applied sciences

MDPI

Article

Speech Sentiment Analysis Using Hierarchical Conformer Networks

Peng Zhao, Fangai Liu * and Xuqiang Zhuang

School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China
* Correspondence:lfa@sdnu.edu.cn

Abstract: Multimodality has been widely used for sentiment analysis tasks, especially for speech sentiment analysis. Compared with the emotion expression of most text languages, speech is more intuitive for human emotion, as speech contains more and richer emotion features. Most of the current studies mainly involve the extraction of speech features, but the accuracy and prediction rate of the models still need to be improved. To improve the extraction and fusion of speech sentiment feature information, we present a new framework. The framework adopts a hierarchical conformer model and an attention-based GRU model to increase the accuracy of the model. The method has two main parts: a local feature learning group and a global feature learning group. The local feature learning group is mainly used to learn the spatio-temporal feature information of speech emotion features through the conformer model, and a combination of convolution and transformer is used to be able to enhance the extraction of long and short-term feature information. The global features are then extracted by the AUGRU model, and the fusion of features is performed by the attention mechanism to access the weights of feature information. Finally, the sentiment is identified by a fully connected network layer, and then classified by a central loss function and a softmax function. Compared with existing speech sentiment analysis models, we obtained better sentiment classification results on the IEMOCAP and RAVDESS benchmark datasets.

Keywords: multimodal; conformer; attention; GRU

Citation: Zhao, P.; Liu, F.; Zhuang, X. Speech Sentiment Analysis Using Hierarchical Conformer Networks. *Appl. Sci.* **2022**, *12*, 8076. https://doi.org/10.3390/app12168076

Academic Editor: Byung-Gyu Kim

Received: 19 July 2022
Accepted: 9 August 2022
Published: 12 August 2022

1. Introduction

Sentiment analysis is one of the important research tasks of natural language processing, and speech sentiment analysis is a very important direction. The original sentiment sub-task mainly focuses on text aspect-level sentiment research. As the times go by, the popularity of short video social media such as Douyin and BILIBILI, people began to express their emotions through forms different from text. By extracting and learning features such as intonation and timbre of speech, we can obtain richer emotional information than simply analyzing text. Following the evolution of deep learning, Speech emotion analysis has been a vital part of HCI [1,2]. Bhardwaj et al. [3] proved the importance of the speech sentiment analysis task and also summarized the results of the current work and future trends.

In the last few years, sentiment analysis has been extensively adopted in different fields and has provided various new approaches. Duan et al. [4] proposed a semi-supervised generative sentiment-based model for sentiment analysis of stock investors. The approach is able to leverage features from training and testing information, taking into account information, sentiment and words. Considering the high computational cost of supervised learning methods, Fares et al. [5] investigated an unsupervised word-level knowledge mapping framework. Bibi et al. [6] presented a novel unsupervised learning framework based on concept and hierarchical clustering. These methods experimentally demonstrate that unsupervised learning resolves the computational cost problem while retaining the same techniques as supervised learning. Abboud et al. [7] performed sentiment analysis on musical compositions and constructed an algorithmic framework that can autonomously

compose and express musical emotions. Sentiment analysis has been well applied and broadened in the field of music. SUN et al. [8] studied the emotional polarity of specific targets, both in terms and sentiment analysis. To catch the multiple interactions of goals and aspects, they built a deep interactive memory network that also works well to form specific memories for different goals.

Some early speech sentiment analysis models only paid more attention to sentiment word information, ignoring the fine-grained feature information transmitted by speech [9]. Shaik et al. [10] presented a new frame to process the speech signal with short-term features to obtain the medium-term features, which are put into the sentiment lexicon for comparison to derive the sentiment lexicality. After the development of deep learning, most research works have paid more attention to the fine-grained features of speech. These models use Convolution Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory Networks (LSTM) or Deep Neural Networks (DNN). Most previous researchers introduced different deep neural network models to analyze emotion in speech modalities. For example, Kwon et al. [11] suggested a model that focuses more on the sentiment information of the raw data, and they sampled from the raw data as the input to the model. Jara-Vera et al. [12] instead used a specific representation of the audio recording as input to the model. Following the extensive expansion of the deep learning technology in the domain of speech sentiment analysis, researchers can extract higher-level features from speech signals and can obtain better accuracy and recognition rates compared to lower-level features [13].

Although deep learning models can better extract high-level features of sentiments, the computational cost and training time of the models also increase [14]. For example, convolution Neural Networks (CNN) have not significantly improved the accuracy and reduced complexity cost of feature processing of speech signals. The Recurrent Neural Network (RNN) solves this problem well, but the training cost is not effectively reduced and the calculation is more complicated. Today, most researchers employ frame representation and connectivity approaches for feature fusion, which only target specific tasks and do not have good generality. The sparsity of the data is also a difficult problem for feature extraction.

To increase the accuracy of speech emotion feature extraction and reduce the computational cost, we introduce a new architecture for speech sentiment analysis using speech sequences in this paper. As shown in Figure 1, the architecture of the framework has two main parts: a local feature learning module and a global feature learning module. First, the raw speech signal data are input into the local feature extraction module, and the spatial features of emotion are extracted by convolution through the conformer, and then the spatial and temporal features of emotion are learned through a hierarchical strategy. Second, the global feature extraction module adaptively adjusts the associated global feature weights based on the relevance of the input features. At last, sentiment classification is performed through the central loss function and the softmax function. The main innovations of our approach include the following:

(1) We present a framework for speech sentiment recognition using hierarchical conformer networks and an attention-based GRU model.
(2) The conformer model learns and extracts the spatio-temporal features of speech signals and the semantic associations of the sentiment information. The Gated Recursive Unit (GRU) of the attention-based network layer adaptively adjusts the weights of the computed global features. The weights of the global features are adjusted according to the local features and the global features.
(3) The conformer model can more accurately select the deep-level features of speech signals without increasing the computational cost. It solves the computational cost issues of traditional models and improves the accuracy significantly.

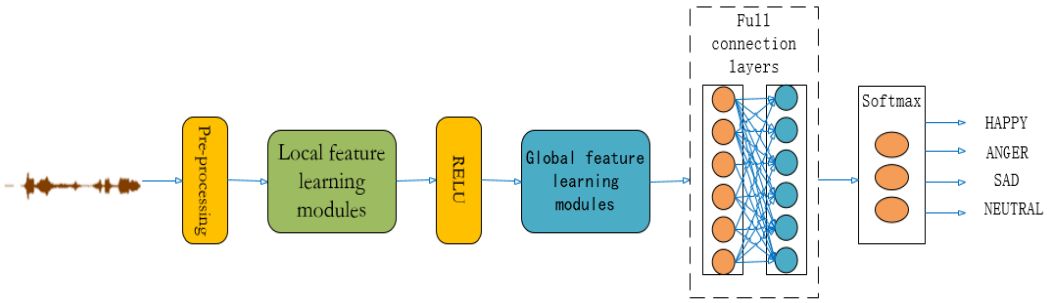

Figure 1. The overall model architecture proposed in this paper.

The other parts of this article work are as follows: Section 2 describes relevant recent work on sentiment analysis; Section 3 introduces our proposed model framework and methodology; Section 4 presents experimental evaluation and discussion; Section 5 reports experimental results and analysis; Section 6 summarizes the work and the future perspectives.

2. Related Works

Most of the recent research work has applied deep learning models for speech sentiment analysis tasks, such as CNNs, RNNs, LSTMs, and DNNs. Deep learning models can indeed learn deeper features in speech signals better than traditional machine learning models, and retain the original low-level features. For example, Kwon et al. [15] presented an AI-aided Deep Stochastic Convolutional Neural Network (DSCNN) architecture. The architecture uses a shared network approach to learn distinguished and discriminable features in the spectral map of the speech signal. The architecture has been enhanced for better implementation. Chernykh et al. [16] proposed a new algorithm for speech sentiment recognition. The algorithm uses the Deep Recurrent Neural Network (DSRNN) and a connectionist time-based classification method. The network is trained on a set of acoustic features computed by small speech intervals. The algorithm can take into account the non-emotional parts contained in emotional words and phrases, and can predict the emotional order of sentences. Kwon et al. [17] presented a unique AI-aided system architecture for Speech Emotion Analysis (SER). The architecture learns local characteristics of each layer by fusing convolution and long short-term memory networks, and then uses folded Gated Recursive Units (GRUs) to learn global features and adjust the weights. For model training, robustness and selection of important features is also one of the important challenges for speech sentiment analysis. In this era of deep learning, researchers have proposed and used various models and methods to improve and solve the difficult task of speech sentiment analysis. Koduru et al. [18] proposed a feature extraction algorithm. This algorithm mainly extracts the acoustic features of the speech signal such as pitch, energy, over-zero rate, Mel frequency cepstral coefficients and discrete wavelet transform to analyze the signal [19]. Chatziagapi et al. [20] used Generative Adversarial Network (GAN)-based data enhancement methods to improve recognition accuracy.

As we all know, audio segment processing in speech sentiment analysis task is a sequential task. There are not only complex spatial features in the audio signal, but also rich temporal features. The order of the audio signal is very important for the model to analyze its sentiment features. Since the development of deep learning, most researchers consider the Long Short-Term Memory (LSTM) network [21] as the perfect model to handle sequential modeling tasks. Considering the acoustic model and the effectiveness of feature extraction, Zhang et al. [22] proposed a novel heterogeneous parallel convolutional Bi-LSTM model. The model takes advantage of heterogeneous parallelism to learn spatio-temporal feature information, which can extract features more effectively. Further, the model can be trained by extracting features at the sentence level or frame level, thus reducing the computational cost. Atila et al. [23] suggested a novel deep architecture based on attention-integrated 3D CNN-LSTM. This architecture transforms the speech signal as MFCC into a speech spectrogram, and then processes the speech spectrogram into 3D data

as input by stacking multiple consecutive frames. The issues of missing low-dimensional features and weak spatio-temporal correlation are effectively solved, and the accuracy of sentiment analysis is improved. Senthilkumar et al. [24] combine a deep belief network and a bidirectional LSTM, and the suggested architecture aims to improve the realization of the affective status of speech by using sequence selection.

Although LSTM and its variants have advanced the process of sentiment analysis tasks, researchers soon found several drawbacks: there are disadvantages in parallel processing, the gradient problem of LSTM for RNNs is only somewhat alleviated, and LSTM is very time-consuming when computing deep networks with large periods. To resolve these problems, Parmar et al. [25] presented a new model transformer for image processing. First, the transformer uses the attention mechanism to effectively solve the long-term dependency problem. Second, it is not a sequential structure such as RNN, so it has good parallelism. Lian et al. [26] proposed the Conversational Transform Network (CT Net), which uses a transformer-based architecture to simulate both intramodal and cross-modal interactions between multimodal features. Multi-headed attention networks help to build context-sensitive and speaker-sensitive dependency models, and to better capture temporal information in discourse. In the task of speech emotion analysis, processing images to obtain local features is very important. The transformer performs generally well in learning local features, so the framework proposed in this paper uses a model conformer [27], which is a fusion of convolution neural network with transformer. The conformer is a parallel two-body network structure, where the CNN branch uses the ResNet [28] structure and the transformer branch uses the ViT [29] structure. Conformer is the first parallel hybrid CNN and transformer network in which the local and global features of each phase will interact with each other through the proposed Feature Coupling Unit (FCU), thus giving conformer the advantages of both. Narayanan et al. [30] proposed a cross-attention conformer architecture designed to enhance contextual modeling of speech signals. O'Malley et al. [31] proposed a conformer-based front-end for speech sentiment recognition and improve its robustness with modules for noise reduction, echo cancellation, and speech enhancement. Li et al. [32] proposed an end-to-end streaming speech sentiment analysis model that employs a conformer layer to improve the problem that delay can lead to quality degradation. In classification, high accuracy can be achieved with smaller parameters and computation, and substantial improvements can be consistently achieved in target and instance segmentation. Thus, conformer has a strong ability to capture local and global information.

3. Methods

In this part, we detail the suggested speech sentiment analysis framework and its important parts, such as the conformer, the local feature learning group and the global feature learning group. Unlike most previous researchers using transformers for feature extraction, we use convolutionally enhanced transformers for feature extraction. This can not only obtain more and more advanced features, but also effectively avoid the problems of excessive training time and computational complexity of the traditional transformer model.

3.1. Preprocessing

Preprocessing is a method of processing speech samples before the extraction of characteristics of the speech signal [33]. We used four signal processing methods to transform the entered speech signal into an image input. Spectrogram, mid-frequency spectrum, cochleogram and windowed fractal dimension methods were used to transform the entered speech signal into a speech image. Figure 2 shows an example of our extracted MFCC features. Compared to the traditional use of a Convolutional Neural Network (CNN) to extract audio signal features, we use the Open SMILE toolkit [34] to extract MFCC features from the audio signal. To reduce the effect of the low Signal-to-Noise Ratio (SNR), we also use a specific window to sample the audio frames. We also use adaptive threshold preprocessing

to denoise and remove the silent part of the input audio. In general, MFCC is an accurate representation of the original audio and is widely used for automatic speech recognition.

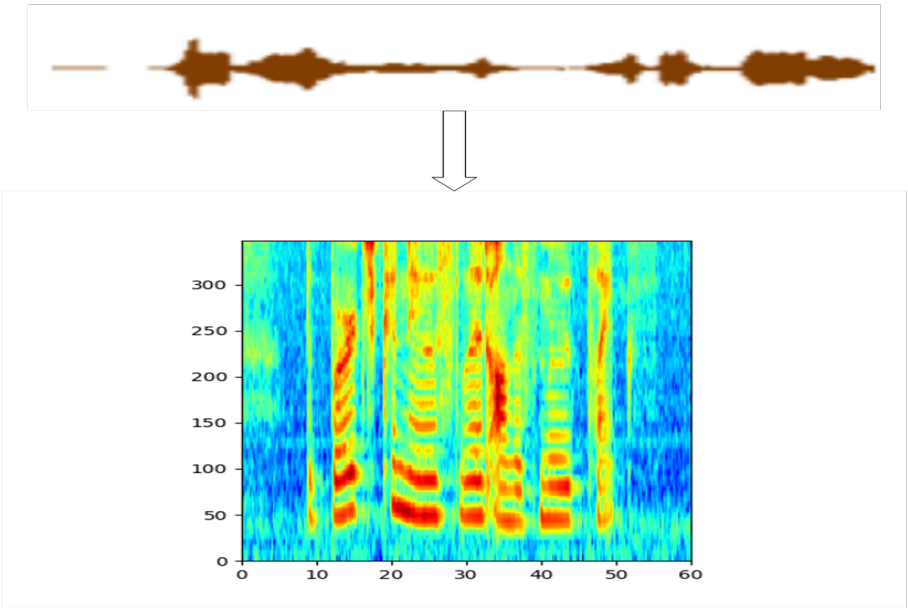

Figure 2. Example of extracting MFCC features from audio signals.

3.2. Local Feature Learning Groups

In the speech emotion analysis task, the order of speech segments is very important for analyzing the emotion of speech. There is a strong correlation between successive audio segments and therefore a great similarity in the sentiment present in them. The framework proposed in this paper also captures the correlations between consecutive audio segments by processing them and using them in the sequence prediction process to identify and analyze their expressed sentiments. As shown in Figure 3, we employ a local feature learning group that contains four identical local feature learning blocks. To sustain spatial and temporal trails, for instance, spatio-temporal information in sequential speech pieces, the conformer structure is used in this framework. Convolution operations are applied to transform input-to-state and state-to-state transitions within the conformer. Conformer catches spatio-temporal trails while convolution is in operation to better explore the correlation between speech segments. If x_i is used to denote the i-th input of the conformer layer, the output y_i is:

$$\tilde{x_i} = x_i + \frac{1}{2}FFN(x_i) \tag{1}$$

$$x_i^{'} = \tilde{x_i} + MHSA(\tilde{x_i}) \tag{2}$$

$$x_i^{''} = x_i^{'} + CONV(x_i^{'}) \tag{3}$$

$$y_i = Layernorm(x_i^{''} + \frac{1}{2}FFN(x_i^{''}) \tag{4}$$

Here, FFN, MHSA and Conv represent the feed forward network, multi-headed self-attentive network and convolution operation, respectively.

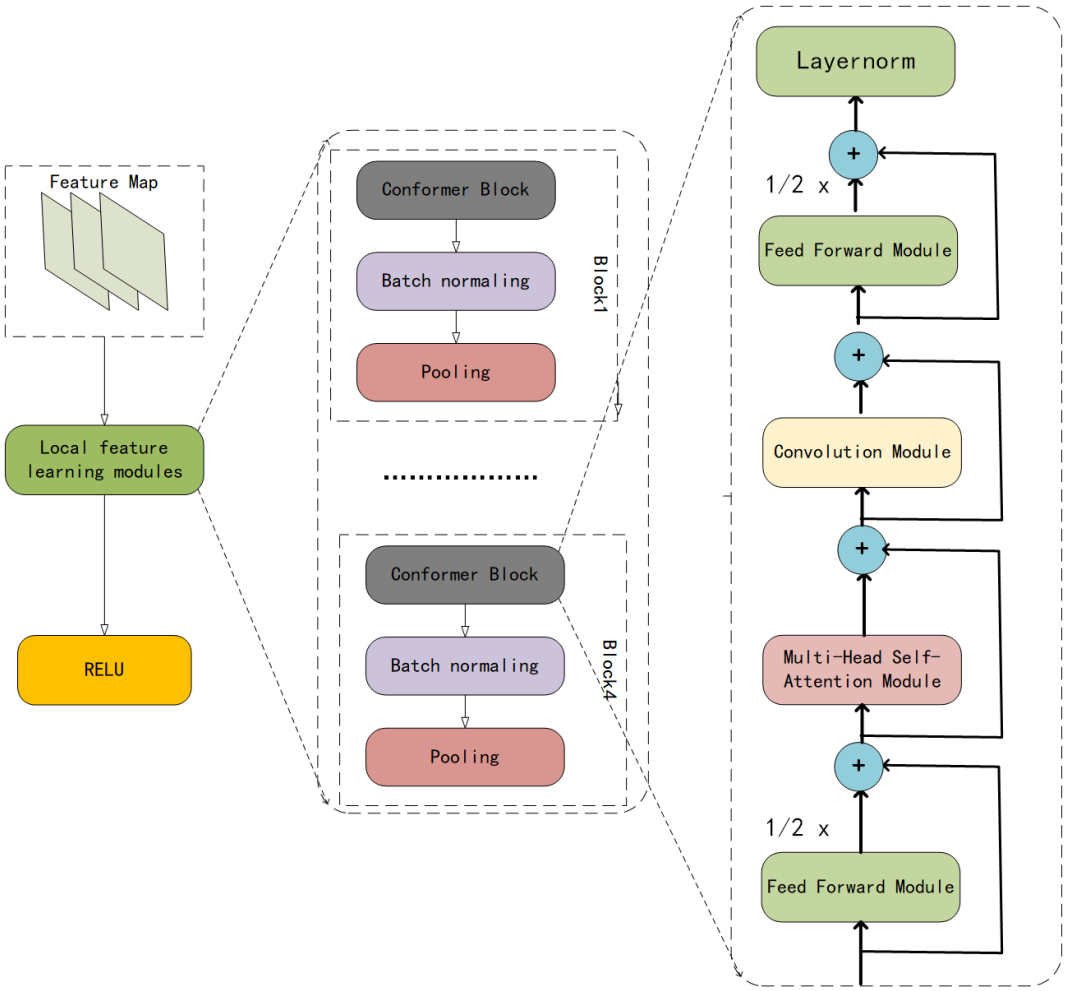

Figure 3. Framework diagram of local feature learning group.

For the calculation of MHSA, we may set $Q \in R^{L\times D}$, $K \in R^{S\times D}$, $V \in R^{S\times D}$, $K = V$, $d_K = D$, The self-attention formula is:

$$SA(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{5}$$

MHSA firstly performs linear dimension-raising operations on Q , K and V (that is, full connection operation, parameters are not shared), output, $\tilde{Q} \in R^{L\times \tilde{D}\times H}$, $\tilde{K} \in R^{S\times \tilde{D}\times H}$, $\tilde{V} \in R^{S\times \tilde{D}\times H}$, $\tilde{K} = \tilde{V}$, $d_{\tilde{K}} = \tilde{D} \times H$, then press "head" to do H times and the final result is spliced.

$$Head_i = SA(\tilde{Q}_i, \tilde{K}_i, \tilde{V}_i) \tag{6}$$

$$O = Concat(Head_1, Head_2, ..., Head_H) \tag{7}$$

Among them $\tilde{Q}_i = R^{L\times \tilde{D}}$, $\tilde{K}_i = R^{S\times \tilde{D}}$, $\tilde{V}_i = R^{S\times \tilde{D}}$, $Concat_{i=1}^{H}\tilde{Q}_i = \tilde{Q}$, $Concat_{i=1}^{H}\tilde{k}_i = \tilde{k}$, $Concat_{i=1}^{H}\tilde{v}_i = \tilde{v}$.

3.3. Global Feature Learning Group

Within the proposed sentiment recognition framework, we use and improve the global feature learning group, as shown in Figure 4. We employ GRU with the Attentional Update Gate (AUGRU) to learn global trails in features and identify long-term contextual dependencies, as shown in Figure 5. GRU acts the same as LSTM in capturing long sequences of semantic association information, which can effectively suppress gradient disappearance or explosion. Both outperform traditional RNNs and have lower computational complexity

compared to LSTM, and GRU networks are frequently applied to time sequence data [35]; however, GRU still cannot completely solve the gradient disappearance problem, while its role as a variant of RNN has a major drawback of RNN structure itself, i.e., it is not parallel computable. AUGRU retains the original dimensional information of the GRU update gate, which determines the importance of each dimension. Based on the differentiation information, it uses the attention score to measure all dimensions of the update gate, with the result that the less the degree of emotion-independent features, the less the impact on the hidden state. AUGRU is more effective in avoiding the interference of useless parts in speech images and increasing the accuracy of feature extraction.

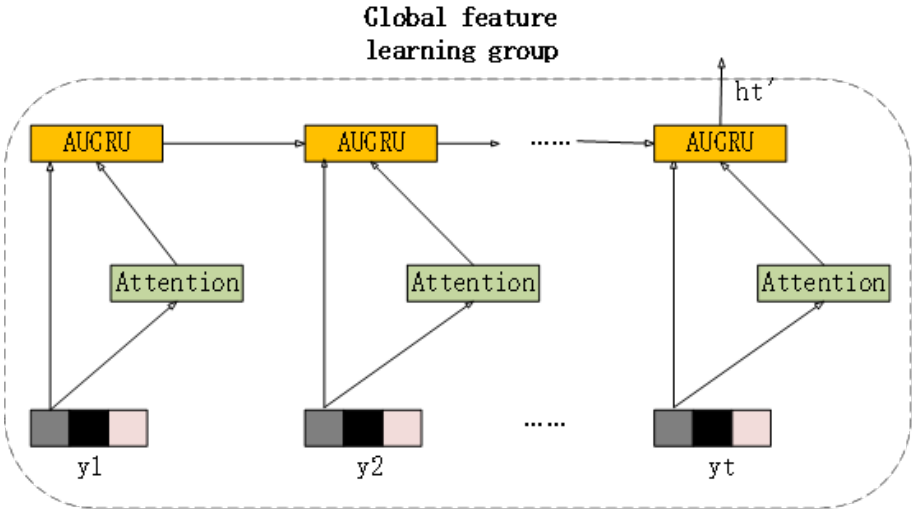

Figure 4. Global feature learning group framework diagram.

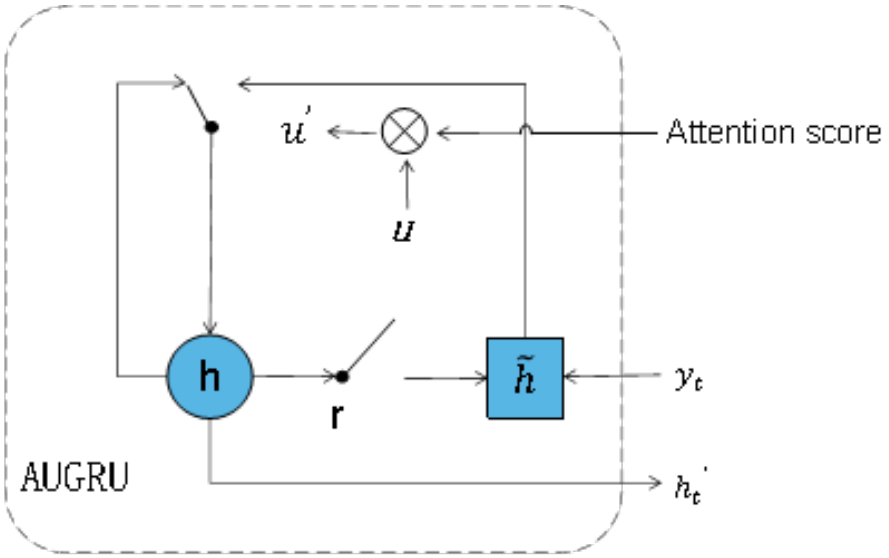

Figure 5. A schematic diagram of AUGRU structure.

The attention mechanism in the global feature learning module can be expressed as:

$$a_t = \frac{exp(y_t W)}{\sum_{i=1}^{t} exp(y_i W)} \tag{8}$$

AUGRU calculates global features can be expressed as:

$$r_t = \sigma(W^r y_t + U^r h_{t-1} + b^r) \tag{9}$$

$$u_t = \sigma(W^u y_t + U^u h_{t-1} + b^u) \tag{10}$$

$$\tilde{h_t} = tanh(W^h y_t + r_t \circ U^h h_{t-1} + b^h) \tag{11}$$

$$h_t = (1 - u_t) \circ h_{t-1} + u_t \circ \tilde{h_t}^{'} \tag{12}$$

$$\tilde{u_t}^{'} = a_t \times u_t^{'} \tag{13}$$

$$h_t^{'} = (1 - \tilde{u_t}^{'}) \circ h_{t-1}^{'} + \tilde{u_t}^{'} \circ \tilde{h_t}^{'} \tag{14}$$

where $u_t^{'}$ is the original update gate of AUGRU, σ is the sigmoid activation function, $\tilde{u_t}^{'}$ is the attentional update gate. $h_t^{'}$, $\circ$ is element-wise product, $\tilde{h_t}^{'}$ and $h_{t-1}^{'}$ are the hidden states. y_t is the input of the global feature learning module and h_t is the t-th hidden state.

3.4. Loss Function

The loss function widely used in sentiment classification tasks is the softmax function [36]; however, we adopt a combination of a central loss function [37] for computing depth features and a softmax function for sentiment classification to enhance the performed of the model and obtain a higher accuracy. The central loss function learns and discovers the centers of the feature vectors of each class, and defines the distances of between features as well as their consistent class centers. We utilize the mean "λ" setting to compute and adjust the update loss function for inter/intra-class distances. We measure the minimum distance between classes with the central loss function and measure the maximum distance between classes with the softmax loss function, which can be expressed as:

$$L_s = -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^t \cdot x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_{y_i}^t \cdot x_i + b_{y_i}}} \tag{15}$$

$$L_c = \frac{1}{2} \sum_{i=1}^{m} ||x_i - c_{y_i}||_2^2 \tag{16}$$

We use "n" and "m" to denote the number of classes and the minimum batch size to the classifier, and c_{y_i} denotes the center of the class y_i in the i-th sample. We use the "λ" notation for central loss to compute the minimum distance required to avoid misclassification in the real-time case. The global loss we use in our model is:

$$L = L_S + \lambda L_C \tag{17}$$

Here "λ" is used as a hyperparameter. Experiments demonstrate that combining the central loss function gives better results.

4. Experimental Assessment

In this subsection, we implement the speech emotion analysis task by converting speech signals into image feature sequences, thereby demonstrating the effectiveness of our proposed speech emotion recognition architecture.

4.1. Datasets

We evaluate our proposed method on different benchmarks such as weighted and unweighted accuracy, precision, recall and F1_score using two commonly used datasets, IEMOCAP [38] and RAVDESS [39] speech corpora. The proposed sentiment recognition framework's capabilities are compared with most advanced models, and the model performance, training, validation accuracy and loss plots for both datasets are shown in Figure 6.

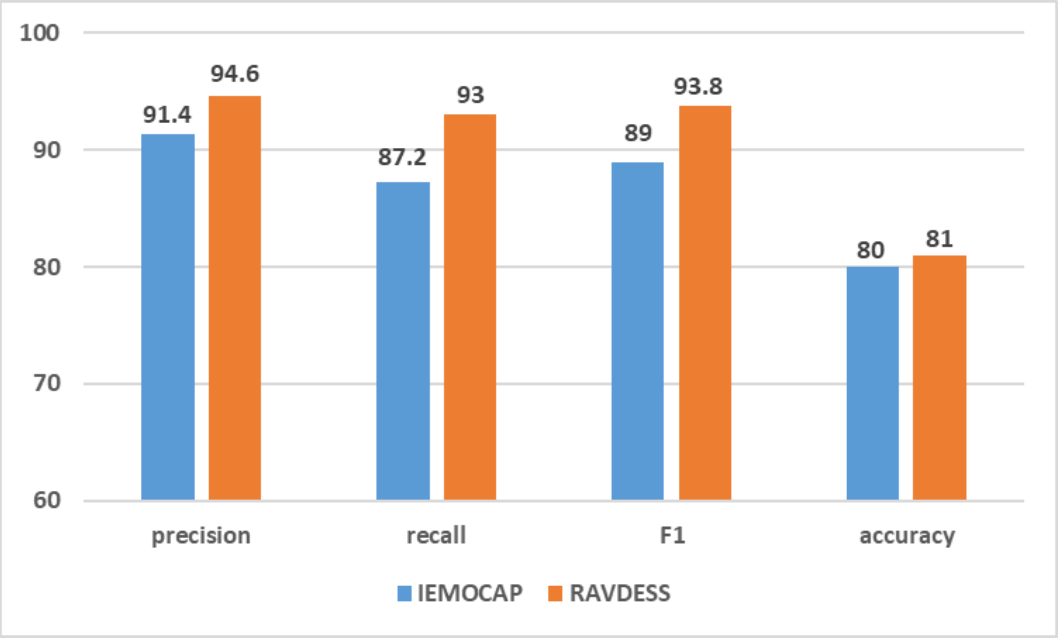

Figure 6. Experimental results on two datasets.

IEMOCAP: Interactive Emotion Binary Motion Capture dataset was proposed in 2008 and widely used in speech sentiment analysis tasks. It is collected by professional actors through hours of dialogue and conversation in different scenes. It consists of five parts and lasts up to 12 h. A total of nine emotions are included in the dataset: anger, happiness, excitement, sadness, depression, fear, surprise, other and neutral states. Table 1 shows the percentage of unused sentiment in the dataset. Each conversation is evaluated by three or more evaluators and when more than half of the evaluators agree, the conversation is marked with the corresponding label. When the evaluators disagree, the conversation is marked "Other". It is more natural to acquire data this way.

RAVDESS: The RAVDESS dataset consists of two parts: the emotional speech part and the song data part. In the speech sentiment analysis task, we mainly focus on the emotional speech part. The emotional voice part is collected by 12 professional male actors and 12 professional actresses performing dialogue performances, with a total of 1440 dialogues. There are eight emotions: calm, happy, sad, angry, fearful, surprised, disgusted and neutral. Table 2 shows the percentage of unused sentiment in the dataset. During data collection, actors were asked to speak two fixed sentences with different sentiment categories; therefore, the balance of the RAVDESS dataset is good, but the naturalness is relatively poor. In this paper, the RAVDESS dataset is used as an auxiliary dataset.

Table 1. The distribution of different emotions in the dataset IEMOCAP.

Class	Utterances	Participation (%)
Anger	1103	19.9
Happiness	1636	29.6
Sad	1084	19.6
Neutral	1708	30.9

Table 2. Distribution of different emotions in the dataset RAVDESS.

Class	Utterances	Participation (%)
Anger	192	13.3
Happiness	192	13.3
Sad	192	13.3
Neutral	96	6.9
Calm	192	13.3
Disgust	192	13.3
Fear	192	13.3
Surprise	192	13.3

4.2. Metrics

The following are the evaluation metrics we use: Precision, Recall, F1_score (F1) and Accuracy(Acc.).

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

$$Recall = \frac{TP}{TP + FN} \tag{19}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{20}$$

$$ACC = \frac{TP + FN}{TP + FP + TN + FN} \tag{21}$$

Among them, TP, FP, TN and FN are true positive, true negative, false positive and false negative, respectively.

4.3. Details

Our suggested model is implemented on the python-based Keras deep learning library. We extract acoustic features using the openSMILE toolbox. We train our model on a machine with a GeForce RTX 2080Ti and optimize it using Adam [40] with an initial learning rate of 0.0001. The batch size is set to 100 and the epochs to 128. For the testing and evaluation of the framework model introduced in this paper, we use 75% of the dataset for training and 25% for testing and validation. For the generalization of the model, we adopt the k-fold cross-validation method [41] to study the reliability and significance of the model. In the k-fold cross-validation technique, we train the model by splitting the database into five folds using an automated technique.

4.4. Baselines

To demonstrate the performance of our proposed model framework, we compare with the following speech sentiment analysis baselines.

CNN+BiLSTM: Kwon et al. [11] proposed a CNN+BiLSTM framework model, and the key contribution lies in the use of normalization techniques to enhance the use of features.

K-Mean+DBN: Senthilkumar et al. [24] proposed a method combining deep belief networks with K-Mean clustering to improve the efficiency of speech emotion recognition tasks.

DCNN: Issa et al. [42] introduced a novel framework that extracts some acoustic features from speech signals as input for emotion recognition and compared to previous work, it takes raw audio samples and does not convert to visual features.

DSCNN: Kwon et al. [15] proposed an AI-assisted Deep Convolution Neural Network (DSCNN) architecture that use a common network approach to learn outstanding and discriminative features from the spectrogram of the voice signal, which are enforced in the previous steps and thus behave better.

Bag-SVM: Bhavan et al. [43] proposed a bagged ensemble consisting of support vector machines and Gaussian kernels, from which a combination of spectral features was extracted, further processed and reduced to the desired feature set.

ConvLSTM+GRUs: Kwon et al. [17] proposed the ConvLSTM model, which performs local feature learning by layering, and combines GRU for global feature learning to improve the accuracy of the model.

HA-ACNN: Xu et al. [44] proposed a multi-head attention-based convolution neural network model, and in order to improve the model accuracy, the authors also studied the impact of noise intensity and time-shift effects on the accuracy.

LSTM-Transformer: Andayani et al. [45] proposed a hybrid model architecture of a hybrid long short-term memory network and transformer encoder, which is able to preserve hidden states and learn long- and short-term dependencies using a multi-head attention mechanism.

5. Results and Analysis

The experimental results of our suggested model framework on two datasets are shown in Figure 6, our introduced method achieves 80% (Accuracy) and 81% (Accuracy) on datasets IEMOCAP and RAVDESS, respectively.

Table 3 presents the experimental results of various benchmark models on the dataset IEMOCAP, including evaluation metrics such as weighted, unweighted, etc., to show the strength of the proposed method. From the experimental point of view, our proposed method produces relatively good results. In particular, our method is significantly more accurate than [17]. This shows that the transformer structure is more suitable than the LSTM model for processing speech sentiment analysis tasks, and the attention-based GRUs are also better than the stacked GRUs for processing features. Comparing with method [15], it can be demonstrated that the combination of convolution and transformer can significantly enhance the extraction of feature information and can reduce its time complexity. In addition, we also compared the methods with and without attention in the model. The experimental results show that the attention mechanism plays a very important role in the deep learning model. To demonstrate the strength of our model, we also classify different emotions on the dataset IEMOCAP, as shown in Figure 7. According to the experimental results, it can be seen that our model is not as effective for the "happy" emotion as other types of emotion. In this regard, by comparing the effects of other methods and models, we conclude that a possible reason is that such emotions are underrepresented in the dataset.

Table 3. Accuracy summary obtained by various models using the IEMOCAP database. WA is the weighted accuracy and UA is the unweighted accuracy.

Method	Year	WA (%)	UA (%)	Acc. (%)
CNN+BiLSTM	2020	74	72	72.5
BiLSTM+DBN	2021	-	-	72.3
DCNN	2020	-	-	64.3
DSCNN	2019	76	72	73.8
HA-ACNN	2021	76.2	76.4	-
ConvLSTM+GRUs	2020	77	75	78
Conformer+AUGRUs (ours)	2022	80	78	80

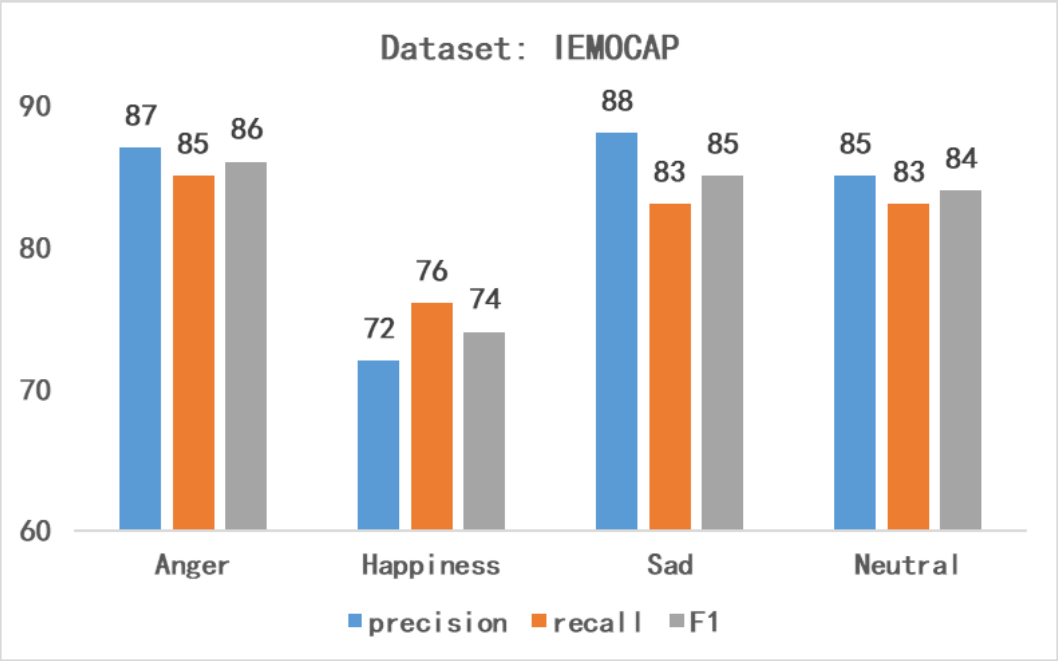

Figure 7. Recognition of different emotions by the proposed method.

Now, the RAVDESS [33] dataset is frequently used in sentiment speech and song recognition systems to test its meaning and effectiveness. Table 4 presents the experimental results of various benchmark models on the dataset RAVDESS, including weighted, unweighted, etc., evaluation metrics to show the strength of the proposed method. From the experimental data, it can be seen that our proposed method significantly outperforms other benchmark models in various metrics. Figure 8 shows the recognition accuracy of our proposed model for different emotions on the dataset RAVDESS. In particular, our method achieves F1_scores of 91% and 90% on the angry and calm emotion classification tasks. The poor performance on the neutral emotion classification task may be due to the fact that the sample rate of this type of emotion in the dataset is too small and under-represented.

Table 4. Accuracy summary of various models obtained using RAVDESS database. WA is the weighted accuracy and UA is the unweighted accuracy.

Method	Year	WA (%)	UA (%)	Acc. (%)
CNN+BiLSTM	2020	81	77	77
BiLSTM+DBN	2021	-	-	77
DCNN	2020	-	-	77.6
DSCNN	2019	68	61	70
HA-ACNN	2021	77.8	77.4	-
ConvLSTM+GRUs	2020	81	80	80
Bag-SVM	2020	-	-	75.7
LSTM-Transformer	2020	77.3	75.6	-
Conformer+AUGRUs (ours)	2022	82.5	80.3	81

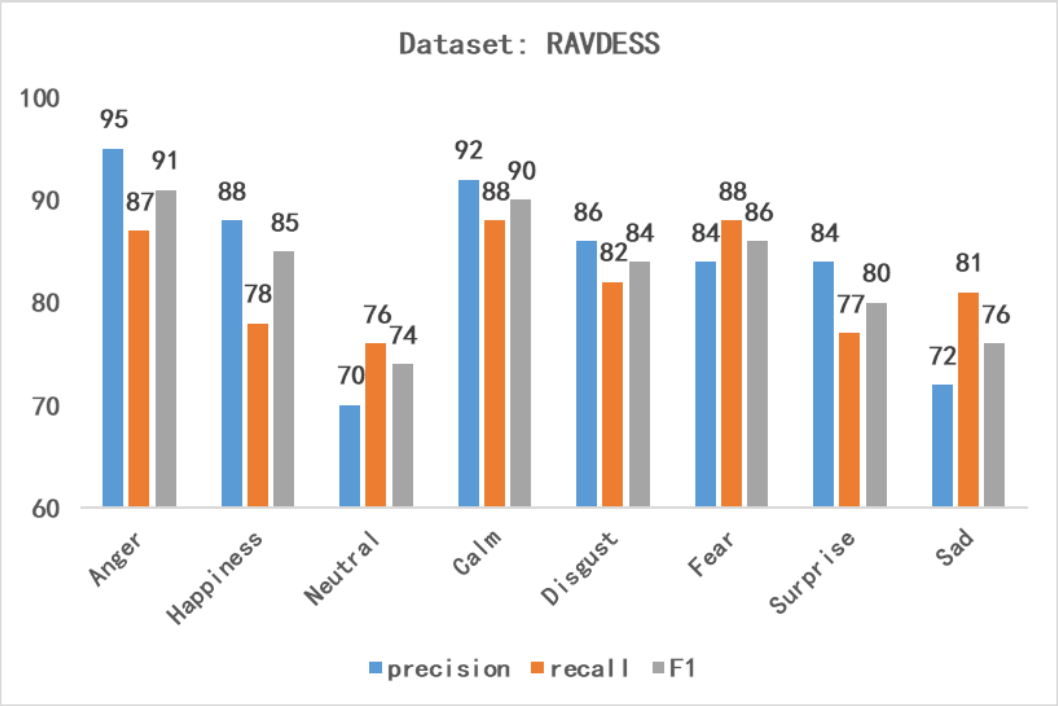

Figure 8. Recognition of different emotions by the proposed method.

Through experiments, a comparison with the method of [43] shows that there is a large gap between models of deep learning and machine learning models in terms of extraction of feature information and accuracy of analysis results. The development of deep models such as CNN and transformer is more favorable to promote the development of sentiment analysis tasks. In contrast to methods [15,42,44], etc., the experimental results show that convolutional operations and attentional mechanisms can produce significant results on sentiment analysis tasks and have great potential for development. We compare with the method proposed in [45]. Experiments show that the long short-term memory network can perform feature extraction with the transformer, but there is still a problem of insufficient feature information. So far as we know, the combination of convolutions and transformers is state-of-the-art for speech sentiment analysis tasks. Experiments also prove that the combination of convolution for local feature extraction and the transformer can better enhance the robustness of the model and reduce the time complexity of model training.

6. Conclusions and Future Direction

In this article, we introduce a new method that aims at a more accurate extraction of local and global features of speech signals. We use a combination of convolution and a transformer to better obtain the spatio-temporal features as well as the long and short-term dependencies of the speech signal. In this approach, we construct corresponding learning groups for local features and global features, respectively. The local feature learning group contains four learning blocks combined by convolution and transformer to learn the spatial contextual relevance of sentiments, using surplus learning strategies in a hierarchical manner. We used these blocks to extract the most salient emotional features. Then, we input these derived features into a GRU network with a fused attention mechanism to recalibrate the global weights using these learned features. We use the fully connected network to further process the feature information after fusion by the attention mechanism, and finally analyze the results by softmax and central loss function. We performed a wide range of experiments on the datasets IEMOCAP and RAVDESS speech libraries to examine and appraise the effectiveness and importance of the suggested model relative to most advanced models. Our suggested system shows excellent results, with IEMOCAP and REVDESS obtaining 80% and 81% recognition accuracy, respectively, clearly showing the

reliability of the model. The experiments demonstrate that the model is effective on various existing systems in terms of sentiment analysis and sentiment classification.

In future work, we plan to implement sentiment analysis tasks in terms of multimodality. We use this relatively novel convolutional and transformer architecture to better analyze the spatio-temporal features of different modalities. Then, we will perform the fusion of different modal features by deep learning methods to achieve the sentiment classification task under multimodality. The method introduced in this paper can also be applied to more speech-like sentiment analysis tasks to further achieve more effective speech sentiment analysis tasks by finding the interconnections between different modalities.

Author Contributions: P.Z. writing—original draft, writing—review and editing. F.L. conceptualization, methodology. X.Z. software, data curation. All authors have read and agreed to the published version of the manuscript.

Funding: We are grateful for the support of the National Natural Science Foundation of Shandong ZR202011020044. We are grateful for the support of the National Natural Science Foundation of China 61772321.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict to interest.

References

1. Schuller, B.W. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* **2018**, *61*, 90–99. [CrossRef]
2. Hossain, M.S.; Muhammad, G. Emotion recognition using deep learning approach from audio–visual emotional big data. *Inf. Fusion* **2019**, *49*, 69–78. [CrossRef]
3. Bhardwaj, V.; Ben Othman, M.T.; Kukreja, V.; Belkhier, Y.; Bajaj, M.; Goud, B.S.; Rehman, A.U.; Shafiq, M.; Hamam, H. Automatic Speech Recognition (ASR) Systems for Children: A Systematic Literature Review. *Appl. Sci.* **2022**, *12*, 4419. [CrossRef]
4. Duan, J.; Luo, B.; Zeng, J. Semi-supervised learning with generative model for sentiment classification of stock messages. *Expert Syst. Appl.* **2020**, *158*, 113540. [CrossRef]
5. Fares, M.; Moufarrej, A.; Jreij, E.; Tekli, J.; Grosky, W. Unsupervised word-level affect analysis and propagation in a lexical knowledge graph. *Knowl. Based Syst.* **2019**, *165*, 432–459. [CrossRef]
6. Bibi, M.; Abbasi, W.A.; Aziz, W.; Khalil, S.; Uddin, M.; Iwendi, C.; Gadekallu, T.R. A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis. *Pattern Recognit. Lett.* **2022**, *158*, 80–86. [CrossRef]
7. Abboud, R.; Tekli, J. Integration of nonparametric fuzzy classification with an evolutionary-developmental framework to perform music sentiment-based analysis and composition. *Soft Comput.* **2020**, *24*, 9875–9925. [CrossRef]
8. Sun, C.; Lv, L.; Tian, G.; Liu, T. Deep interactive memory network for aspect-level sentiment analysis. *ACM Trans. Asian Low Resour. Lang. Inf. Process. TALLIP* **2020**, *20*, 1–12. [CrossRef]
9. Li, J.; Deng, L.; Haeb-Umbach, R.; Gong, Y. Fundamentals of speech recognition. In *Robust Automatic Speech Recognition*; Academic Press; Waltham, MA, USA, 2016; pp. 9–40.
10. Shaik, R.; Venkatramaphanikumar, S. Sentiment analysis with word-based Urdu speech recognition. *J. Ambient. Intell. Humaniz. Comput.* **2022**, *13*, 2511–2531. [CrossRef]
11. Mustaqeem; Sajjad, M.; Kwon, S. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access* **2020**, *8*, 79861–79875.
12. Jara-Vera, V.; Sánchez-Ávila, C. Cryptobiometrics for the Generation of Cancellable Symmetric and Asymmetric Ciphers with Perfect Secrecy. *Mathematics* **2020**, *8*, 1536. [CrossRef]
13. Khalil, R.A.; Jones, E.; Babar, M.I.; Jan, T.; Zafar, M.H.; Alhussain, T. Speech emotion recognition using deep learning techniques: A review. *IEEE Access* **2019**, *7*, 117327–117345. [CrossRef]
14. Zhang, J.; Jiang, X.; Chen, X.; Li, X.; Guo, D.; Cui, L. Wind power generation prediction based on LSTM. In Proceedings of the 2019 4th International Conference on Mathematics and Artificial Intelligence, Chegndu, China, 12–15 April 2019; pp. 85–89.
15. Kwon, S. A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors* **2019**, *20*, 183.
16. Chernykh, V.; Prikhodko, P. Emotion recognition from speech with recurrent neural networks. *arXiv* **2017**, arXiv:1701.08071.
17. Kwon, S. CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM network. *Mathematics* **2020**, *8*, 2133.

18. Koduru, A.; Valiveti, H.B.; Budati, A.K. Feature extraction algorithms to improve the speech emotion recognition rate. *Int. J. Speech Technol.* **2020**, *23*, 45–55. [CrossRef]
19. Likitha, M.; Gupta, S.R.R.; Hasitha, K.; Raju, A.U. Speech based human emotion recognition using MFCC. In Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 22–24 March 2017; pp. 2257–2260.
20. Chatziagapi, A.; Paraskevopoulos, G.; Sgouropoulos, D.; Pantazopoulos, G.; Nikandrou, M.; Giannakopoulos, T.; Katsamanis, A.; Potamianos, A.; Narayanan, S. Data Augmentation Using GANs for Speech Emotion Recognition. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 171–175.
21. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
22. Zhang, H.; Huang, H.; Han, H. A Novel Heterogeneous Parallel Convolution Bi-LSTM for Speech Emotion Recognition. *Appl. Sci.* **2021**, *11*, 9897. [CrossRef]
23. Atila, O.; Şengür, A. Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition. *Appl. Acoust.* **2021**, *182*, 108260. [CrossRef]
24. Senthilkumar, N.; Karpakam, S.; Devi, M.G.; Balakumaresan, R.; Dhilipkumar, P. Speech emotion recognition based on Bi-directional LSTM architecture and deep belief networks. *Mater. Today Proc.* **2022**, *57*, 2180–2184. [CrossRef]
25. Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; Tran, D. Image transformer. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 4055–4064.
26. 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]
27. Peng, Z.; Huang, W.; Gu, S.; Xie, L.; Wang, Y.; Jiao, J.; Ye, Q. Conformer: Local features coupling global representations for visual recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 367–376.
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
29. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* 2022, *in press*. [CrossRef] [PubMed]
30. Narayanan, A.; Chiu, C.C.; O'Malley, T.; Wang, Q.; He, Y. Cross-attention conformer for context modeling in speech enhancement for ASR. In Proceedings of the 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 13–17 December 2021; pp. 312–319.
31. O'Malley, T.; Narayanan, A.; Wang, Q.; Park, A.; Walker, J.; Howard, N. A conformer-based asr frontend for joint acoustic echo cancellation, speech enhancement and speech separation. In Proceedings of the 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 13–17 December 2021; pp. 304–311.
32. Li, B.; Gulati, A.; Yu, J.; Sainath, T.N.; Chiu, C.C.; Narayanan, A.; Chang, S.Y.; Pang, R.; He, Y.; Qin, J.; et al. A better and faster end-to-end model for streaming asr. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–12 June 2021; pp. 5634–5638.
33. Kurpukdee, N.; Kasuriya, S.; Chunwijitra, V.; Wutiwiwatchai, C.; Lamsrichan, P. A study of support vector machines for emotional speech recognition. In Proceedings of the 2017 8th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES), Chonburi, Thailand, 7–9 May 2017; pp. 1–6.
34. Eyben, F.; Weninger, F.; Gross, F.; Schuller, B. Recent developments in opensmile, the munich open-source multimedia feature extractor. In Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain, 21–25 October 2013; pp. 835–838.
35. Qin, Y.; Song, D.; Chen, H.; Cheng, W.; Jiang, G.; Cottrell, G. A dual-stage attention-based recurrent neural network for time series prediction. *arXiv* **2017**, arXiv:1704.02971.
36. Wang, M.; Lu, S.; Zhu, D.; Lin, J.; Wang, Z. A high-speed and low-complexity architecture for softmax function in deep learning. In Proceedings of the 2018 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), Chengdu, China, 26–30 October 2018; pp. 223–226.
37. Akbari, A.; Awais, M.; Bashar, M.; Kittler, J. How does loss function affect generalization performance of deep learning? Application to human age estimation. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 141–151.
38. 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]
39. 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*, e0196391. [CrossRef]
40. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
41. Fushiki, T. Estimation of prediction error by using K-fold cross-validation. *Stat. Comput.* **2011**, *21*, 137–146. [CrossRef]
42. Issa, D.; Demirci, M.F.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control* **2020**, *59*, 101894. [CrossRef]
43. Bhavan, A.; Chauhan, P.; Hitkul; Shah, R.R. Bagged support vector machines for emotion recognition from speech. *Knowl. Based Syst.* **2019**, *184*, 104886. [CrossRef]

44. Xu, M.; Zhang, F.; Zhang, W. Head fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset. *IEEE Access* **2021**, *9*, 74539–74549. [CrossRef]
45. Andayani, F.; Theng, L.B.; Tsun, M.T.; Chua, C. Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files. *IEEE Access* **2022**, *10*, 36018–36027. [CrossRef]