applied sciences

MDPI

Article

A Feature Fusion Model with Data Augmentation for Speech Emotion Recognition

Zhongwen Tu [1,*], Bin Liu [2], Wei Zhao [3], Raoxin Yan [2] and Yang Zou [2]

[1] Educational Service Center, Communication University of China, Beijing 100024, China
[2] School of Information and Engineering, Communication University of China, Beijing 100024, China
[3] School of Data and Intelligence, Communication University of China, Beijing 100024, China
* Correspondence: bytuzhongwen@cuc.edu.cn

Abstract: The Speech Emotion Recognition (SER) algorithm, which aims to analyze the expressed emotion from a speech, has always been an important topic in speech acoustic tasks. In recent years, the application of deep-learning methods has made great progress in SER. However, the small scale of the emotional speech dataset and the lack of effective emotional feature representation still limit the development of research. In this paper, a novel SER method, combining data augmentation, feature selection and feature fusion, is proposed. First, aiming at the problem that there are inadequate samples in the speech emotion dataset and the number of samples in each category is unbalanced, a speech data augmentation method, Mix-wav, is proposed which is applied to the audio of the same emotion category. Then, on the one hand, a Multi-Head Attention mechanism-based Convolutional Recurrent Neural Network (MHA-CRNN) model is proposed to further extract the spectrum vector from the Log-Mel spectrum. On the other hand, Light Gradient Boosting Machine (LightGBM) is used for feature set selection and feature dimensionality reduction in four emotion global feature sets, and more effective emotion statistical features are extracted for feature fusion with the previously extracted spectrum vector. Experiments are carried out on the public dataset Interactive Emotional Dyadic Motion Capture (IEMOCAP) and Chinese Hierarchical Speech Emotion Dataset of Broadcasting (CHSE-DB). The experiments show that the proposed method achieves 66.44% and 93.47% of the unweighted average test accuracy, respectively. Our research shows that the global feature set after feature selection can supplement the features extracted by a single deep-learning model through feature fusion to achieve better classification accuracy.

Keywords: speech emotion recognition; data augmentation; feature selection; multi-head attention; features fusion

check for updates

Citation: Tu, Z.; Liu, B.; Zhao, W.; Yan, R.; Zou, Y. A Feature Fusion Model with Data Augmentation for Speech Emotion Recognition. *Appl. Sci.* **2023**, *13*, 4124. https://doi.org/10.3390/app13074124

Academic Editors: Yoshinobu Kajikawa and Cheng-Yuan Chang

Received: 28 February 2023
Revised: 21 March 2023
Accepted: 22 March 2023
Published: 24 March 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

There may be semantic barriers between different languages, but phonetic emotion is a common expression of human beings, and we can identify the emotion of the speaker of a language that we do not understand without relying on semantics. Because of the great convenience of language communication and the cross-linguistic, cross-cultural and cross-regional commonality of emotion, phonetic emotion can help to express ourselves. It is natural to think of realizing this form on a computer [1]. Emotional intelligence is an important part of robot intelligence. It aims to promote natural interaction with machines through direct voice interaction [2] and making robots have humanoid intelligence and emotion is an important factor [3].

The research of SER is inseparable from the high-quality emotion speech dataset. The quality of the emotion speech dataset directly determines the performance upper limit of the emotion recognition system trained by it. In recent years, image classification, speech recognition, natural language processing and other tasks have developed rapidly because of the huge benchmark dataset. Due to the subjectivity and complexity of speech

emotion tasks, the collection of speech emotion data is limited. At present, the amount of dataset data in the field is relatively small [4–6], up to tens of thousands. The number of dataset samples is not as large as Imagenet [7] and Audioset [8], and is still not completely unified between performances, affective models and collection methods. In view of the lack of data, data augmentation is a convenient and feasible method. The commonly used speech data augmentation methods are speed disturbance [9], pitch adjustment [10], the addition of noise [11], increasing the amount of data by changing the audio rate, fundamental frequency, or adding noise to the original voice. In addition, Chatziagapi et al. applied Generative Adversarial Networks (GANs) to generate spectrograms, enhanced the training data of samples, and conducted experiments on IEMOCAP and FEEL–25k emotion database [12]. Xu et al. used the Vocal Tract Length Perturbation (VTLP) method to enhance the speech emotion classification model [13]. The authors in [14] proposed pre-training neural networks named Pretrained Audio Neural Networks (PANNs) for audio scene classification, using two data enhancement methods, Mixup [15] and SpecAugment [16]. Researchers have constantly explored audio data enhancement methods, from traditional methods to modern methods such as neural networks. We have also explored and migrated the methods suitable for speech synthesis, which can improve this problem.

Deep learning accelerates the development of speech emotion. However, as the first step in speech signal processing–feature extraction, various manually designed low-level features are still widely used in SER [17–19]. The discovery of a speech acoustic feature that can effectively carry speech emotion information has become the key to the improvement of SER. At present, Mel Frequency Cepstral Coefficient (MFCC) and Log-Mel spectrograms are widely used in SER research [20–22] because that they can better describe the time–frequency correlation of emotional details [23].

At present, Convolution Neural Network (CNN) [24], Long Short-Term Memory (LSTM) [25], attention mechanism and other methods are mainly used for depth feature extraction in audio classification tasks and speech emotion recognition [20–23,26–28], which has achieved great improvement compared with the traditional model. However, on the one hand, existing models do not fully consider how to establish effective connections between speech; on the other, models are often based on a single-input feature classification, the accuracy of model is often unsatisfactory, and the ability of the model needs to be further improved [29].

Therefore, aiming at the SER classification task of discrete models, this paper proposes an emotional speech data augmentation method, using the machine-learning method to select from several global feature sets of speech emotion to reduce the dimensionality of the selected feature set, and conducts the feature fusion combined with the features extracted from the deep-learning model MHA-CRNN. In this paper, our main contributions are as follows:

(1) We propose an emotional speech data augmentation method, Mix-wav, migrated from the data augmentation method Mixup in the image recognition task. Experiments on two datasets, IEMOCAP and CHSE-DB, show that the proposed method improves the recognition accuracy of the model.
(2) We propose a novel speech emotion feature-extraction model MHA-CRNN based on multi-head attention mechanism. Combining the advantages of CNN and LSTM effectively extracts the spectral features of speech and uses the multi-head attention mechanism to further pay attention to the emotion-related information in the time and frequency domain.
(3) We use four global feature sets to conduct emotion classification experiment on IEMOCAP, chose the feature set with the best performance, and use two methods to conduct feature dimension reduction experiments to find out the feature subset with the best performance.
(4) We use the selected global statistical features with the features extracted by the MHA-CRNN model to perform feature fusion. We tested our suggested SER model on benchmark IEMOCAP and self-built dataset CHSE-DB. It achieved 66.44% and 93.47%

(Unweighted Average Accuracy, WAA), respectively. In the comparative analysis, our system showed our performed recognition results.

The rest of this paper is structured as follows. The second section describes some related work in the field of SER. Section 3 describes our proposed data augmentation method. Section 4 describes the method and model architecture proposed in our paper. Section 5 describes the experimental setup, results and related analyses. Section 6 describes the conclusions of our paper and future work.

2. Related Work

After decades of research, speech emotion recognition research has formed a relatively mature process framework and a certain research scale. Before the development of deep-learning methods, the more widely used classic machine-learning models in the field of speech emotion recognition include Naive Bayes classifier, Gaussian Mixture Model (GMM) [30], Hidden Markov Model (HMM) [31], Support Vector Machines (SVM) [32], etc. In recent years, with the development of deep-learning methods, these have been widely used in the field of SER. At present, there is no unified standard for the establishment of the speech emotion database, and the database is often prone to problems such as small data volume and unbalanced data with different labels, which makes the model prone to overfitting. The effective improvement of the impact of data problems on the model is part of our work.

Data augmentation is an effective way to prevent overfitting. Mixup is an effective supervised learning method for image data augmentation. It trains the model with a combination of a pair of sample inputs and their targets, making the model more robust to adversarial samples [15]. Mixup can be easily applied to image classification tasks where the target is a single label. The authors in [14] used the Mixup method on the audio scene classification task, and its input was the Log-Mel spectrogram, which improved the Area Under the Curve (AUC) index by 3% compared with the non-Mixup method. MixSpeech [33] improved Mixup and applied it to automatic speech recognition tasks. The input was the Log-Mel spectrogram and MFCC, and the label was a text sequence. It obtained a good Word Error Rate (WER) of 4.7% on the "Wall Street Journal" dataset, which is 10.6% higher than the statement error rate of the data augmentation method SpecAugment. It can be seen that the improved Mixup method has achieved good results in enhancing the characteristic data of the two-dimensional audio spectrum, but its application and effectiveness in the original one-dimensional audio data have not been verified. We try to improve the Mixup method and apply it to the original one-dimensional audio to realize data enhancement.

Feature extraction is the key to emotion recognition, and the quality of features has a great impact on the performance of SER algorithms [34]. The selection of a dataset with strong emotional representation ability and more robustness from various feature sets is a requirement for researchers before solving the learning model.

Deep Neural Networks (DNNs) have shown promising performance in extracting high-level features of SER recent years. Tzirakis [35] experimented with audio waveforms as input, first using CNN to extract representations, and then putting them into LSTM for emotion recognition. Lee et al. [36] applied a Recurrent Neural Network (RNN) to learn the long-term temporal relationship of SER. Since the attention mechanism greatly improves the accuracy of machine translation [37], its application in speech emotion recognition is also widely used. Chen et al. [21] took the Log-Mel spectrogram and its first and second-order difference coefficients as input, and proposed a three-dimensional Convolutional Recurrent Neural Network model (ACRNN) with hierarchical attention for speech emotion classification. Nediyanchath et al. [38] proposed a multi-head attention deep-learning speech emotion recognition network based on Logarithmic Mel Filter Bank Energy (LMFBE) spectral features. The design of a model with better recognition performance is also our aim.

Another problem with input features is that most studies tend to use only a single spectral feature as input. Only the local feature extraction of emotion in speech is considered, but the coherence of speech as a whole is not considered. Compared with low-level spectral features, there is another type of acoustic feature for speech signal processing called global features, also known as high-level statistical features. Generally, a feature set is formed by applying a global statistical function to the acoustic features of each frame in a speech segment to obtain the statistical values of all frame-level features in the speech segment. Common global feature sets used in speech emotion recognition tasks include the IS09_emotion [39], IS10_paraling [40], IS13_ ComParE [41] feature sets, etc. We find a relatively suitable global feature to integrate into the recognition part through experimental tests.

In our work, we propose an emotional speech data augmentation method, Mix-wav, for a classification task based on the discrete emotion model, which enhances and balances the dataset samples, and then propose MHA-CRNN using CNN combined with a multi-head attention mechanism. Using LSTM from the Log-Mel spectrum extracts discriminative emotional features and explores the parameters of multi-head attention. Then, the optimal emotion-related global features are selected from the four global features, and are fused with the features extracted by the above model. The experimental results show that our proposed method can improve the performance of the system effectively.

3. Proposed Methods

The methods proposed in this paper include audio data augmentation Mix-wav, LightGBM-based global fusion feature selection method, speech emotion feature model MHA-CRNN, and its feature fusion model. This paper first proposes a time-domain emotional speech data enhancement method to solve the problem that emotional datasets are often small in scale. In addition, a deep-learning model, MHA-CRNN, is proposed to further extract the emotional information in the time–frequency domain from the spectral features, and the LightGBM method is used to screen and reduce the dimensionality of the HSF, select the most effective features, and compare the frame-level acoustic features with the obtained high-level statistical features, which are combined to obtain a fused sentiment feature vector.

The following subsections explain the proposed method in detail. Figure 1 presents the suggested method.

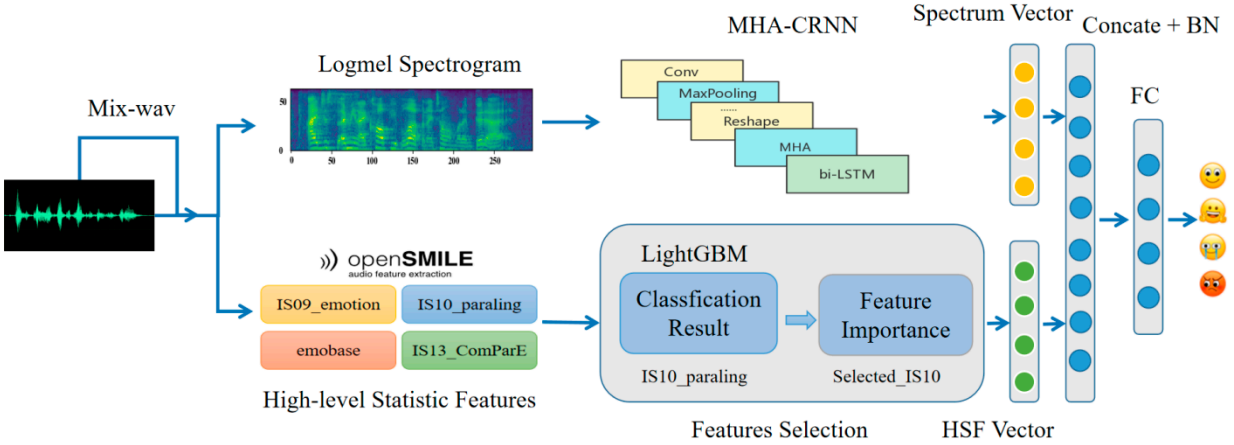

Figure 1. Proposed method. Includes data enhancement, feature extraction, global feature selection, feature fusion and classification.

3.1. Mix-Wav

The Mixup can be demonstrated as:

$$X_{mix} = \lambda X_i + (1 - \lambda) X_j \quad (1)$$

$$Y_{mix} = \lambda Y_i + (1 - \lambda) Y_j \quad (2)$$

The meaning of each variable is as follows:

- X_i and Y_i are the ith input and label of the data sample;
- X_{mix} and Y_{mix} represent the mixed data composed of a pair of original data samples;
- $\lambda \sim \text{Beta}(\alpha, \beta)$, $\alpha, \beta \in (0, \infty)$, λ and $1 - \lambda$ are a set of correlation weights used in Figure 2.

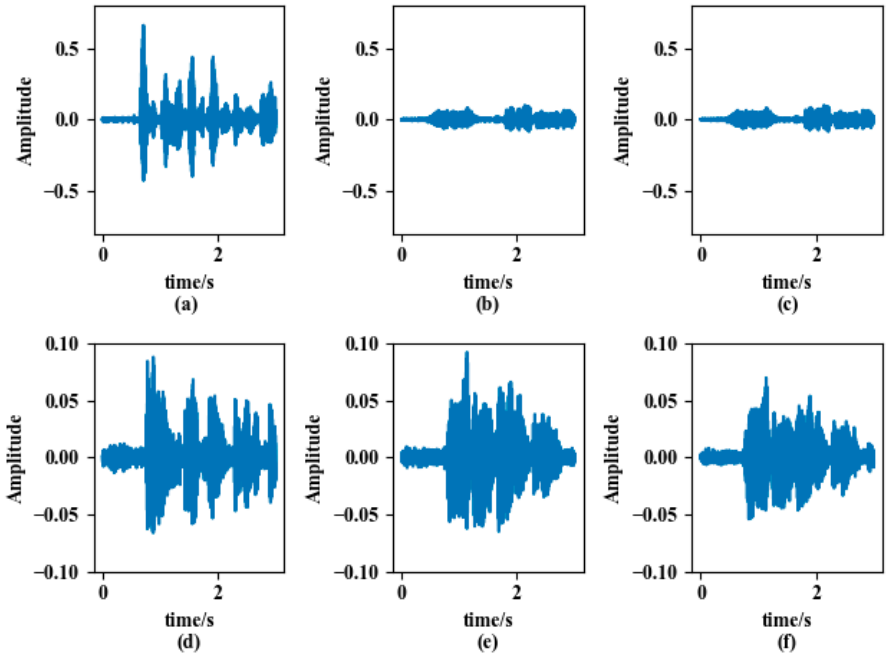

Figure 2. Two categories of emotional speech and Mix-wav-enhanced speech. (**a**,**b**) are the emotional speech waveforms of two neutral categories, (**c**) is the audio waveform enhanced with Mix-wav, where λ is 0.9660, so the generated speech (**c**) comes from (**a**) has more ingredients. (**d**–**f**) are the emotional speech of the sad category, using the same weight.

We control the contribution of the two samples to the enhanced sample's weight, typically $\alpha = \beta$; the hyperparameter α controls the strength of the interpolation between sample-enhanced samples.

Mixup has proved its effectiveness in many tasks of speech data augmentation tasks, but its objects are all two-dimensional spectral image features, based on which data augmentation is performed. No work has yet demonstrated the effectiveness of data augmentation on raw audio or one-dimensional sequence data. In many scenarios, we need to enhance the original data. For example, the input of the audio feature processing tool openSMILE [42] is a wav file, which requires the Mixup method to be applied to one-dimensional audio sequence data for data augmentation.

In response to the above problems, this paper proposes a data augmentation method Mix-wav applied to emotional speech tasks: it inherits the calculation method of Mixup mixed data, and after identifying the mixing between different emotional classes from the sense of hearing, it will cause the problem of inaccurate emotional expression of sound samples. Here it is applied to two audio samples with the same emotional label for calculation, which means $Y_i = Y_j$. Only use formula (1) to mix the data of the audio sample. Two audio input samples are weighted and output using weighting coefficients λ and $1 - \lambda$ obeying beta distribution, where X_i and X_j are audio data of the same emotion category.

3.2. Structure of MHA-CRNN

This model is inspired by the CRNN model often used in SER research. The general practice is to use convolutional neural networks to extract high-level features in the spectrogram. The MHA-CRNN model here adds a multi-head attention mechanism to further extract the time domain. The salient emotional information in the frequency domain features then uses Bidirectional LSTM (Bi-LSTM) to process the time-series data and finally use several fully connected layers to complete the classification.

In our method, the original corpus is divided and filled into 3 s segments as input, Log-Mel spectrogram features are extracted, and then modified VGGish [43] is used to extract deep feature extraction from Log-Mel spectrograms. VGGish is a pre-training model on the large-scale audio classification task set Audioset [8] released by Google in 2018. The audio features obtained as a feature extractor have been proven to perform better than artificially designed features on a variety of audio tasks. To adapt to the SER task, a longer time audio input is used to retain more emotional information. In addition, the dataset samples used for speech emotion are usually small, which makes it easy to produce overfitting problems, Moreover, the output of the VGGish convolution module is maximized, and the extracted features are not compressed. In this paper, a dropout layer is added to the VGGish model to avoid overfitting. The flatten layer and the full connection layer after the last pooling layer of the model is removed.

After that, a multi-head attention layer is added to project the high-level features extracted by CNN to the attention subspace vectors of different attention heads, learn the salient emotional representation between features from the temporal and spatial dimensions, and jointly output each attention head for further extracted features.

Finally, the salient emotional features extracted by the attention layer are input into two Bi-LSTM with a length of 128 cells to process the timing information and output the hidden state of the last time step. After that, it is connected to a fully connected layer with 256 nodes, and finally the output is classified using a SoftMax layer. Figure 3 is a schematic diagram of the proposed sentiment classification model MHA-CRNN framework.

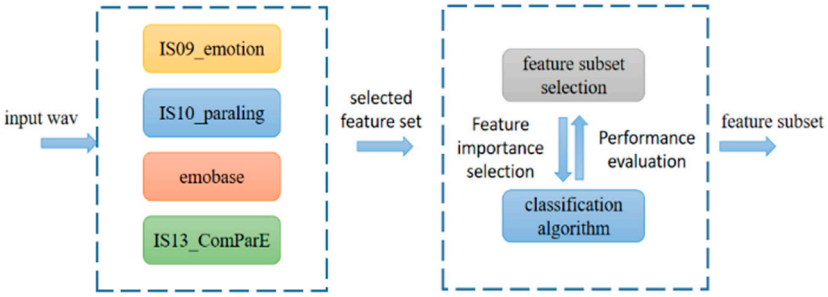

Figure 3. Emotion recognition model MHA-CRNN framework. Contains three large convolutional pooling blocks, two dropout layers, a multi-head attention layer, a Bi-LSTM, and a fully connected layer.

3.3. Fusion Global Feature Selection

In this part, we choose to use LightGBM, which is a framework for implementing the Gradient Boosting Decision Tree (GBDT) algorithm. Before LightGBM was put forward, the most famous GBDT tool was XGBoost, which is a decision tree algorithm based on the pre-sorting method. However, it costs a lot in time and space. LightGBM supports efficient parallel training, and has the advantages of faster training speed, lower memory consumption, better accuracy, distributed support and the rapid processing of massive data.

First, the IS09_emotion, IS10_paraling, IS13_ComParE, and emobase feature sets of IEMOCAP are extracted using openSMILE, and the configuration file is modified to make

the frame length 25 ms and the frame shift 10 ms, which is consistent with the division of the Log-Mel feature.

Then, the classification experiment is carried out on the extracted four global feature sets using the LightGBM [44] classification algorithm, and the feature set with the best classification performance and lower dimensionality is selected by comparison. Then, the selected feature set is used for the five-fold cross-validation classification experiment. The two feature importance indexes—"split" of splitting times and "gain" of splitting information gain on the nodes in each experiment—are accumulated as the feature selection criteria, and the feature selection experiment is carried out.

Then, we use the selected IEMOCAP feature set data and LightGBM to conduct a 50% cross-emotion classification experiment. We accumulate the feature importance scores obtained each time according to the above two feature importance indexes, and finally obtain the feature importance score ranking of each feature. We select the N features with the largest score as the subset after feature selection, and then train a LightGBM classifier for experiments to compare the performance of each feature subset after selection. According to the feature set performance and feature dimension, we select the feature set subset that can effectively reduce feature redundancy. Table 1 shows the overall experimental framework.

Table 1. Phase feature selection method based on LightGBM.

Input Log-Mel Spectrogram [Num_Frames, 64]
Conv2D 3 × 3@64 ReLu
MaxPooling2D 2 × 2
Conv2D 3 × 3@128 ReLu
MaxPooling2D 2 × 2
Dropout (0.25)
(Conv2D 3 × 3@256 ReLu)*2
MaxPooling2D 2 × 2
Dropout (0.25)
(Conv2D 3 × 3@512 ReLu)*2
MaxPooling2D 2 × 2
Multi-Head Attention 8 × 512
Bi-LSTM(2 × 128)
Dense 256

3.4. Feature Fusion

In many tasks, fusing features of different scales is an important means to improve the performance of the model. Relying on a single feature-extraction method, the effect is often not satisfactory. Due to the particularity of speech signal processing, the methods of feature extraction are different, and the feature granularity and focus are also different, such as frequency domain and time-domain features, frame level, global statistical features, and so on. Because of different feature sources, the focus of feature representation of emotional information is also different. The network model using a single feature input may bring the problem of insufficient use of the emotional information of the original speech signal.

To solve the above problems, this section integrates the emotion classification model based on the Multi-Head Attention (MHA-CRNN) mechanism proposed above with the global statistical features extracted by the stage feature selection method of the popular machine-learning method LightGBM, to improve the speech emotion recognition effect of the overall model, as shown in Figure 1.

The purpose of the feature fusion model is to combine the advantages of different features to complement each other. In this paper, two-dimensional Log-Mel spectrum and one-dimensional global statistical features are used. The two-dimensional Log-Mel spectrum takes continuous frames as units, extracts the spectrum characteristics of each frame of audio, and represents the spectrum space information of a whole sentence. One-

dimensional global statistical features, which are the prosodic features and sound quality features of each frame of audio in a sentence, complement each other.

One of the most commonly used methods of feature fusion is splicing or addition. For example, in ResNet and FPN, the method of adding elements is used for feature fusion, while in DenseNet, the method of splicing is used for feature fusion [45]. Aiming at the characteristics that the two features complement each other, this paper adopts the method of splicing the global statistical features through linear layer mapping and the output of MHA-CRNN, a deep-learning model based on multi-head attention mechanism. To make the input fusion features have the same contribution to the final classification as the output of MHA-CRNN in this step, the number of linear layer units after fusion features is 256, which is the same as the dimensions of the Bi-LSTM output vector.

4. Experiment Setup

4.1. Environmental Dependence

The experiment runs on Keras v.2.3.1 which using TensorFlow v.1.13.1 as backend, NVIDIA RTX 2060 and the machine toolkit is LightGBM.

4.2. Dataset Augmentation

Mix-wav has an important parameter controlling the mixing strength of samples α, the range of α is $[0, \infty]$. Here, the authors in [15] proved $\alpha = 0.2$ can bring better performance than other values, and two papers [15,30] applying Mixup method have used it $\alpha = 0.2$, which is considered to be the optimal α parameter.

In the specific data augmentation method, a generation multiple is used to control the number of samples generated by mixing. Two samples are randomly selected from the same category of audio each time, and Mix-wav is used to generate an enhanced emotional speech data of the same category. In addition to data augmentation, Mix-wav is also used for training sample category balance. For the emotion class with few samples in the training data, the generation multiple greater than that of other classes is used, so that each type of sample in the final augmentative dataset basically achieves balance. on IEMOCAP, there are relatively few samples of sadness and anger categories. We set a larger generation multiple for these two types of samples, so that the total amount of training data after the initial training data of each emotion category plus Mix-wav enhancement is roughly the same.

Among them, there are 5531 $\times$ 0.8 = 4424 training samples of IEMOCAP, and 6400 $\times$ 0.8 = 5120 samples for CHSE-DB. When setting the number generated by Mix-wav, we use a value roughly equivalent to the total number of samples in the training set. Below is one of the experiments of the five-fold cross-validation, with the number of samples of each type in the training set before and after applying Mix-wav augmentation, as shown in Table 2. After the slash is our result after data enhancement by the Mix-wav.

Table 2. Comparison of data quantity before and after use Mix-wav.

	Neutral	Happy	Sad	Angry	Total
IEMOCAP	1375/2750	1303/2606	853/2559	893/2679	4424/10,594
CHSE-DB	1627/3254	1572/3144	1605/3210	1596/3192	6400/12,800

4.3. Evaluation Metrics

In the SER task, as with other multi-classification problems, the accuracy and confusion matrix are also used to evaluate the quality of the model.

IEMOCAP dataset is an unbalanced label dataset, but we use the data augmentation method to make the number of samples close to each other. Unweighted Average Accuracy (UAA) and Weighted Average Accuracy (WAA) are mainly used for evaluation. WAA is the weighted average of the accuracy of different emotion categories, and its weight is directly

proportional to the number of samples in each category [46] The UAA is the average of the accuracy of different emotion categories, and the weight of each category is equal.

$$WAA = \frac{\sum_{j=1}^{c} N_j \times Accuracy}{\sum_{j}^{C} N_j} \quad (3)$$

$$UAA = \frac{1}{C}\sum_{j=1}^{C} Accuracy_j \quad (4)$$

The meaning of variables in the formula is as follows:

- C represents the total number of emotion categories in the experiment;
- N_j represents the total number of samples of the JTH emotion category;
- $Accuracy_j$ represents the prediction accuracy of the JTH emotion category, which can be obtained by dividing the number of correct predictions by the total number of samples.

4.4. Parameter Setup

The experimental data processing method is as follows: speech greater than 3 s is randomly segmented into a segment of 3 s, samples less than 3 s are zero filled, the frame length is 25 ms, the frame shift is 10 ms, the number of Mel filter banks is 64, and the offset of 0.01 is added to make the extracted features of the filled part not 0, so as to obtain the Log-Mel spectrum with the input of [64,298] dimensions. The experiments use the stochastic gradient descent (SGD) optimizer with a learning rate of 0.001, learning rate decay (decay) of 10–6, momentum of 0.9, batch size of 32, the maximum number of training rounds (num_epochs) of 150, and the model with the highest UAA is saved during training. Each set of experiments was subjected to a five-fold cross-validation with a training, validation, and test set sample ratio of 7:1:2, and the UAA and WAA with confusion matrix were reported as the results for comparison and analysis.

5. Experiment Results

5.1. Corpus Data

5.1.1. IEMOCAP

IEMOCAP is a multimodal English speech emotion database recorded by the University of Southern California. It has emotion data in video and audio formats [5]. It contains five sessions. Each session is recorded by a pair of speakers (one male and one female) in scripted and improvised performance. At the same time, two emotion models are used for labeling: two-dimensional activation valence emotion model and discrete emotion model. The discrete emotion categories include anger, happiness, excitement, sadness, disappointment, fear, surprise, neutral and other nine categories, including 10,039 audio samples, with an average duration of 4.5 s and a sampling rate of 16 KHz.

In this paper, the improvisation and performance part of the IEMOCAP dataset is used. The emotional categories total 5531 sentences of neutral, excited, sad and angry.

5.1.2. CHSE-DB

CHSE-DB [47] is a speech emotion dataset recorded by the State Key Laboratory of Media Integration and Communication of Communication University of China. Four postgraduates (two men and two women) majoring in broadcasting and anchoring of Communication University of China, according to 500 different content texts, express seven levels of emotions: neutral, happy, very happy, sad, very sad, angry and very angry in the way of broadcasting performance. There are 2000 audio in each category, and 14,000 clips in total.

5.2. Mix-Wav

In this section, the training data are processed with Mix-wav and without Mix-wav, respectively. The proposed MHA-CRNN model is used to classify the four emotions and conduct five-fold cross-validation. Figures 4 and 5 show the test UAA of IEMOCAP and CHSE-DB.

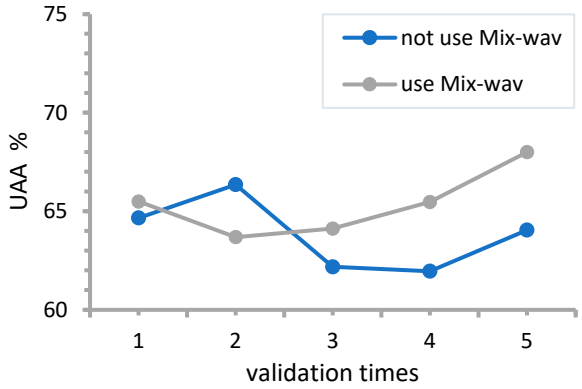

Figure 4. Comparison of tests before and after using Mix-wav on IEMOCAP.

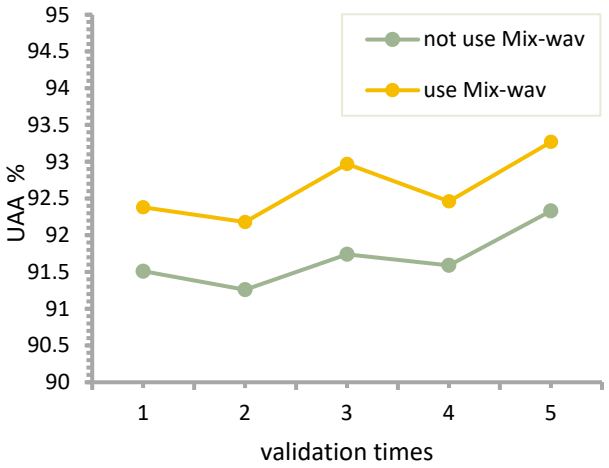

Figure 5. Comparison of tests before and after using Mix-wav on CHSE-DB.

As can be seen from Figure 4, in the two-fold cross-validation, there is a set of abnormal data, and the training without data augmentation achieved 66.35% accuracy in the test, which is different from the results of the other four experiments. The possible reason for this is that the total sample of the dataset is small, and when the data are divided, they deviate from the overall sample distribution, and the internal variance of the training set is larger than that of the validation set, resulting in a larger error. Therefore, in the calculation, we counted the unweighted test accuracies the remaining four times. The average correct rates of the tests with and without Mix-wav on the IEMOCAP training set were obtained as 65.77% and 63.21%, respectively, an improvement of 2.56%, and the classification confusion matrix for one of the sets of experiments is reported as a comparison, as shown in Figure 5.

Furthermore, it can be seen from Figure 5 that the test results after data enhancement with Mix-wav are better than those without Mix-wav. The average accuracy of the five tests on CHSE-DB with and without Mix-wav was 92.65% and 91.68%, respectively, with an average improvement of nearly 1%.

Tests on two datasets, IMEMOCAP and CHSE-DB, show that the Mixup method can be well migrated to audio data enhancement, and the improved Mix-wav shows better recognition performance in both databases.

5.3. Fusion Feature Selection

In the first stage, we first extracted the IS09_emotion, IS10_paraling, IS13_ ComParE and emobase four global feature sets, and the LightGBM algorithm with the same configuration is used to conduct a five-fold cross-validation emotion classification experiment on each feature set. According to the average test accuracy of the five-fold classification experiments, the feature set with the best classification result is selected, and the results are shown in Figure 6.

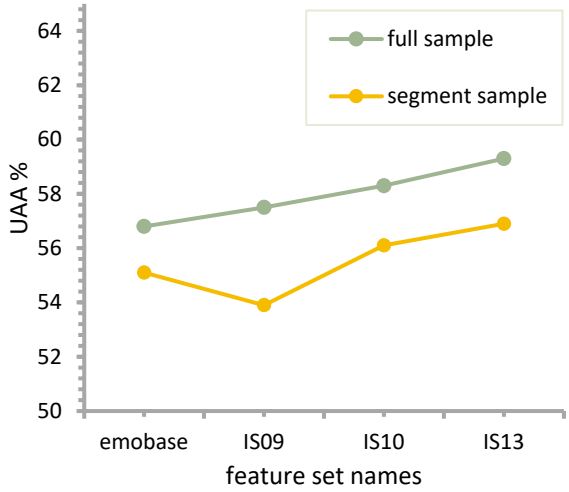

Figure 6. Comparison of four global feature set classification results.

The IS13_ComParE feature set achieves the highest recognition accuracy when two different data are used as input, and the IS10_paraling feature set comes second, but the difference between its recognition ability and that of IS13_ComParE feature set is not much, within 1% in both cases. Considering that the feature dimensionality of IS13_ComParE is much larger than that of IS10_paraling, the relative performance improvement is not much, and IS10_paraling is finally selected as the fused global feature set here.

In the second stage, we extracted the IS10_paraling feature set on the IEMOCAP dataset and used the following two statistical values as the method for feature importance selection while performing classification experiments on the LightGBM model: (1) "split": follow the split node when using a feature as a number of split attributes. (2) "gain": the total information gain that the feature brings to the split. A five-fold cross-validation was performed to accumulate the importance of the features obtained in the above two ways for each classification experiment. Here, 1500, 1200, and 800 features with the largest values were selected as the feature subset dimension for feature selection, and the five-times feature importance accumulation values were sorted from largest to smallest. The feature subsets were selected using feature subscripts of the first 1500, 1200, and 800 statistical values obtained, the three feature subsets obtained from feature selection were used again for classification experiments, and the experimental results are shown in Figure 7.

The original IS10_paraling feature set of 1582-dimensional features is known to achieve an average recognition accuracy of 58.26% on IEMOCAP. It can be seen that after two methods of feature selection, the recognition accuracy of the original feature set is basically maintained using its 1500-dimensional and 1200-dimensional feature subsets. In addition, the 1200-dimensional feature subset obtained using the "split" filtering method achieves 59.3% recognition accuracy, which is 1% higher than the average accuracy of the original feature set. The 1200-dimensional feature subset obtained by the above method is used as the result of the feature selection experiment.

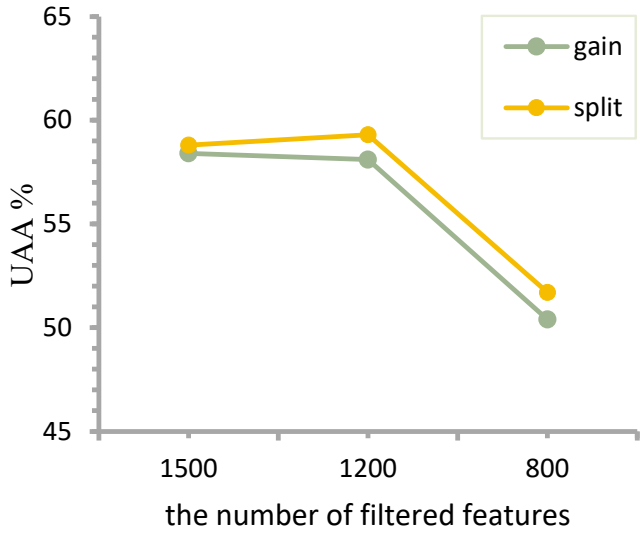

Figure 7. Classification performance of feature subsets obtained using two feature selection methods.

We compare and select the global features, and determine the features with better relative performance and lower dimensions, which is beneficial to the subsequent feature fusion step and can improve the model recognition performance. This will be shown in detail later.

5.4. MHA-CRNN and Feature Fusion Model

(1) A key parameter of the multi-head attention mechanism is the number of attention heads, which determines how many linear transformations and self-attention operations are performed on the input, and the core the determination of the best num_ heads. To verify the effect of num_heads on model performance, four values of 1, 4, 8, and 16 are selected for the experiments, and the size of each head is 128, 256, 512 and 1024, to be consistent with the input vector dimension before transformation. We use the IEMOCAP on four types of samples—neutral, happy, sad, and angry—with a total of 5531 data, and the training set, validation set, and test set in the ratio of 7:1:2. We conducted a five-fold crossover experiment, and reported the UAA for the five experiments. A comparison of the experimental results is shown in Figure 8.

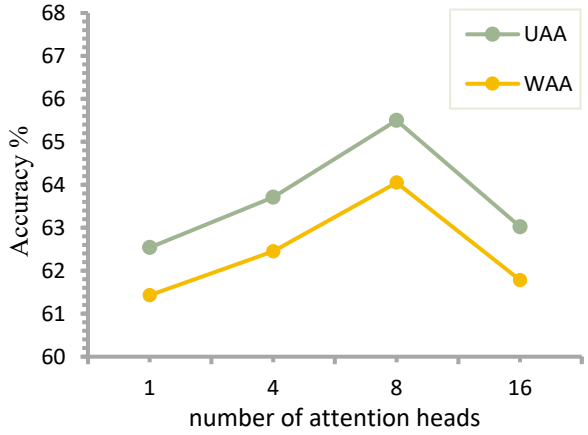

Figure 8. Comparison of the number of different attention heads in the MHA-CRNN structure.

The model achieves the highest recognition rate when the number of attention heads is 8, and the lowest when the number of attention heads is 1. When the number of multiple attention heads is 1, it is equivalent to the original self-attention model. It can be seen

that the multi-head attention mechanism has achieved better results than the single head self-attention mechanism. When the number of heads is 8, the highest recognition accuracy is 65.5%, and when the number of heads is 16, the lowest is 63.02%, but both are higher than the recognition accuracy of 62.54% obtained by single head self-attention mechanism.

Moreover, increasing the number of attention heads does not always lead to performance improvement. As the number of attention heads increases from 4 to 16, the experimental results show an increasing and then decreasing trend, and the performance is optimal at eight heads. This indicates that more heads are not better, and that there is a saturation value for dividing the attention subspace (number of heads), and that too many or too few heads can affect performance. The number of attention heads determines the number of spatial vectors for linear transformations of the input. The larger the number, the more attention analysis can be performed on the input from more scales, but too many spatial transformations will affect the retention of the original feature information of the input, so choosing the appropriate number of attention heads is an important step in adjusting the attention parameters.

(2) Based on the above experiments, after determining that the optimal number of attention heads is equal to 8, the influence of the size of attention head (head_size) on the performance of the model is further explored. Four attention head sizes are selected for the experiment. The attention head size determines the length of the linear mapping of the input vector in the multi-head attention mechanism. Finally, the length of the output of the multi-head attention mechanism is equal to the size of the attention head multiplied by the number of heads. The experimental comparison of different attention head sizes is shown in Figure 9.

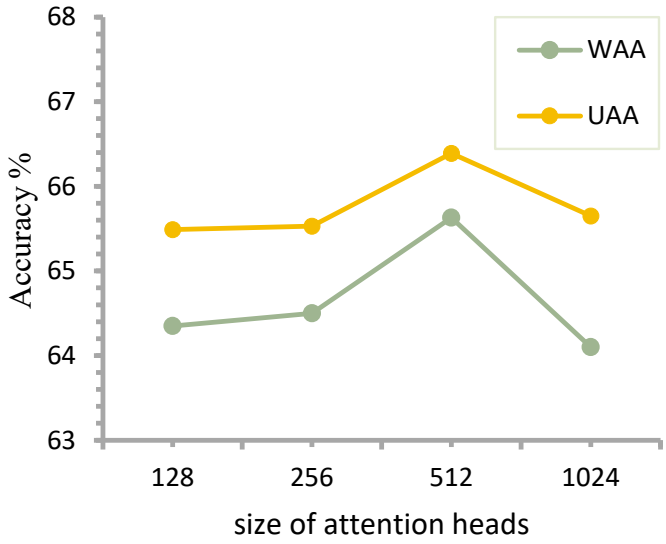

Figure 9. Comparison of different attentional head sizes in MHA-CRNN structures.

Here, the input of the multi-head attention mechanism layer is [19,1024], When the number of heads is 8, the head size is 128 so that the input and output dimensions of the multi-head attention mechanism layer are equal. When head size is larger than 128, the later Bi-LSTM will obtain longer input of time-series sentiment information. As shown in Figure 9, the classification results obtained when the number of attention heads is 256, 512, and 1024 are all better than the case when the head size is 128. However, the increase of attention head size does not always lead to performance improvement, and either too large or too small will lead to performance variation. This proves that choosing the right attention head size is also an important step in tuning the attention parameters. The size of the attention head reflects the dimensionality of the attention vector obtained by linearly mapping the input vector. A small dimensionality of the mapping may result in the loss of

feature information, while too high a dimensionality will make the representation of the original emotional input information sparse and make it difficult for the attention head to focus its attention.

(3) To compare the proposed MHA-CRNN model as well as the fusion model, this section uses the above two models with the affective speech data enhancement method Mix-wav and conducts a comparison experiment on the speech emotion benchmark dataset IEMOCAP. Here, the multi-head attention parameters are num heads equal to 8 and head size equal to 512. The fusion input in the fusion model is the 1200-dimensional IS10_paraling global features after feature selection. Figure 10 shows the test UAA of each of the five-fold cross-validation experiments for both models.

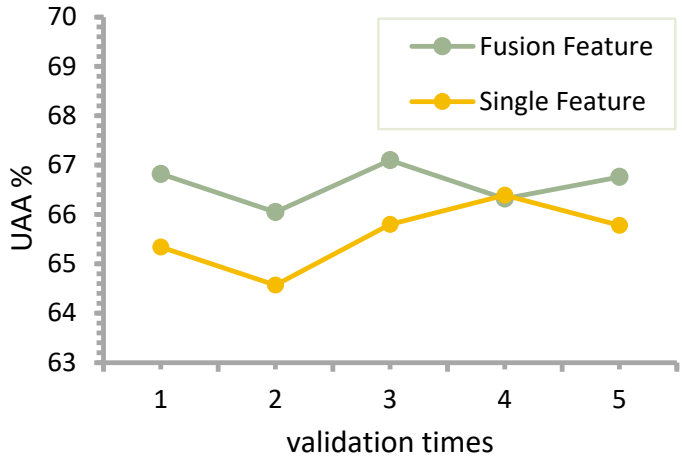

Figure 10. Comparison of MHA-CRNN and its feature fusion model in IEMOCAP.

(4) After a series of different experimental comparisons and parameter adjustments, we achieved a result of 66.44% on the common four-classification task on the benchmark dataset IEMOCAP, which is better than most research methods using similar models. To compare with more studies in this paper, we report the UAA and WAA methods of the MHA-CRNN in IEMOCAP database and the feature fusion model based on MHA-CRNN, as shown in Table 3.

Table 3. The four categories of tasks on the dataset IEMOCAP are compared with other works.

Method	UAA	WAA%	Year
CRNN + Attention [21]	64.74	/	2018
CNN + LSTM [48]	61.7	/	2018
SPP + LSTM + fusion [49]	64.6	64.5	2019
Blocked-CRNN [50]	62.3	/	2020
Selective MTL [51]	59.47	56.87	2022
MHA-CRNN	65.57	64.45	2022
MHA-CRNN + fusion	66.44	65.17	2022

This shows the confusion matrix of the best one-time classification results on CHSE-DB and IEMOCAP, as shown in Figure 11. It can be seen that a large amount of data falls on the diagonal of the confusion matrix, i.e., the model classification is mostly correct. The darker the color on the diagonal, the better the task prediction for the category. With the method described in this paper, on the IEMOCAP dataset, the anger class has achieved the highest recognition rate, with a rate of 73.1%. It is difficult to distinguish between neutral and happy emotions, and the probability of mis-classification. One possible reason is the corpus sample problem, and the other is that the model has not yet extracted discriminative emotional features, which needs to be further improved. On the CHSE-DB dataset, the

classification performance is significantly better, probably due to the fact that the vocalists of the CHSE-DB dataset are professional performers and the quality of the sample recordings is better than that of IEMOCAP. The performance on both Chinese and English speech emotion datasets proves the effectiveness of the proposed method.

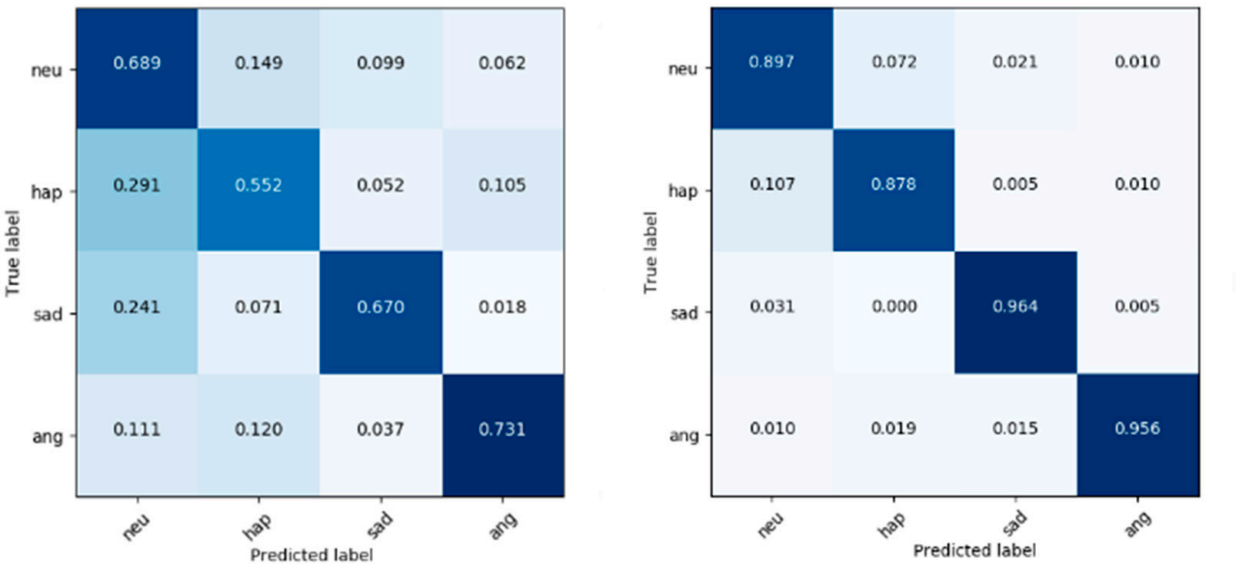

Figure 11. Confusion matrix for IEMOCAP and CHSE-DB Dataset.

6. Conclusions

In this paper, an emotion recognition method based on speech data and feature fusion is proposed. It has been verified on the public IEMOCAP and the self-built CHSE-DB dataset. First, in view of the small scale of each speech emotion dataset and the uneven classification of samples, the Mixup method in the field of image recognition is migrated and applied to one-dimensional audio sequence data. Then, on the one hand, to solve the problem that the optimal speech emotion feature set is still unclear, and in order to extract more representative fusion features, the LightGBM method is used to screen several global feature sets. On the other hand, based on the conventional use of the convolutional neural network, the multi-head attention mechanism is used to learn the emotional salient features, and then the long short-term memory network is used to analyze the time series. The information is analyzed and output. Finally, by fusing the extracted global statistical feature vector and the time–frequency domain vector, the emotion expression features from the two granularities are complemented, and the emotion discrimination ability of the model is effectively improved. The unweighted average test accuracy of 66.44% and 93.47% was obtained on the IEMOCAP and CHSE-DB dataset, respectively. According to the experimental results, our proposed method has higher accuracy compared with previous studies in the literature. Our proposed data enhancement method is widely applicable to multi-label audio databases and can alleviate the problems of insufficient and unbalanced data. At present, we are still lacking in the optimal set and feature fusion of emotional acoustic features, which will be the focus of our next work. In the future, we will also consider improving the network structure to improve the recognition rate and generalization ability. Considering the important position of emotion recognition in human–computer interaction, the improvement of the speed of system recognition is also a problem we need to solve in the future.

Author Contributions: B.L. conducted the research study, and wrote the paper; W.Z. helped to edit the paper; Z.T. made suggestions to this paper and guided the research study. R.Y. and Y.Z. Participated in the revision and inspection of the article. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Research Funds for the Central Universities.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data and materials are available upon request.

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

References

1. Akçay, M.B.; Oğuz, K. Speech Emotion Recognition: Emotional Models, Databases, Features, Preprocessing Methods, Supporting Modalities, and Classifiers. *Speech Commun.* **2020**, *116*, 56–76. [CrossRef]
2. 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]
3. Lv, Z.; Poiesi, F.; Dong, Q.; Lloret, J.; Song, H. Deep Learning for Intelligent Human–Computer Interaction. *Appl. Sci.* **2022**, *12*, 11457. [CrossRef]
4. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A Database of German Emotional Speech. In Proceedings of the Interspeech 2005, Lisbon, Portugal, 4–8 September 2005; ISCA: Sydney, Australia; pp. 1517–1520. [CrossRef]
5. 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]
6. 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] [PubMed]
7. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Li, F.-F. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255. [CrossRef]
8. Gemmeke, J.F.; Ellis, D.P.W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 776–780. [CrossRef]
9. Ko, T.; Peddinti, V.; Povey, D.; Khudanpur, S. Audio Augmentation for Speech Recognition. In Proceedings of the Interspeech 2015, Dresden, Germany, 6–10 September 2015; ISCA: Sydney, Australia; pp. 3586–3589. [CrossRef]
10. Shahnawazuddin, S.; Dey, A.; Sinha, R. Pitch-Adaptive Front-End Features for Robust Children's ASR. In Proceedings of the Interspeech 2016, San Francisco, CA, USA, 8–12 September 2016; ISCA: Sydney, Australia; pp. 3459–3463. [CrossRef]
11. Tóth, L.; Kovács, G.; Van Compernolle, D. A perceptually inspired data augmentation method for noise robust cnn acoustic models. In Proceedings of the 20th International Conference, SPECOM 2018, Leipzig, Germany, 18–22 September 2018; pp. 697–706. [CrossRef]
12. Lieskovská, E.; Jakubec, M.; Jarina, R.; Chmulík, M. A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism. *Electronics* **2021**, *10*, 1163. [CrossRef]
13. Xu, M.; Zhang, F.; Cui, X.; Zhang, W. Speech Emotion Recognition with Multiscale Area Attention and Data Augmentation. *arXiv* **2021**, arXiv:2102.01813.
14. Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; Plumbley, M.D. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *arXiv* **2020**, arXiv:1912.10211. [CrossRef]
15. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond Empirical Risk Minimization. *arXiv* **2018**, arXiv:1710.09412.
16. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.-C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 2613–2617. [CrossRef]
17. Sahidullah, M.; Saha, G. Design, Analysis and Experimental Evaluation of Block Based Transformation in MFCC Computation for Speaker Recognition. *Speech Commun.* **2012**, *54*, 543–565. [CrossRef]
18. Deng, L.; O'Shaughnessy, D. *Speech Processing: A Dynamic and Optimization-Oriented Approach*; CRC Press: Boca Raton, FL, USA, 2003.
19. Ravikumar, K.M.; Rajagopal, R.; Nagaraj, H.C. An Approach for Objective Assessment of Stuttered Speech Using MFCC Features. *Digit. Signal Process. J.* **2009**, *9*, 1687–4811.
20. Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2227–2231. [CrossRef]
21. Chen, M.; He, X.; Yang, J.; Zhang, H. 3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition. *IEEE Signal Process. Lett.* **2018**, *25*, 1440–1444. [CrossRef]
22. Jiang, P.; Fu, H.; Tao, H.; Lei, P.; Zhao, L. Parallelized Convolutional Recurrent Neural Network With Spectral Features for Speech Emotion Recognition. *IEEE Access* **2019**, *7*, 90368–90377. [CrossRef]

23. Yan, M.; Lou, X.; Chan, C.A.; Wang, Y.; Jiang, W. A semantic and emotion-based dual latent variable generation model for a dialogue system. *CAAI Trans. Intell. Technol.* **2023**, 1–12. [CrossRef]
24. Yan, M.; Xiong, R.; Shen, Y.; Jin, C.; Wang, Y. Intelligent generation of Peking opera facial masks with deep learning frameworks. *Herit. Sci.* **2023**, *11*, 20. [CrossRef]
25. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
26. Sitaula, C.; He, J.; Priyadarshi, A.; Tracy, M.; Kavehei, O.; Hinder, M.; Withana, A.; McEwan, A.; Marzbanrad, F. Neonatal Bowel Sound Detection Using Convolutional Neural Network and Laplace Hidden Semi-Markov Model. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 1853–1864. [CrossRef]
27. Xie, Y.; Liang, R.; Liang, Z.; Huang, C.; Zou, C.; Schuller, B. Speech Emotion Classification Using Attention-Based LSTM. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1675–1685. [CrossRef]
28. Lian, Z.; Liu, B.; Tao, J. CTNet: Conversational Transformer Network for Emotion Recognition in IEEE/ACM Transactions on Audio. *Speech Lang. Process.* **2021**, *29*, 985–1000. [CrossRef]
29. Al-onazi, B.B.; Nauman, M.A.; Jahangir, R.; Malik, M.M.; Alkhammash, E.H.; Elshewey, A.M. Transformer-Based Multilingual Speech Emotion Recognition Using Data Augmentation and Feature Fusion. *Appl. Sci.* **2022**, *12*, 9188. [CrossRef]
30. Nwe, T.L.; Foo, S.W.; Silva, L.C. Speech emotion recognition using hidden Markov models. *Speech Commun.* **2003**, *41*, 603–623. [CrossRef]
31. Ververidis, D.; Kotropoulos, C. Emotional Speech Classification Using Gaussian Mixture Models and the Sequential Floating Forward Selection Algorithm. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6–9 July 2005; IEEE: New York, NY, USA, 2005; pp. 1500–1503. [CrossRef]
32. Chavhan, Y.; Dhore, M.L.; Yesaware, P. Speech Emotion Recognition Using Support Vector Machine. *IJCA* **2010**, *1*, 8–11. [CrossRef]
33. Meng, L.; Xu, J.; Tan, X.; Wang, J.; Qin, T.; Xu, B. MixSpeech: Data Augmentation for Low-Resource Automatic Speech Recognition. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021.
34. Min, F.; Hu, Q.; Zhu, W. Feature Selection with Test Cost Constraint. *Int. J. Approx. Reason.* **2014**, *55*, 167–179. [CrossRef]
35. Tzirakis, P.; Trigeorgis, G.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1301–1309. [CrossRef]
36. Lee, J.; Tashev, I. High-Level Feature Representation Using Recurrent Neural Network for Speech Emotion Recognition. In Proceedings of the INTERSPEECH, Dresden, Germany, 6–10 September 2015.
37. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
38. Nediyanchath, A.; Paramasivam, P.; Yenigalla, P. Multi-Head Attention for Speech Emotion Recognition with Auxiliary Learning of Gender Recognition. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7179–7183. [CrossRef]
39. Schuller, B.; Steidl, S.; Batliner, A. The Interspeech 2009 Emotion Challenge. In Proceedings of the INTERSPEECH, Brighton, UK, 6–10 September 2009; pp. 312–315. [CrossRef]
40. Schuller, B.; Steidl, S.; Batliner, A.; Burkhardt, F.; Devillers, L.; Müller, C.; Narayanan, S.S. The INTERSPEECH 2010 Paralinguistic Challenge. In Proceedings of the Interspeech 2010, Chiba, Japan, 26–30 September 2010; pp. 2794–2797. [CrossRef]
41. Schuller, B.; Steidl, S.; Batliner, A.; Vinciarelli, A.; Scherer, K.; Ringeval, F.; Chetouani, M.; Weninger, F.; Eyben, F.; Marchi, E.; et al. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In Proceedings of the INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, 25–29 August 2013. [CrossRef]
42. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Ottawa ON, Canada, 25–29 October 2010. [CrossRef]
43. Hershey, S.; Chaudhuri, S.; Ellis, D.P.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Plakal, M.; Platt, D.; Saurous, R.A.; Seybold, B.; et al. CNN Architectures for Large-scale Audio Classification. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 131–135. [CrossRef]
44. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
45. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 2261–2269. [CrossRef]
46. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
47. Tu, Z.; Liu, B.; Zhao, W.; Cao, B. Establishment of Chinese Speech Emotion Database of Broadcasting. In Proceedings of the 2021 International Conference on Culture-oriented Science & Technology (ICCST), Beijing, China, 18–21 November 2021; pp. 603–606. [CrossRef]
48. Etienne, C.; Fidanza, G.; Petrovskii, A.; Devillers, L.; Schmauch, B. CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation. *arXiv* **2018**, arXiv:1802.05630.

49. Zeng, Y.; Mao, H.; Peng, D.; Yi, Z. Spectrogram based multi-task audio classification. *Multimed Tools Appl.* **2019**, *78*, 3705–3722. [CrossRef]
50. Zhou, H.; Liu, K. Speech Emotion Recognition with Discriminative Feature Learning. In In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 4094–4097. [CrossRef]
51. Zhang, H.; Mimura, M.; Kawahara, T.; Ishizuka, K. Selective Multi-Task Learning For Speech Emotion Recognition Using Corpora Of Different Styles. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 7707–7711. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.