale MFCC features and develops a multi-unit attention module for re-weighting the features in terms of time, frequency, and channel dimensions; Ref. [19] designs a dual-stream representation from pre-trained and spectrum-based encoders and proposes an interactive attention module for fusing the intermediate features; Ref. [47] extracts tri-stream acoustic information of MFCCs, spectrograms, and high-level information; and [68] fuses static and dynamic features in a hierarchical network using three different modules.

Table 4. Current achievements on the IEMOCAP database.

	WA (%)	UA (%)	Year
[63]	66.92	64.51	2020
[64]		64.30	2020
[68]	70.50	72.50	2021
[37]	65.4	66.9	2021
[66]		70.46	2021
[47]	71.64	72.70	2022
[61]		66.64	2022
[69]	82.46	80.45	2022
[62]	71.65	72.50	2023
[70]	70.50	71.51	2023
[19]	72.48	73.32	2023
[41]	68.85	71.89	2024
(HuBERT + TextRCNN) + CAF	69.65	70.88	2024

5.3. On the RAVDESS Database

Table 5 shows the ablation study on the RAVDESS database. Fine-tuned HuBERT is the best single-stream method (WA, 0.7351; UA, 0.7341; F1S, 0.7324), and fine-tuned LLMs lead to superior results compared to fully trained models.

Using the dual-stream encoder HuBERT + DPCNN, CAF achieves the best performance (WA, 0.8186; UA, 0.8275; F1S, 0.8284). It leads to higher values than the other three fusion modules. However, 10 out of the 16 dual-stream representation frameworks cause worse results than the single-stream model of fine-tuned HuBERT.

Compared to the best single-stream representation framework of fine-tuned HuBERT, the dual-stream representation HuBERT + DPCNN enriches the emotion encoding and leads to obvious improvement when CAF performs feature fusion. The metric values are increased from 0.7351 to 0.8186 (WA, $\approx 0.08 \uparrow$), from 0.7341 to 0.8275 (UA, $\approx 0.09 \uparrow$), and from 0.7324 to 0.8284 (F1S, $\approx 0.10 \uparrow$).

Table 6 lists representative works on the RAVDESS database. It is found that the studies in [46,62] obtain the best performance, followed by [40,60,69], and our work. Notably, Ref. [40] uses pre-trained Wav2Vec2 without parameter fine-tuning; Ref. [62] is a temporal-aware bidirectional multi-scale network with single-stream temporal emotional modeling and its performance surpasses the second-best network [60], with an improvement of more than 4.5%.

Table 5. Ablation study on the RAVDESS database.

		WA	UA	F1S
TextRCNN (full training)		0.5661	0.5614	0.5604
DPCNN (full training)		0.6562	0.6484	0.6638
HuBERT (fine-tuning)		**0.7351**	**0.7341**	**0.7324**
Wav2Vec2 (fine-tuning)		0.7118	0.7158	0.7066
HuBERT + DPCNN	CON	0.7292	0.7294	0.7219
	SUM	0.7708	0.7676	0.7638
	FiLM	0.7257	0.7169	0.7277
	CAF	**0.8186**	**0.8275**	**0.8284**
HuBERT + TextRCNN	CON	0.6667	0.6754	0.6816
	SUM	0.6841	0.6634	0.6635
	FiLM	0.7604	0.7522	0.7545
	CAF	**0.7812**	**0.7881**	**0.7868**
Wav2Vec2 + DPCNN	CON	0.7118	0.7151	0.7075
	SUM	0.7083	0.6913	0.7082
	FiLM	0.7014	0.6912	0.6923
	CAF	**0.7674**	**0.7729**	**0.7717**
Wav2Vec2 + TextRCNN	CON	0.7188	0.7191	0.7165
	SUM	0.7083	0.7024	0.7014
	FiLM	0.6528	0.6551	0.6572
	CAF	**0.7292**	**0.7229**	**0.7198**

Table 6. Current achievement on the RAVDESS database.

	WA (%)	UA (%)	Year
Deep-CNN [64]		71.61	2020
CNN-LSTM [65]		53.08	2020
QCNN [66]		77.87	2021
TSP-INCA [60]	87.43	87.43	2021
Wav2Vec2-PT [40]		84.3	2021
MLAnet [69]	84.48	83.55	2022
TIM-Net [62]	92.08	91.93	2023
DS-CTFN [46]	97.7	97.6	2023
NWS graph [67]	76.45	75.79	2024
(HuBERT + DPCNN) + CAF	81.86	82.75	2024

6. Discussion

A novel SER framework is developed for addressing the challenges in emotion encoding and feature fusion of speech signals. It embraces dual-stream representation for comprehensive quantification of emotional cues and designs a CAF module for effective integration of encoded features (Figure 1). Specifically, the dual-stream encoder employs a fully trained network and a pre-trained network, and the CAF merges the encoder output (Figure 2) for recognizing subjective emotions.

The effectiveness of the proposed SER framework has been verified by its promising performance. On EmoDB, it achieves a top-tier result, even if it is slightly inferior to the best one (Table 2). On IEMOCAP, the framework is competitive with the state-of-the-art works (Table 4). On RAVDESS, there is still room for improvement when using the proposed model (Table 6). On the databases, screening of the literatur suggests that Refs. [46,62] remains the best, followed by the proposed framework. Two other notable networks are proposed in [60,69]. The former employs multi-level wavelet decomposition and coefficient selection, and the latter uses domain-specific features and multi-unit attention-based reweighting. Interestingly, both [62,69] present single-stream representations and multi-scale MFCC features are analyzed, while the proposed network and [46,60] develop dual-stream representations. The comparison result inspires us that combining comprehensive analysis and dual-stream representation could further improve the performance.

The ablation study suggests that CAF outperforms the other three fusion modules (Tables 1, 3 and 5). It should be admitted that "SUM" and "CON" are widely used. However, significant inconsistency exists in the scales and semantics of feature representation. For instance, in the current study, dual-stream output features from audio signals and spectrogram images express different kinds of embedding information (Figure 1). To address this problem, FiLM [49] develops a feature-wise affine transformation layer to influence network computation; however, it seems to fail in feature integration, as shown by its inferior performance. Different from FiLM, the proposed CAF employs matrix transformation to tackle the inconsistency by representing features in a unified space, and the weights and biases of the transformation matrix are split, aligned, and then, added according to the dimensions of the input features. Its superiority has been verified in the ablation study and the literature screening study, while its generalization on dual-stream or dual-modality feature fusion requires further investigation.

Dual-stream encoding combined with the CAF module achieves improved recognition performance compared to single-stream representation (Tables 1, 3 and 5). In addition to the importance of the CAF module, dual-stream representation is helpful for embedding emotional cues. Among single-stream representations, fine-tuned LLMs (HuBERT and Wav2Vec2) are consistently better than fully trained text processing networks (TextRCNN and DPCNN). The reasons are manifold. Firstly, LLMs have been pre-trained with the corpus LibriSpeech [57] which consists of around 1000 h of speech audio, 803 million tokens, and 900,000 distinct words. Secondly, LLMs embedded with advanced modules and novel loss functions and training strategies have been verified to have good generalization on many affective computing tasks [5], even if trained with a relatively small number of data samples towards a specific downstream task. In addition, TextRCNN and DPCNN are primarily designed for text processing, and text is a modality different from audio signals. In the current study, their variants are dedicated to finding emotional cues in audio signals, while their encoded features seem not to be sufficient for expressing subjective emotion patterns. Therefore, more attention could be paid to the modification of text processing models in our future work.

There are several limitations in the current study. Firstly, no additional studies are conducted to select the most promising encoders. In the dual-stream encoders, many CNNs [71] could be used to replace text processing networks for extracting in-depth features, and many cutting-edge models, such as GPT [43,72] and emotional text-to-speech models [41], could be employed for improved feature embedding of the audio signals. Secondly, in addition to FiLM [49], more dual- or multi-stream fusion modules [19,45,47,60] could be investigated for broadening our understanding on proper feature integration. Thirdly, how to cut the speech audio signals into segments remains an unsettled problem [73]. To find the best way indeed requires empirical experience and massive experiments, while it is meaningful in the field of speech signal analysis. Fourthly, speech emotion can be represented in many kinds of representations, including but not limited to audio signals, MFCC features, and spectrogram images [47,48,73]. For improving the performance in emotion recognition and affective computing, multi-stream networks combined with multi-branch fusion modules might be promising. In addition, different data sample augmentation methods, such as random mixing, adversarial training, transfer learning, and curriculum learning, could be applied to improve the training sample number and to enrich the data diversity [41]. Last but not least, ensembles of existing SER models would be beneficial to further enhance performance, although how to properly combine these models remains challenging [74].

7. Conclusions

Speech emotion recognition remains challenging. In this study, to unearth the emotional cues in audio signals, a dual-stream encoder is designed. One stream uses full training of a text processing network to encode the log-Mel spectrogram images, and the other applies fine-tuning of a pre-trained language model. To avoid conflicts during feature

integration, a cross-attention fusion module is proposed. Extensive experiments on three databases have verified the effectiveness of the dual-stream encoder, the fusion module, and the framework for speech emotion prediction. For further enhancing the performance, it could be promising to combine multi-stream networks and multi-branch fusion modules.

Author Contributions: Conceptualization, S.Y. and J.M.; data curation, J.M. and Y.C.; formal analysis, W.F., B.Z. and Y.X.; funding acquisition, W.F., Y.X. and Q.S.; investigation, S.Y., B.Z., H.Y. and Y.X.; methodology, S.Y., J.M., Y.C., B.Z. and H.Y.; project administration, Q.S.; software, J.M., Y.C. and B.Z.; supervision, Q.S.; validation, W.F. and Y.C.; visualization, S.Y., J.M. and Y.C.; writing—original draft, S.Y. and J.M.; writing—review and editing, W.F., H.Y., Y.X. and Q.S. All authors have read and agreed to the published version of the manuscript.

Funding: The work was in part supported by the National Key Research and Develop Program of China (Grant No. 2022ZD0115901, and 2022YFC2409000), the National Natural Science Foundation of China (Grant No. 62177007, and U20A20373), the China-Central Eastern European Countries High Education Joint Education Project (Grant No. 202012), the Shenzhen Science and Technology Program (Grant No. KQTD20180411185028798), and the Medium- and Long-term Technology Plan for Radio, Television and Online Audiovisual (Grant No. ZG23011). The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability Statement: The dataset supporting the current study is available online (EmoDB, http://emodb.bilderbar.info/start.html, accessed on 29 May 2024 ; IEMOCAP, https://sail.usc.edu/iemocap/, accessed on 29 May 2024; RAVDESS, https://zenodo.org/records/1188976, accessed on 29 May 2024).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

SER	Speech emotion recognition
CAF	Cross-attention fusion
EEG	Electroencephalogram
MFCCs	Mel-frequency cepstral coefficients
RNNs	Recurrent neural networks
LLMs	Large language models
BERT	Bidirectional encoder representations from Transformers
HuBERT	Hidden-unit BERT
Wav2Vec	Wave to vector
Wav2Vec2	Wave to vector 2.0
TextRCNN	Recurrent convolutional neural network for text classification
DPCNN	Deep pyramid convolutional neural networks
EmoDB	Berlin Database of Emotional Speech
IEMOCAP	Interactive Emotional Dyadic Motion Capture
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
SUM	Summarization
CON	Concatenation
FiLM	Feature-wise linear modulation
WA	Weighted accuracy
UA	Unweighted accuracy
F1S	F1-score
CNN	Convolutional neural network
Bi-LSTM	Bidirectional long short-term memory

References

1. Kheddar, H.; Hemis, M.; Himeur, Y. Automatic speech recognition using advanced deep learning approaches: A survey. *Inf. Fusion* **2024**, *109*, 102422. [CrossRef]
2. Houssein, E.H.; Hammad, A.; Ali, A.A. Human emotion recognition from EEG-based brain-computer interface using machine learning: A comprehensive review. *Neural Comput. Appl.* **2022**, *34*, 12527–12557. [CrossRef]
3. Wu, H.; Xu, H.; Seng, K.P.; Chen, J.; Ang, L.M. Energy efficient graph-based hybrid learning for speech emotion recognition on humanoid robot. *Electronics* **2024**, *13*, 1151. [CrossRef]
4. Tan, L.; Yu, K.; Lin, L.; Cheng, X.; Srivastava, G.; Lin, J.C.-W.; Wei, W. Speech emotion recognition enhanced traffic efficiency solution for autonomous vehicles in a 5G-enabled space-air-ground integrated intelligent transportation system. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 2830–2842. [CrossRef]
5. Wang, Y.; Song, W.; Tao, W.; Liotta, A.; Yang, D.; Li, X.; Gao, S.; Sun, Y.; Ge, W.; Zhang, W.; et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Inf. Fusion* **2022**, *83*, 19–52. [CrossRef]
6. Egger, M.; Ley, M.; Hanke, S. Emotion recognition from physiological signal analysis: A review. *Electron. Notes Theor. Comput. Sci.* **2019**, *343*, 35–55. [CrossRef]
7. Wang, K.; Wu, Q.; Song, L.; Yang, Z.; Wu, W.; Qian, C.; He, R.; Qiao, Y.; Loy, C.C. MEAD: A large-scale audio-visual dataset for emotional talking-face generation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 700–717.
8. 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]
9. Hashem, A.; Arif, M.; Alghamdi, M. Speech emotion recognition approaches: A systematic review. *Speech Sommunication* **2023**, *154*, 102974. [CrossRef]
10. Zhang, T.; Tan, Z. Survey of deep emotion recognition in dynamic data using facial, speech and textual cues. *Multimed. Tools Appl.* **2024**, 1–40. [CrossRef]
11. Liu, F.; Yang, P.; Shu, Y.; Yan, F.; Zhang, G.; Liu, Y.J. Emotion dictionary learning with modality attentions for mixed emotion exploration. *IEEE Trans. Affect. Comput.* **2023**, 1–15. [CrossRef]
12. Tan, Y.; Sun, Z.; Duan, F.; Solé-Casals, J.; Caiafa, C.F. A multimodal emotion recognition method based on facial expressions and electroencephalography. *Biomed. Signal Process. Control.* **2021**, *70*, 103029. [CrossRef]
13. Liu, W.; Qiu, J.-L.; Zheng, W.-L.; Lu, B.L. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Trans. Cogn. Dev. Syst.* **2021**, *14*, 715–729. [CrossRef]
14. Atmaja, B.T.; Sasou, A.; Akagi, M. Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. *Speech Commun.* **2022**, *140*, 11–28. [CrossRef]
15. Khan, M.; Gueaieb, W.; El Saddik, A.; Kwon, S. MSER: Multimodal speech emotion recognition using cross-attention with deep fusion. *Expert Syst. Appl.* **2024**, *245*, 122946. [CrossRef]
16. Tang, Y.; Hu, Y.; He, L.; Huang, H. A bimodal network based on audio-text interactional-attention with arcface loss for speech emotion recognition. *Speech Commun.* **2022**, *143*, 21–32. [CrossRef]
17. Zhang, J.; Liu, Z.; Liu, P.; Wu, B. Dual-waveform emotion recognition model for conversations. In Proceedings of the IEEE International Conference on Multimedia and Expo, Shenzhen, China, 5–9 July 2021; pp. 1–6.
18. Tripathi, S.; Beigi, H. Multi-modal emotion recognition on IEMOCAP with neural networks. *arXiv* **2018**, arXiv:1804.05788.
19. Hu, Y.; Hou, S.; Yang, H.; Huang, H.; He, L. A joint network based on interactive attention for speech emotion recognition. In Proceedings of the IEEE International Conference on Multimedia and Expo, Brisbane, Australia, 10–14 July 2023; pp. 1715–1720.
20. Schuller, B.W. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. Acm* **2018**, *61*, 90–99. [CrossRef]
21. Zhang, Z.; Liang, X.; Qin, W.; Yu, S.; Xie, Y. matFR: A MATLAB toolbox for feature ranking. *Bioinformatics* **2020**, *36*, 4968–4969. [CrossRef]
22. Zhang, X.; Xiao, H. Enhancing speech emotion recognition with the improved weighted average support vector method. *Biomed. Signal Process. Control.* **2024**, *93*, 106140. [CrossRef]
23. Guizzo, E.; Weyde, T.; Leveson, J.B. Multi-time-scale convolution for emotion recognition from speech audio signals. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020; pp. 6489–6493.
24. Sha, N.; Yang, W.; Wei, F.; Lu, Z.; Chen, M.; Ma, C.; Zhang, L.; Shi, H. Speech emotion recognition using RA-GMLP model on time-frequency domain features extracted by TFCM. *Electronics* **2024**, *13*, 588. [CrossRef]
25. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput.* **2015**, *7*, 190–202. [CrossRef]
26. Eyben, F.; Wollmer, M.; Schuller, B. OpenSMILE: The munich versatile and fast open-source audio feature extractor. In Proceedings of the ACM International Conference on Multimedia, Firenze Italy, 25–29 October 2010; pp. 1459–1462.
27. Rouast, P.V.; Adam, M.T.; Chiong, R. Deep learning for human affect recognition: Insights and new developments. *IEEE Trans. Affect. Comput.* **2019**, *12*, 524–543. [CrossRef]

28. Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic speech emotion recognition using recurrent neural networks with local attention. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, LA, USA, 5–9 March 2017; pp. 2227–2231.
29. Liu, M.; Raj, A.N.J.; Rajangam, V.; Ma, K.; Zhuang, Z.; Zhuang, S. Multiscale-multichannel feature extraction and classification through one-dimensional convolutional neural network for Speech emotion recognition. *Speech Commun.* **2024**, *156*, 103010. [CrossRef]
30. Zhang, S.; Zhang, S.; Huang, T.; Gao, W. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Trans. Multimed.* **2017**, *20*, 1576–1590. [CrossRef]
31. Lu, Z.; Cao, L.; Zhang, Y.; Chiu, C.C.; Fan, J. Speech sentiment analysis via pre-trained features from end-to-end ASR models. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020; pp. 7149–7153.
32. Kim, E.; Shin, J.W. DNN-based emotion recognition based on bottleneck acoustic features and lexical features. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12–17 May 2019; pp. 6720–6724.
33. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional Transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
34. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460. [CrossRef]
35. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. wav2vec: Unsupervised pretraining for speech recognition. *arXiv* **2019**, arXiv:1904.05862.
36. 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.
37. Xia, Y.; Chen, L.W.; Rudnicky, A.; Stern, R.M.; Temporal context in speech emotion recognition. *Interspeech* **2021**, 3370–3374.
38. Chen, L.W.; Rudnicky, A. Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
39. Sun, C.; Zhou, Y.; Huang, X.; Yang, J.; Hou, X. Combining wav2vec 2.0 fine-tuning and ConLearnNet for speech emotion recognition. *Electronics* **2024**, *13*, 1103. [CrossRef]
40. Pepino, L.; Riera, P.; Ferrer, L. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv* **2021**, arXiv:2104.03502.
41. Ma, Z.; Wu, W.; Zheng, Z.; Guo, Y.; Chen, Q.; Zhang, S.; Chen, X. Leveraging speech PTM, text LLM, and emotional TTS for speech emotion recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Seoul, Korea, 14–19 April 2024; pp. 11146–11150.
42. Baevski, A.; Hsu, W.; Xu, Q.; Babu, A.; Gu, J.; Auli, M. Data2vec: A general framework for self-supervised learning in speech, vision and language. *Int. Conf. Mach. Learn.* **2022**, *162*, 1298–1312.
43. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S. GPT-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
44. Feng, T.; Narayanan, S. Foundation model assisted automatic speech emotion recognition: Transcribing, annotating, and augmenting. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Seoul, Republic of Korea, 14–19 April 2024; pp. 12116–12120.
45. Chen, M.; Zhao, X. A multi-scale fusion framework for bimodal speech emotion recognition. *Interspeech* **2020**, 374–378.. [CrossRef]
46. Tellai, M.; Gao, L.; Mao, Q. An efficient speech emotion recognition based on a dual-stream CNN-transformer fusion network. *Int. J. Speech Technol.* **2023**, *26*, 541–557. [CrossRef]
47. 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 IEEE International Conference on Acoustics, Speech and Signal Processing, Virtual, 7–13 May 2022; pp. 7367–7371.
48. Lope, J.; Graña, M. An ongoing review of speech emotion recognition. *Neurocomputing* **2023**, *528*, 1–11. [CrossRef]
49. Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; Courville, A. FiLM: Visual reasoning with a general conditioning layer. *Aaai Conf. Artif. Intell.* **2018**, *32*, 3942–3951. [CrossRef]
50. Johnson, R.; Zhang, T. Deep pyramid convolutional neural networks for text categorization. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 562–570.
51. Habimana, O.; Li, Y.; Li, R.; Gu, X.; Yu, G. Sentiment analysis using deep learning approaches: An overview. *Sci. China Inf. Sci.* **2020**, *63*, 1–36. [CrossRef]
52. Lai, S.; Xu, L.; Liu, K.; Zhao, J. Recurrent convolutional neural networks for text classification. *Aaai Conf. Artif. Intell.* **2015**, *29*, 2267–2273. [CrossRef]
53. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3942–3951.
54. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. *Interspeech* **2005**, *5*, 1517–1520.
55. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]

56. Livingstone, S.R.; Russo, F.A. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLoS ONE* **2018**, *13*, 0196391. [CrossRef] [PubMed]
57. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, South Brisbane, QLD, Australia, 19–24 April 2015; pp. 5206–5210.
58. Yang, S.W.; Chi, P.H.; Chuang, Y.S.; Lai, C.I.J.; Lakhotia, K.; Lin, Y.Y.; Liu, A.T.; Shi, J.; Chang, X.; Lin, G.T.; et al. Superb: Speech processing universal performance benchmark. *arXiv* **2021**, arXiv:2105.01051.
59. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
60. Tuncer, T.; Dogan, S.; Acharya, U.R. Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques. *Knowl.-Based Syst.* **2021**, *211*, 106547. [CrossRef]
61. Hou, M.; Zhang, Z.; Cao, Q.; Zhang, D.; Lu, G. Multi-view speech emotion recognition via collective relation construction. *IEEE ACM Trans. Audio, Speech, Lang. Process.* **2021**, *30*, 218–229. [CrossRef]
62. 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 IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
63. Yi, L.; Mak, M.W. Improving speech emotion recognition with adversarial data augmentation network. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *33*, 172–184. [CrossRef]
64. Issa, D.; Demirci, M.F.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control.* **2020**, *59*, 101894. [CrossRef]
65. Parry, J.; Palaz, D.; Clarke, G.; Lecomte, P.; Mead, R.; Berger, M.; Hofer, G. Analysis of deep learning architectures for cross-corpus speech emotion recognition. *Interspeech* **2019**, 1656–1660.
66. Muppidi, A.; Radfar, M. Speech emotion recognition using quaternion convolutional neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada, 6–11 June 2021; pp. 6309–6313.
67. Soltani, R.; Benmohamed, E.; Ltifi, H. Newman-Watts-Strogatz topology in deep echo state networks for speech emotion recognition. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108293. [CrossRef]
68. Cao, Q.; Hou, M.; Chen, B.; Zhang, Z.; Lu, G. Hierarchical network based on the fusion of static and dynamic features for speech emotion recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada, 13 May 2021; pp. 6334–6338.
69. Liu, K.; Wang, D.; Wu, D.; Liu, Y.; Feng, J. Speech emotion recognition via multilevel attention network. *IEEE Signal Process. Lett.* **2022**, *29*, 2278–2282. [CrossRef]
70. Chen, W.; Xing, X.; Xu, X.; Pang, J.; Du, L. Speechformer++: A hierarchical efficient framework for paralinguistic speech processing. *IEEE ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 775–788. [CrossRef]
71. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6999–7019. [CrossRef] [PubMed]
72. Zhang, T.; Meng, J.; Yang, Y.; Yu, S. Contrastive learning penalized cross-entropy with diversity contrastive search decoding for diagnostic report generation of reduced token repetition. *Appl. Sci.* **2021**, *14*, 2817. [CrossRef]
73. Zhu, B.; Li, X.; Feng, J.; Yu, S. VGGish-BiLSTM-attention for COVID-19 identification using cough sound analysis. In Proceedings of the International Conference on Signal and Image Processing, Wuxi, China, 8–10 July 2023; pp. 49–53.
74. López-Gil, J.; Garay-Vitori, N. Assessing the effectiveness of ensembles in speech emotion recognition: Performance analysis under challenging scenarios. *Expert Syst. Appl.* **2024**, *243*, 122905. [CrossRef]