electronics

MDPI

Article

Speech Emotion Recognition Based on Deep Residual Shrinkage Network

Tian Han [1,2], Zhu Zhang [1,2,*], Mingyuan Ren [1], Changchun Dong [1], Xiaolin Jiang [1] and Quansheng Zhuang [2]

[1] Department of Artificial Intelligence, Jinhua Advanced Research Institute, Jinhua 321013, China; htopen@foxmail.com (T.H.); rmy2000@126.com (M.R.); hitdongcc@163.com (C.D.); jlynner@163.com (X.J.)
[2] School of Measurement and Communication Engineering, Harbin University of Science and Technology, Harbin 150080, China; izqshust@tom.com
* Correspondence: zhuzhang.zz@foxmail.com

Abstract: Speech emotion recognition (SER) technology is significant for human–computer interaction, and this paper studies the features and modeling of SER. Mel-spectrogram is introduced and utilized as the feature of speech, and the theory and extraction process of mel-spectrogram are presented in detail. A deep residual shrinkage network with bi-directional gated recurrent unit (DRSN-BiGRU) is proposed in this paper, which is composed of convolution network, residual shrinkage network, bi-directional recurrent unit, and fully-connected network. Through the self-attention mechanism, DRSN-BiGRU can automatically ignore noisy information and improve the ability to learn effective features. Network optimization, verification experiment is carried out in three emotional datasets (CASIA, IEMOCAP, and MELD), and the accuracy of DRSN-BiGRU are 86.03%, 86.07%, and 70.57%, respectively. The results are also analyzed and compared with DCNN-LSTM, CNN-BiLSTM, and DRN-BiGRU, which verified the superior performance of DRSN-BiGRU.

Keywords: speech emotion recognition; mel-spectrogram; DRSN; BiGRU

Citation: Han, T.; Zhang, Z.; Ren, M.; Dong, C.; Jiang X.; Zhuang, Q. Speech Emotion Recognition Based on Deep Residual Shrinkage Network. *Electronics* **2023**, *12*, 2512. https://doi.org/10.3390/electronics12112512

Academic Editor: Manohar Das

Received: 6 May 2023
Revised: 27 May 2023
Accepted: 31 May 2023
Published: 2 June 2023

1. Introduction

Communication through language is the most fundamental and significant aspect of human interaction. Emotional communication in particular facilitates a deeper, more nuanced understanding between individuals, leading to more precise and genuine expressions of feelings. Regrettably, this emotional connection is often absent in human–computer interaction, limiting the level of engagement and satisfaction in such scenarios. Consequently, it becomes critical to imbue computers with the capacity to comprehend human emotions conveyed through vocal expression.

The origins of research on emotion recognition of speech date back to the 1980s [1,2]. The conventional method for emotion recognition of speech involves the classification of speech emotions, which commonly employs classification methods, such as KNN [3], SVM [4,5], HMM [6], GMM [7], Decision Tree [8], and ANN [9] to name a few. These classification methods employ feature extraction of speech, with commonly used features falling into three categories, prosodic features, sound quality features, and spectral features. Prosodic features include pitch, short-term energy, and duration [10,11], while spectral features include MFCC, LPCC, LFPC, GFCC, Formants, etc. [12]. Sound quality features include jitter, shimmer, and HNR [13–15]. Research on speech emotion recognition has traditionally focused on the use of these features or optimization of traditional models. In recent years, deep learning has emerged as a prominent alternative to the traditional models, showcasing superior performance across various fields, including the realm of speech emotion recognition. Recent applications of DNN, RNN, CNN, LSTM, and other network models have reaped fruitful outputs in voice emotion recognition [16,17].

Through speech emotion recognition technology, people can more conveniently interact with computers and achieve more intelligent services and management. At the

same time, voice emotion recognition technology can also help people better understand their emotional state, and timely detect and solve psychological problems. Therefore, the research and application of speech emotion recognition technology is of great significance for improving people's quality of life and work efficiency. At present, speech emotion recognition technology is not yet perfect, and there are two main areas that need to be improved which are searching for robust and universal emotional features of speech, and constructing strong models with high recognition accuracy. In this paper, a speech emotion recognition method is proposed which is composed of Mel-spectrogram feature extraction and Deep Residual Shrinkage Network with Bidirectional Gated Recurrent Unit (DRSN-BiGRU). In Section 3.1, the Mel-spectrogram is introduced in detail and adopted as a feature for speech emotion recognition, as it contains both time-domain and frequency-domain information of speech, and can be used as input for convolution layers. In Section 3.2, DRSN is proposed in detail which is used to recognize emotions from Mel-spectrogram. In this part, residual shrinkage building unit, soft thresholding and self-attention mechanism are introduced and analyzed, which are key components of the network. In Section 3.3, BiGRU is proposed as the last part of the model, which allows the network to not only obtain information from forward propagation, but also utilize information in the reverse direction, making more important features fully utilized. In order to verify the speech emotion recognition method, experiments on three emotional speech datasets (CASIA dataset, IEMOCAP dataset, and MELD dataset) are carried out and results are presented in Section 4. In Section 5, the speech emotion recognition method is discussed and compared with three other methods.

2. Related Work

Speech emotion recognition methods based on speech acoustic features typically extract some manually designed features from speech signals and feed these features into classifiers to complete recognition tasks. In 2018, Zhang used AlexNet designed for the ImageNet LSVRC-2010 to recognize emotion from speech, and in the research EMO-DB was used to verify the performance of the network. The accuracy of four emotions (angry, sad, happy, and neutral) are over 80%, which was about 20% higher than SVM. Then the same experiments were carried out in three other databases (RML, Enterface05, and Baum-1s), and AlexNet all got better results than SVM [18]. In 2019, Sun introduced a novel approach to recognize speech emotion utilizing a DNN-decision tree-SVM model. DNNs extract bottleneck features that are then employed to train each SVM in the decision tree. The findings of the experiment indicate that the proposed technique yields an average emotion recognition rate 6.25% higher than the traditional SVM classification and 2.91% better than the DNN-SVM classification [19]. In 2019, Zhao proposed a novel approach to emotion recognition from speech using two CNN and LSTM networks, one 1D CNN LSTM network and one 2D CNN LSTM network. These networks were specifically designed to learn both local and global emotional features from speech and log-Mel spectrogram data. The research found that the hybrid model could leverage the strengths of both types of networks while overcoming the limitations of both [17]. In 2020, Huang analyzed the differences in speech signal features of different emotional categories and ranked the contribution of features. The study found that the fundamental frequency (F0), sound intensity, and duration have a greater contribution to emotional categories [20]. In 2020, Atmaja used an 88-dimensional Ge MAPS feature set as the input feature of the Long Short-Term Memory (LSTM) network, and discussed the impact of different loss functions on the performance of dimensional emotion recognition [21]. In 2021, Cai proposed a multi task learning framework that simultaneously performs automatic speech recognition and speech emotion classification tasks. This method is an end-to-end deep neural network model for speech emotion recognition based on wav2vec [22]. In 2020, Yeh used end-to-end ASR to extract ASR-based representations for speech emotion recognition, and designed a decomposition domain adaptation method on a pre-trained ASR model to improve speech recognition rate and recognition accuracy of the target emotion corpus [23]. In 2020, Bakhshi used raw speech time-domain signals and frequency-domain information as

inputs to the deep Conv-RNN network, effectively extracting emotional representations of speech signals and achieving end-to-end dimensional emotion recognition [24]. In 2020, Sun used residual convolution neural networks to capture emotional information from raw speech signals and classify them. In addition, this method also incorporates speaker gender information to further improve recognition rate [25]. In 2020, a novel framework was introduced by Sajjad for detecting emotions from speech. The proposed methodology utilized key sequence segment selection based on a radial-based function network (RBFN) similarity measurement in clusters to arrive at the selected sequence. This sequence was converted into a spectrogram through the application of the STFT algorithm and subsequently fed into a CNN for feature extraction. To ensure precise recognition performance, normalization of CNN features was undertaken prior to their use in training a deep bi-directional long short-term memory (BiLSTM) model to learn temporal information for recognizing the final state of emotion [26]. In 2020, Issa introduced a novel architecture that utilizes a combination of Mel-frequency cepstral coefficients, chromagram, Mel-scale spectrogram, Tonnetz representation, and spectral contrast features to extract information from sound files for emotion recognition. The one-dimensional Convolution Neural Network is employed to classify the features, and an incremental method is employed to optimize the model for enhanced classification accuracy [27].

In recent studies, it has been further demonstrated that introducing attention mechanisms and using bidirectional long short-term memory networks can further improve recognition accuracy. In 2021, Wang used a multimodal transformer with shared weights to capture dependencies between modalities, which also achieved good results. Finally, the success of the pre-trained model in speech recognition tasks has led to its increasing attention in research on speech emotion recognition [28]. In 2022, Zou utilized the Wav2Vec pre-training model to extract deep features of speech and combined them with traditional acoustic features for emotion recognition [29]. The review of the above research is summarized in Table 1.

Currently, speech emotion recognition mainly relies on data-driven deep learning models. However, there are essential differences and inexplicability between the emotional representation space extracted by deep learning models and the human understandable space. The inexplicability of these deep learning models poses challenges for establishing secure and reliable speech emotion recognition tasks. In addition, emotion recognition involves knowledge from multiple disciplines, such as perception, cognition, and psychology. How to integrate these related knowledge into deep model algorithms is a major challenge in speech emotion recognition technology. Therefore, constructing a combination of knowledge driven and data driven speech emotion recognition will be another direction for the development of speech emotion recognition.

Table 1. Literature survey on techniques for speech emotion recognition.

Year and Author	Models	Datasets	Contributions
2018 Zhang [18]	DCNN	EMO-DB RML eNTERFACE BAUM	Application of AlexNet in SER
2019 Sun [19]	Decision Tree DNN SVM	CASEC	Bottlenect features extracted by DNNs
2019 Zhao [17]	LSTM CNN	EMO-DB IEMOCAP	1D and 2D DNN LSTM network
2020 Sun [25]	R-CNN	FAU eNTERFACE	End-to-end SER with RAW data

Table 1. *Cont.*

Year and Author	Models	Datasets	Contributions
2020 Huang [20]	NetVALD	IEMOCAP	Rank different features
2020 Atmaja [21]	LSTM	IEMOCAP	88-dimensional features
2021 Cai [22]	Wav2Vec	IEMOCAP	Multi task learning framework
2020 Yeh [23]	ASR model	IEMOCAP LibriSpeech	ASR based representation of speech
2020 Bakhshi [24]	DNN	RECOLA	End-to-end model of predicting continuous emotions
2020 Sajjad [26]	BiLSTM	EMO-DB IEMOCAP RAVDESS	Key sequence segment selection based on RBFN
2020 Issa [27]	CNN	EMO-DB IEMOCAP RAVDESS	A framework of features combination
2021 Wang [28]	Transformer	IEMOCAP	Multi modal transformer with sharing weights
2022 Zou [29]	CNN BiLSTM Wav2Vec2	IEMOCAP	Multi-level acoustic information with co-attention module

3. Materials and Methods

3.1. Mel-Spectrogram of Speech

Speech is a signal within the time domain that employs traditional features for recognition and classification. These features include short-term energy, zero-crossing rate, pitch, formant, MFCC, and LPCC, either in frequency or time domain. In order to produce superior results, it is necessary for the machine learning model to integrate both domain features and establish an appropriate trade-off. This presents significant challenges for model and feature selection. With the help of deep learning algorithms and their more intricate network structures, such as highlighting key features during the training process and filtering redundant information, more precise results can be produced. The spectrogram represents the speech signal in two dimensions of spectral information, which illustrates characteristics in both frequency and time domains. The Mel-filter function converts the energy spectrum to the Mel-frequency, which better imitates the human ear principle. By blending the spectrogram and Mel-filter advantages, this paper introduced the Mel-spectrogram feature for speech emotion recognition. Figure 1 depicts the feature extraction flow chart of Mel-spectrogram.

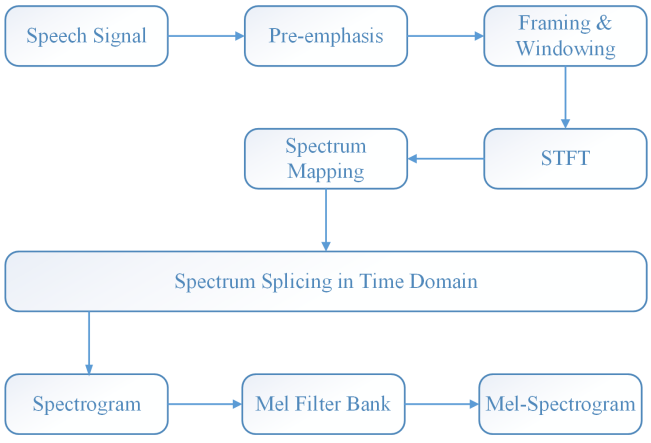

Figure 1. Flow chart of Mel-spectrogram extraction.

As shown in Figure 1, the feature extraction of Mel-spectrogram needs the following processes. (1) Pre-emphasis and frame-based windowing technology. The sampling frequency of voice signal is 16,000 Hz, the number of sampling points in each voice frame is N = 512, the pre-emphasis coefficient is 0.97, and the frame step is 200. (2) Fast Fourier transform. Each frame of speech signal divided by step (1) is very short, so it is considered that the frequency is in a stable state within a frame of time. The fast Fourier transform of each frame of speech signal $y_i(n)$ is calculated according to Equation (1) to obtain its spectrum. The multi-frame spectrum is spliced on the time dimension to form spectrogram.

$$F_{\mathrm{i}}(k) = \sum_{n=0}^{N-1} y_i^{'}(n) e^{\frac{-j2\pi}{N} kn} \tag{1}$$

The power spectrum of speech signal P is calculated by Equation (2).

$$P = \frac{|F_i(k)|^2}{N} \tag{2}$$

In which, $F_i(k)$ is the Fourier transform output of the ith frame of speech signal $y_i^{'}(n)$.

(3) Mel-filter bank. Mel-filter bank is used to convert frequency into Mel-frequency to simulate human perception. By passing the energy spectrum through multiple groups of Mel-filters bank, Mel-filters bank can approximate the human auditory system and have a certain choice of frequency, which can directly ignore the frequency signal that you do not want to perceive. The structure of Mel-filter is shown in Figure 2. Mel-spectrogram can be obtained from the spectrogram after passing through Mel-filter bank.

In this paper, the number of Mel-filter banks M = 40 is selected, and the response at the center frequency is 1, which gradually decreases to 0 at both ends, until the center frequency of two adjacent filters is reached, where the response is 0, and the interval between them increases with the increase of m value. Mel-filter banks can be constructed from Equation (3). The energy spectrum P of step (2) is passed through the Mel-filter bank $H_m(k)$ to obtain the Mel-spectrogram, which is shown in Figure 3.

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le k < f(m) \\ 1 & k = f(m) \\ \frac{f(m+1) - k}{f(m+1) - f(m)} & f(m) < k \le f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{3}$$

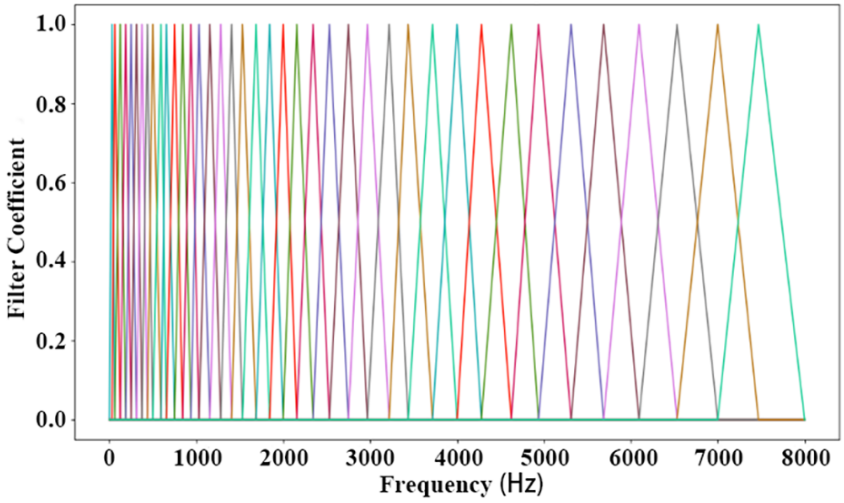

Figure 2. Mel-filter bank.

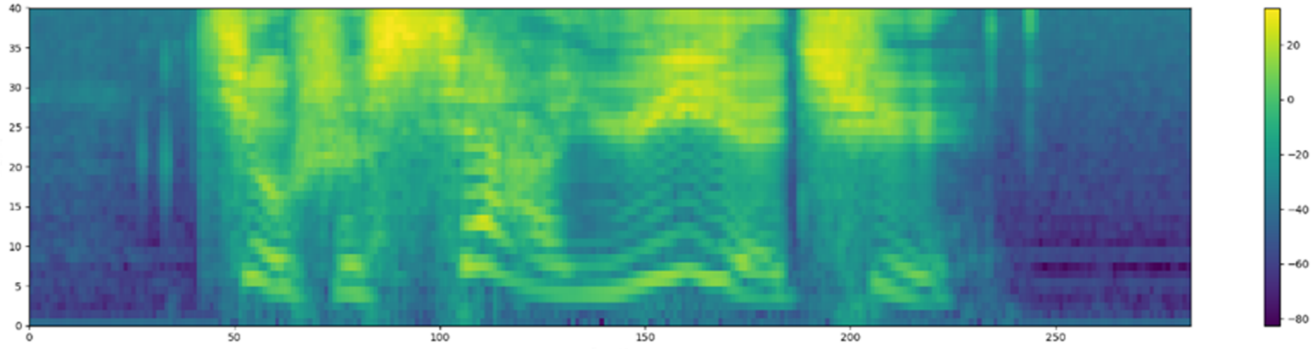

Figure 3. Mel-spectrogram.

It is generally believed that the more primitive the information, the more complete the information, and direct conversion of data generally leads to information loss. In theory, if the input is waveform data, there is a network that can automatically achieve FFT calculation, and it can also achieve any form of feature extraction. However, based on experience, manually defined transformations, such as spectrum or MFCC, are considered to perform well in speech emotion recognition tasks, so it is unnecessary to try another neural network to discover a better feature, which requires additional computational costs. The characteristic of Mel-spectrogram is that it contains both time-domain and frequency-domain information, and traditional transformations, such as FFT and Mel-filter, are also included. Therefore, Mel-spectrogram contains a large amount of original information, and some necessary transformations reduce the computational burden of the model. In addition, Mel-spectrogram can directly link convolutional layers, which makes it very suitable as an input for DRSN. Therefore, Mel-spectrogram is selected as input feature in this research.

3.2. Deep Residual Shrinkage Network

The notion behind the incorporation of the residual shrinkage module into the deep residual network has led to the development of the deep residual shrinkage network, as exemplified in Figure 4. The deep residual shrinkage network (DRSN) is a sophisticated deep neural network that employs attention mechanism and soft threshold function to remove noise in data. The self-attention mechanism utilized by the DRSN methodically selects relevant features while eliminating ineffectual features and noise, thereby elevating the deep neural network's capacity to extract valuable features from noisy data.

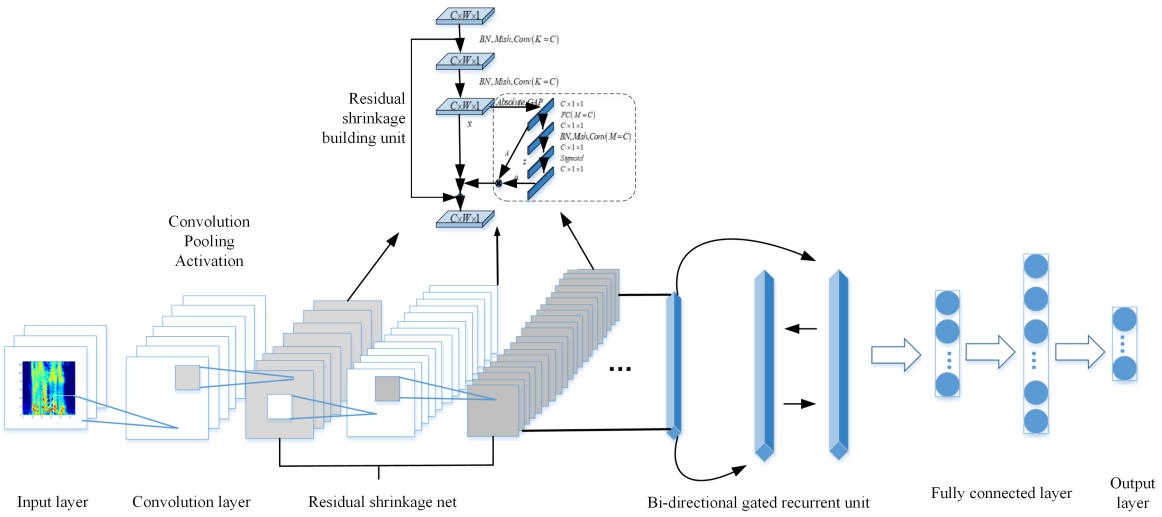

Figure 4. Structure of deep residual shrinkage network.

Convolution layer is an important component of the model, which can extract local features and reduce the number of weights through local connections and weight sharing of convolution kernels. The size of convolution kernels is an important parameter in building a model. Common convolution kernel sizes include 3×3, 5×5, and 7×7. A large convolution kernel size can increase the receptive field of the model, but the convolution calculation will also increase. If the size of the convolution kernel is small, the convolution calculation can be reduced, but then the receptive field becomes smaller. In the research of VGG-19, it is analyzed that stacking two layers of 3×3 convolution layer can be equivalent to 5×5 receptive field, and stacking three layers of 3×3 convolution layer can be equivalent to 7×7 receptive field [30]. When the receptive field is the same, multiple 3×3 convolution layers use multiple non-linear activation function, which will increase the non-linear expression ability, thus providing more complex pattern learning. In addition, using 3×3 convolution kernels can reduce parameters. Assuming that there are currently three layers of 3×3 convolution kernels stacked in a convolution layer with C input and output channels, its total number of parameters is $3 \times 3 \times 3 \times C \times C = 27C^2$. For one convolution layer with the same size of receptive field, the convolution kernel is 7×7. If the number of input and output channels are both C, the total number of its parameters is $7 \times 7 \times C \times C = 49C^2$. Overall, the stacking of 3×3 convolution kernel not only increases the depth of the network but also reduces the number of parameters, so the stacking of layers with 3×3 convolution kernel is selected for the network in this paper.

3.2.1. Residual Shrinkage Building Unit

Residual Shrinkage Building Unit (RSBU) is the basic unit of DRSN. The structure of residual shrinkage module is shown in Figure 5.

As shown in Figure 5, the DRSN is composed of two-batch normalization (BN), two activation functions (Mish), two convolution layers (Conv), one identity mapping, and one soft threshold learning subnetwork. There is a sub-network in each part and its function is to learn a set of thresholds independently. Its learning process is as follows:

(1) Take absolute values for all input features;
(2) An feature A is obtained after global average pooling (GAP);
(3) Input A into a fully connected network. The sigmoid function is on the last layer, and the output is a coefficient α in the range of 0 to 1;
(4) αA is the threshold to be learned.

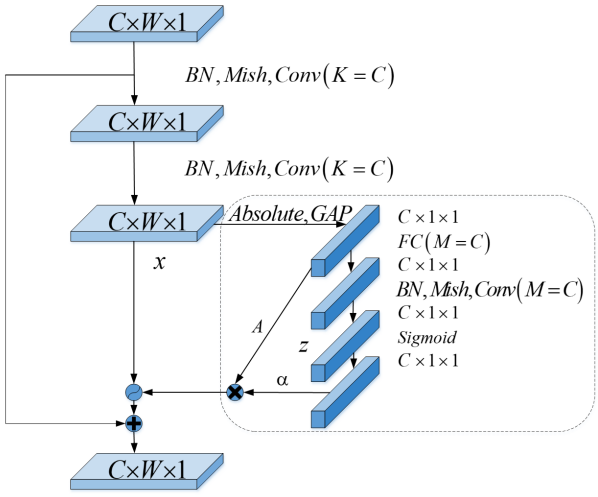

Figure 5. Residual shrinkage building unit.

In this way, the threshold is guaranteed to be positive and not too large. Different feature maps learn different thresholds, so the above subnetworks can be used as an

attention mechanism, the soft threshold converts the observed invalid features to zero; important features noted are preserved.

3.2.2. Soft Thresholding

In this paper, the soft threshold function is used to remove the features close to 0 and retain the positive and negative features. The soft threshold function is shown in Equation (4). In the equation, x is the input feature, y is the output feature, and thr is the threshold.

$$y = \begin{cases} x - thr & x > thr \\ 0 & -thr \le x \le thr \\ x + thr & x < -thr \end{cases} \tag{4}$$

Equation (5) is derived from Equation (4), and the soft threshold function becomes 1 or 0 after derivation.

$$\frac{\partial y}{\partial x} = \begin{cases} 1 & x > thr \\ 0 & -thr \le x \le thr \\ 1 & x < -thr \end{cases} \tag{5}$$

The process of soft threshold can also be converted into Equation (6), in which, x is the input data of the sub-network, thr is the threshold after network learning, and $sign(x)$ is the symbolic function.

$$soft = (x, a) = sign(x) * \max\{|x| - thr, 0\} \tag{6}$$

The key of soft thresholding is to obtain the range of threshold independently. The deep neural network has good performance in self-learning, so the combination of soft thresholding and deep neural network can well distinguish effective features from irrelevant features.

3.2.3. Self-Attention Mechanism

The attention mechanism principle involves facilitating neural network models to independently learn input features and assign weights to them. This approach allows the model to allocate ample computing resources to acquire crucial features, leading to improved performance.

The attention mechanism involves a process of key-value mapping through the query operation of vector matrix. This entails computing the similarity between dimension vectors and deriving the corresponding weight values, which are then normalized via the SoftMax function. Subsequently, the weight value is multiplied to the vector of each matrix dimension and their sum is obtained to formulate the final attention matrix. When $K = V = Q$, the process is referred to as self-attention and can be represented by Equation (7).

$$Attention(Q, K, V) = soft\max\left\{\frac{QK^T}{\sqrt{d_k}}\right\} \tag{7}$$

In which, $Q = (q_1, q_2, ..., q_l) \in R^{n \times d}$, $K = (k_1, k_2, ..., k_l) \in R^{n \times d}$, $V = (v_1, v_2, ..., v_n) \in R^{n \times d}$, d is the dimension of a single vector, n is the number of input vectors, which are obtained by linear transformation of input matrix X. K^T is the transposition of K, and d_k is a matrix with dimension k for adjusting inner product. The input sequence linear mapping process formula is as follows:

$$\begin{cases} Q = W_q X \\ K = W_k X \\ V = W_v X \end{cases} \tag{8}$$

In which, w_q, w_k, and w_v are three linear mapping parameter matrices, which are self learned in model training process.

3.3. Bi-Directional Gated Recurrent Unit

The gated recurrent unit (GRU) is a type of lightweight cyclic neural network that differs from other neural networks due to its internal "gate" structure. This unique structure enables the network to determine which data are relevant and which data can be discarded based on their relationship, facilitating effective data transmission within the network and effectively controlling internal information. As a result, the GRU partially addresses the issue of long-term dependence in neural networks. The internal structure of the GRU consists of three main components, the reset gate, update gate, and hidden state, which work together to extract temporal information and obtain long-term dependencies. The individual functions of these gates are described in further detail below, while the internal structure of the GRU is displayed in Figure 6.

(1) Reset gate and update gate

The input for the reset and update gates is obtained through the fully connected layer that operates on the current input x_t and the hidden state of the previous time step h_{t-1}. An activation function is applied to transform the input of the reset gate into a value between 0 and 1. This value is then used to determine the relevancy of the input. Specifically, an input with a value of 0 will be discarded whereas an input with a value of 1 will be retained. The product of the reset output and the previous hidden state is calculated element-wise. The update gate regulates the extent to which the previous state information is incorporated into the current state. Both the reset and update gates influence the candidate hidden state of the current neuron simultaneously. The reset gate can be computed as per Equation (9).

$$r_t = sigmoid\left([w_{hr}|w_{xr}] \otimes [h_{t-1}|x_t]^T\right) \tag{9}$$

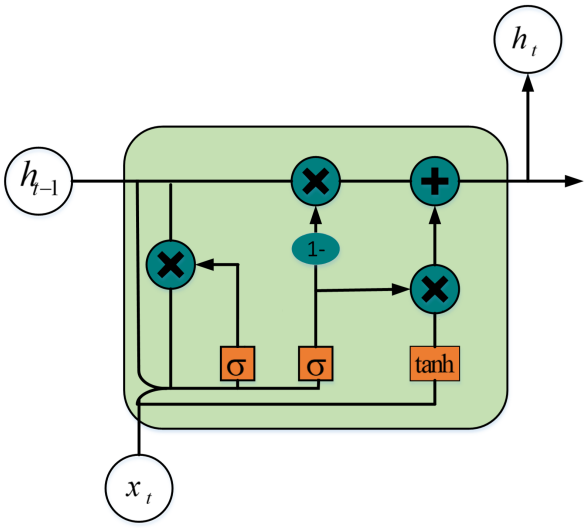

Figure 6. Structure diagram of gated recurrent unit.

In which $\otimes$ is cross product of matrix, t is the current time point, $t-1$ is the last time point, x_t is the input sequence of current time point, h_{t-1} is the hidden state of last time point, and w_{hr} and w_{xr} are the hidden state of last time and weight matrix of input sequence at current time point. According to the current information selection of the reset gate, the candidate hidden state is calculated by Equation (10).

$$\widetilde{h_t} = \tanh\left(w_{\widetilde{h}} \otimes [h_{t-1} \circ r_t)|x_t]\right) \tag{10}$$

In which ∘ is the Hadamard operator, h_t is the hidden state and w_h is the weight matrix of candidate hidden state. The output of update gate is calculated by Equation (11), where the data should be updated.

$$z_t = sigmoid\left(\left[w_{hz}|w_{xz}\right] \otimes \left[h_{t-1}|x_t\right]^T\right) \tag{11}$$

In which z_t is the output of update gate at current time. According to update gate, the retained information of hidden state at the last time point is calculated by Equation (12).

$$h^*_{t-1} = (1 - z_t) \circ h_{t-1} \tag{12}$$

(2) Hidden status

The hidden status is the output of the GRU through the candidate hidden status and the status information of the last moment reserved by the update gate, which is calculated by Equation (13).

$$h_t = \left(z_t \circ h'_t\right) \oplus h^*_{t-1} \tag{13}$$

In which, $\oplus$ is matrix addition operator.

(3) Bi-directional gated recurrent unit

The structure of BiGRU is shown in Figure 7.

For GRU, the information obtained at present is only affected by the information at the last moment. Generally, the status of hidden state transmission in GRU is from front to back. However, for the extraction of speech affective feature, we need to pay attention not only to the information at the current moment and the information at the previous moment, but also to the time state from the back to the front. It can be seen from Figure 7 that the hidden state of the current time step in the BiGRU network is affected by the current time step input sequence, the forward hidden state, and the backward hidden state. x_t is the input of BiGRU, $x_t \in R^{n \times d}$ (n is the number of small batch of input samples, d is the number of input moments). It is assumed that the forward hidden state is $\overrightarrow{h_t} \in R^{n \times h}$ and the backward hidden state is $\overleftarrow{h_t} \in R^{n \times h}$ which are calculated by Equations (14) and (15).

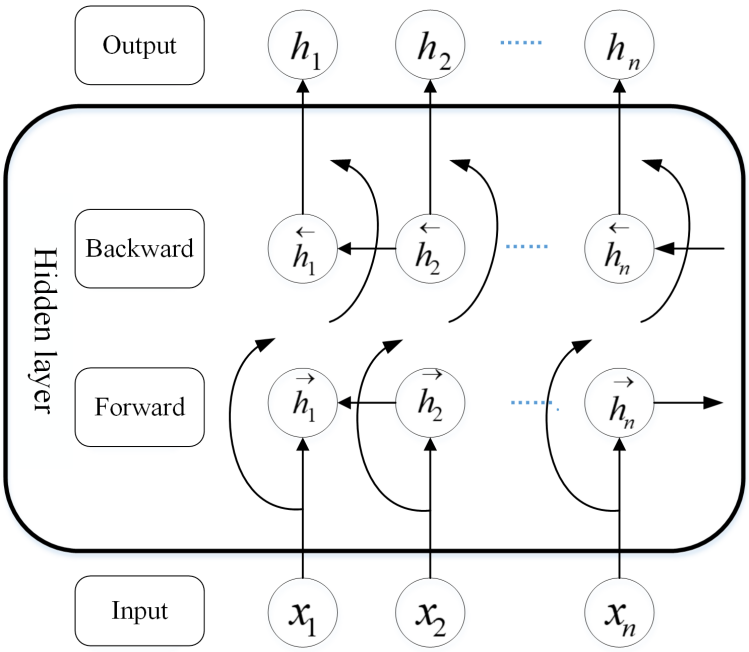

Figure 7. Structure of bidirectional gated recurrent unit.

$$\overrightarrow{h_t} = sigmoid\left(x_t w_{xh}^{(f)} + h_{t-1}\overrightarrow{w_{hh}^{(f)}} + b_h^{(f)}\right) \tag{14}$$

$$\overleftarrow{h_t} = sigmoid\left(x_t w_{xh}^{(b)} + \overleftarrow{h_{t-1}} w_{hh}^{(b)} + b_h^{(b)}\right) \tag{15}$$

In which, $w_{xh}^{(f)} \in R^{d\times h}$, $w_{hh}^{(f)} \in R^{h\times h}$, $w_{xh}^{(d)} \in R^{d\times h}$, $w_{hh}^{(b)} \in R^{h\times h}$ are weight matrix. Then the calculated $\overrightarrow{h}$ and $\overleftarrow{h}$ are connected to obtain the hidden state $h_t \in R^{n\times 2h}$ at current time, and the output layer $o^t \in R^{n\times q}$ is calculated by Equation (16).

$$o_t = h_t w_{hq} + b_q \tag{16}$$

4. Results

This paper used the DRSN to classify the emotion of the datasets, and use Tensorflow 2.2.0 deep learning framework and Python 3.8.0 to build the model. The hardware for the experiment are Intel Core i5-12750H CPU, RTX 3060 GPU, 32GB memory, and the parameter settings are shown in Table 2 below.

Table 2. Model training parameters.

Parameter	Value
Learning rate	10^{-4}
Epochs	160
Batch size	64
Optimization method	RMS Prop

Speech data were converted into Mel-spectrogram features as the input parameters of the network model, and DRN-BiGRU and DSRN-BiGRU were tested on CASIA and IEMOCAP speech emotion datasets, respectively. The performance of depth residual shrinkage module on the datasets are analyzed.

4.1. Datasets Introduction

The CASIA Chinese emotional corpus was curated by the esteemed Institute of Automation at the Chinese Academy of Sciences. The repository comprises of recordings by four proficient speakers and encompasses six distinct emotional states, namely anger, happiness, sadness, fear, surprise, and neutrality, totalling an impressive 9600 pronunciations. A noteworthy aspect of the corpus is that, while it contains 300 identical texts, the different emotional renderings of each exhibit the diverse acoustic and rhythmic performances through precise comparative analysis. In addition, the corpus features 100 distinct texts, which the literal meaning suggests as having inherent emotional leaning, facilitating more consistent and accurate emotion depiction. For our experiment, we handpicked the corpus's identical text for use as the experimental data, where the first 200 readings were assigned to the training set and the residual 100 allocated for testing. Furthermore, the experiment conducted partitioning of the dataset [31].

The Interactive Emotional Binary Motion Capture Database (IEMOCAP) was collected by the Speech Analysis and Interpretation Laboratory (SAIL) of the University of Southern California (USC). This database is an invaluable resource for the research and modeling of multimodal and expressive human communication, as it contains over 12 h of data recorded from 10 actors engaged in binary conversation scenes, with detailed motion capture information that identifies their facial expressions and hand movements. Additionally, the database includes interactive settings that stimulated specific emotions (such as happiness, anger, sadness, and depression) in emotionally scripted and naturally occurring conversation scenarios [32].

The MELD dataset, derived from the Emotion Lines dataset, features over 1400 conversations and 13,700 sentences drawn from the classic TV series Friends. In addition to the pure text dialogue present in the original dataset, MELD expands it by incorporating multimodal data, such as video and audio, including an average of 9.5 sentences per conversation and an average of 3.6 s per sentence. Each sentence is assigned one of seven emotional labels, including anger, disgust, sadness, happiness, neutrality, surprise, and fear, with corresponding emotional labels that are categorized as either positive, negative, or neutral [33].

The model proposed in this paper is trained by datasets, so the quality of data is important for the performance of the model. Three datasets are compared in Table 3 and there are enough data for the three datasets. Only CASIA dataset has a small number of speakers. In order to verify the performance of the model, all the three datasets will be used for experiments.

Table 3. Datasets ingrotuction.

Name	Speakers	Language	Data Volume	Emotions
CASIA	4 proficient speakers	Chinese	9600 pronunciations	anger, happy, sad, fear, surprise, neutral
IEMOCAP	10 actors	English	12 h	happy, angry, neutral, sad
MELD	many actors	English	13,700 sentences	happy, angry, sad, fear, surprise, neutral

4.2. Activation Funtion

The common activation function in deep learning are ReLU and Mish. This paper selects Mish as the activation function of the model which is shown in Equation (17). The positive value of Mish has no upper limit, which solves the problem of saturation caused by the upper limit. Compared with ReLU, whose negative values are all set to 0, Mish's slightly negative value theory can make gradient flow better. The smooth nature of Mish's can make the information in the input neural network more in-depth, thus improving the accuracy and generalization.

$$Mish(x) = x \times \tanh(\ln(1 + e^x)) \tag{17}$$

The activation function curve is shown in Figure 8. When the input data are negative, some small negative inputs can be retained as negative outputs to improve the interpretability and gradient flow of the network. When the input data are positive, linear operations are performed on the output and input to improve the convergence rate of the model.

The result of model with different activation is shown in Table 4. From the result, it can be seen that Mish performs better than ReLU on all three datasets. So the model proposed in this paper will select Mish as activation function.

Table 4. Comparison of actication functions.

Activation Function	CASIA	IEMOCAP	MELD
ReLU	84.62%	85.79%	69.6%
MiSH	86.03%	86.07%	70.57%

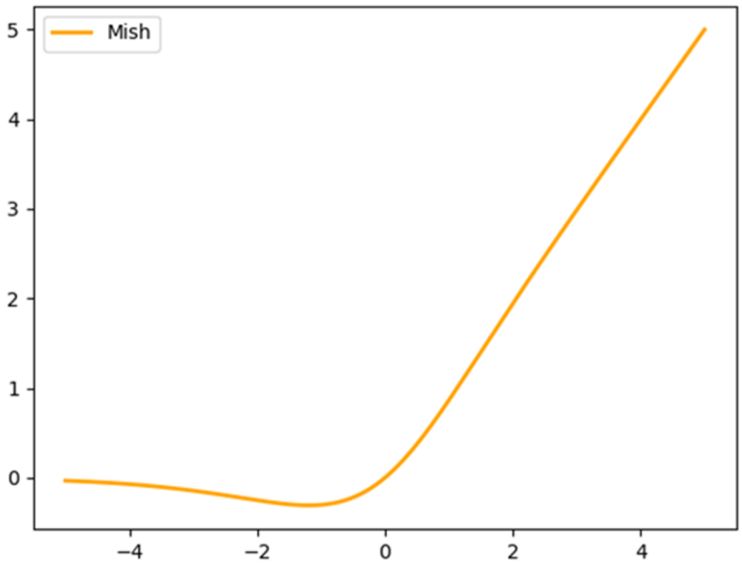

Figure 8. Mish function curve.

4.3. Optimization of RSBU Quantity

This paper outlines the composition of the DRSN-BiGRU model, incorporating multiple depth residual shrinkage building units and bidirectional gated recurrent units. The number of residual shrinkage building units employed correlates positively with the amount of speech emotional feature information extracted. However, an excess of residual shrinkage units can impede the network convergence speed, and only selective units can extract pivotal features and eliminate redundant ones. Thus, an experiment was conducted to determine the optimal number of residual shrinkage building units by comparing the model structures of varied units. The accuracy of the resulting network on CASIA and IEMOCAP datasets is illustrated in Figure 9.

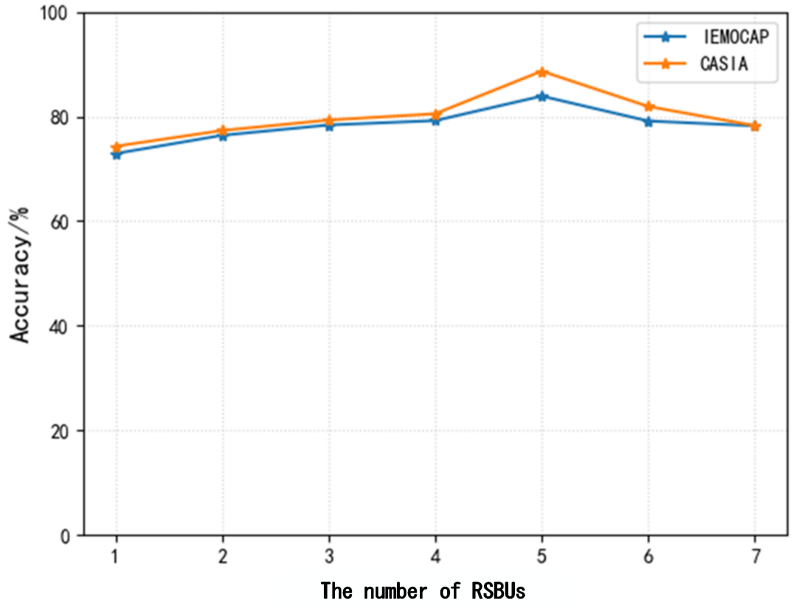

Figure 9. Influence of RSBU quantity.

Upon conducting an analysis, it was discovered that the utilization of a single residual shrinkage unit resulted in accuracy rates of 74.30% and 72.90% for CASIA and IEMOCAP, respectively. The addition of one more unit yielded improved accuracy rates of 79.31% and 78.37% for the respective datasets. Increased usage up to five residual shrinkage units provided the network with peak accuracy rates of 88.45% and 83.26%. However, adding further residual shrinkage building units cannot yield significant improvements and, in some cases, even resulted in decreased accuracy. This suggests that the initial residual shrinkage units are effective in extracting feature information, and subsequent additions only serve to complicate the network structure and disrupt the model. Based on the comprehensive analysis, it was determined that utilizing five residual shrinkage units results in the best performance for speech emotion recognition, and, thus, the DRSN-BiGRU in this paper employs the use of five residual shrinkage units.

4.4. Performance of DBSN-BiGRU

The DRSN-BiGRU model underwent training on the IEMOCAP dataset for 160 epochs. The accuracy curve depicting the learning process is displayed in Figure 10. It is evident from the learning curve that the performance of the model exhibited rapid enhancement during the initial 40 epochs of training. Subsequently, the performance level gradually stabilized and eventually reached a state where it no longer fluctuated significantly, manifesting a highly effective learning outcome. During the training process, the training curve and testing curve are almost synchronous, and there is no over fitting training.

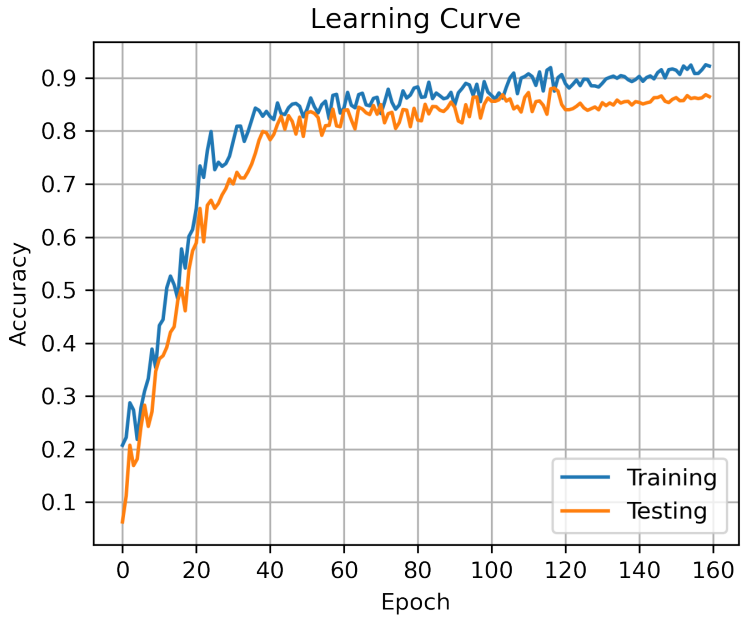

Figure 10. Learning curve of the models on IEMOCAP dataset.

The confusion matrix of testing result on IEMOCAP is shown in Figure 11. The highest accuracy is 90.2% for happy emotion and the lowest accuracy is 83.52% for angry emotion. The mean accuracy on IEMOCAP is 86.07%.

The DRSN-BiGRU model also underwent training on the CASIA dataset for 160 epochs. The accuracy curve depicting the learning process is displayed in Figure 12. It is evident from the learning curve that the performance of the model exhibited rapid enhancement during the initial 30 epochs of training which is similar to IEMOCAP dataset. Subsequently, the performance level gradually stabilized and eventually reached a state where it no longer fluctuated significantly, manifesting a highly effective learning outcome. During the training process, the training curve and testing curve are almost synchronous, and there is no over fitting training.

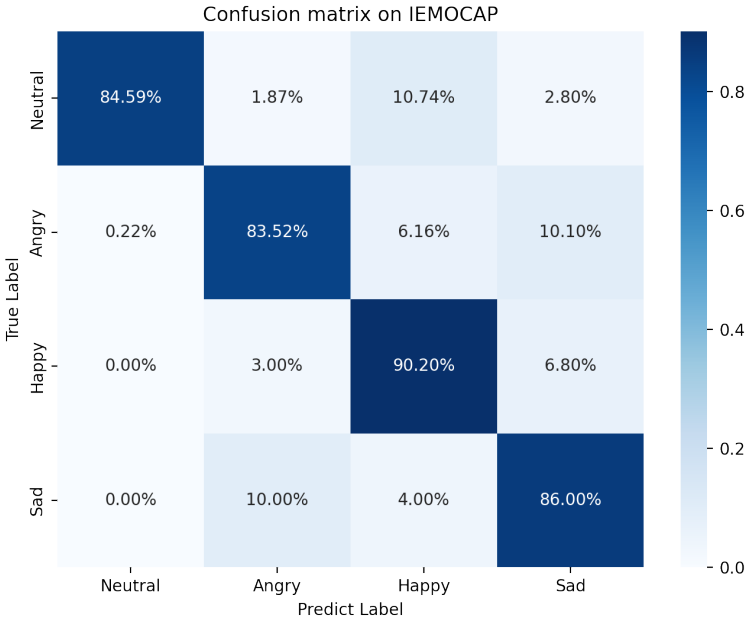

Figure 11. Confusion matrix of DRSN-BiGRU on IEMOCAP.

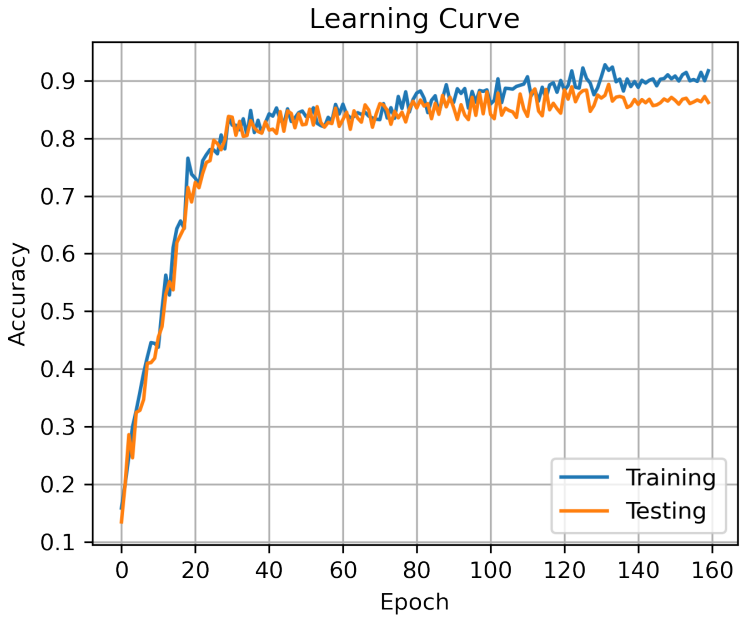

Figure 12. Learning curve of the models on CASIA dataset.

The confusion matrix of testing result on CASIA dataset is shown in Figure 13. The highest accuracy is 92.88% for happy emotion and the lowest accuracy is 78.65% for fear emotion. The mean accuracy on IEMOCAP is 86.03%.

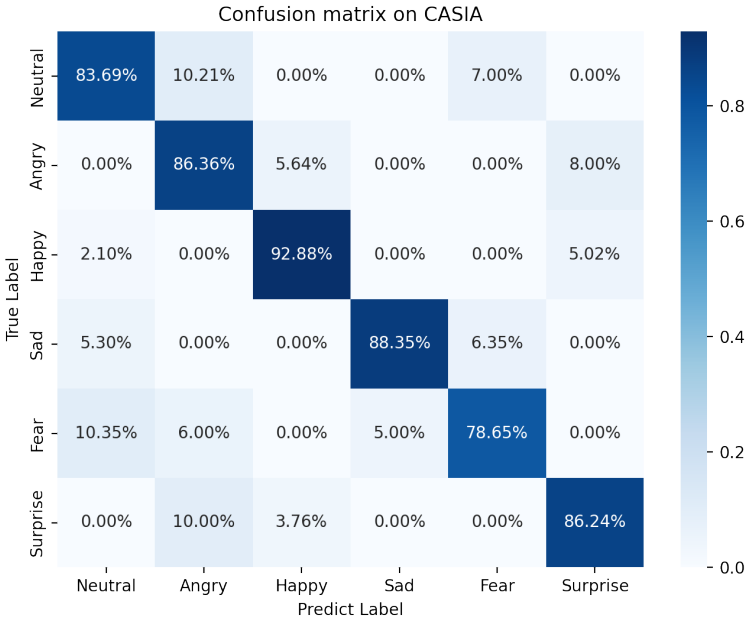

Figure 13. Confusion matrix of DRSN-BiGRU on CASIA dataset.

The same experiment are also carried out on MELD dataset and the accuracy curve depicting the learning process is displayed in Figure 14. It is evident from the learning curve that 160 epochs are also enough for MELD dataset and there is no over fit training.

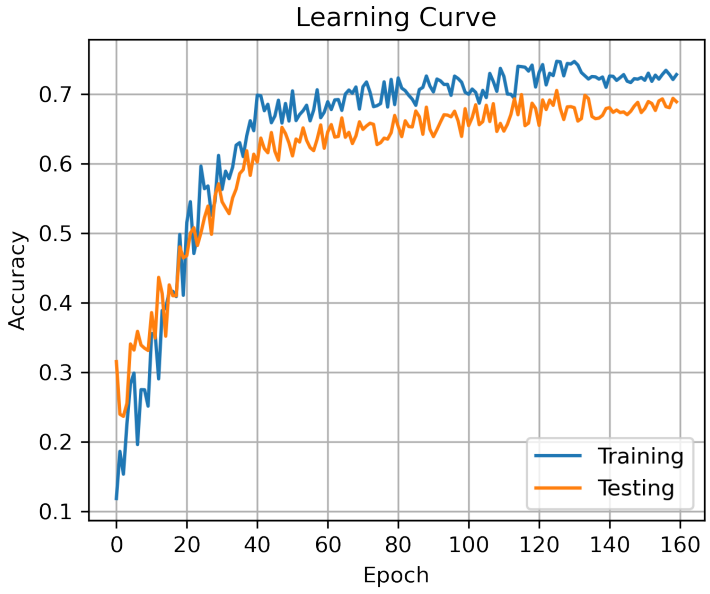

Figure 14. Learning curve of the models on MELD dataset.

In natural situations, human emotions are more multifaceted, influenced by environmental stimuli, rapid emotional shifts, mixed emotions, and subjective hidden emotions. Thus, recognizing emotions in real-life situations is more challenging. MELD dataset extracted from the Friends television series closely resembles natural speech patterns. The confusion matrix of the DRSN-BiGRU network model tested on the MELD dataset is illustrated in Figure 15.

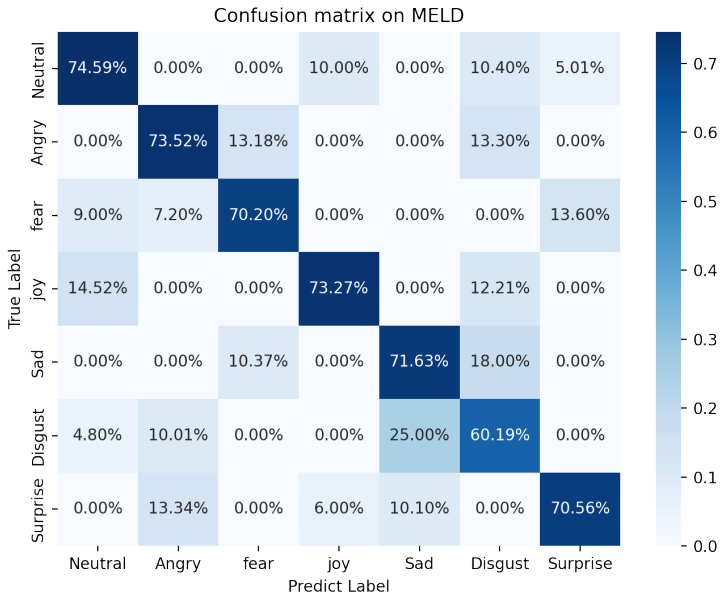

Figure 15. Confusion matrix of DRSN-BiGRU on MELD.

Upon analyzing the confusion matrix, it is evident that DRSN-BiGRU demonstrates a recognition accuracy of 70.57% on the MELD dataset. The model's accuracy is comparatively lower than CASIA and IEMOCAP due to the complexity of speech emotion recognition in natural situations, especially in the case of a seven-classification problem as in the MELD dataset. The DRSN-BiGRU performs exceptionally well in neutral, angry, and happy emotions, with a recognition accuracy of 74.59%, 73.52%, and 73.27%, respectively. Nevertheless, the model exhibits poor recognition on "disgust" emotion, with a recognition accuracy rate of only 60.19%. During the classification process of "disgust" emotion, the model has a 25% probability of categorizing it as "sad", a 10.01% probability of categorizing it as "angry", and a 4.8% probability of categorizing it as "neutral".

To further evaluate DRSN-BiGRU's efficacy in natural situations, the model was compared to DCNN-LSTM, CNN-BiGRU, and DRN-BiGRU, and their results are presented in next section.

5. Discussion

5.1. Results Discussion

From the comfusion matrix on IEMOCAP (Figure 10), it can be seen that anger and sadness are the most obvious confusing emotions. Because anger and sadness are both negative emotions, there is a portion of anger speech that does not erupt and also shows a relatively low voice, which is similar to the characteristics of sadness, leading to confusion between the two. Sometimes the speech of anger status depends on the speaker's personality. Some introverted people tend to be relatively introverted and less likely to burst out when angry, while some extroverted people tend to express aggressive behavior when angry, such as faster speaking speed, louder volume, and higher pitch. The differences in personality create differences in the expression of angry speech, which poses significant challenges in emotion recognition.

Compared to the IEMOCAP dataset, the CASIA dataset has increased the number of emotional labels, but the recognition accuracy has not decreased. This is because there are only four speakers in the CASIA dataset and the speech style is more unified, which, to some extent, compensates for the defect of more emotional labels. Another pair of easily confused emotions is neutrality and fear, which is manifested in the experimental results as 7% of neutrality being recognized as fear, and 10% of fear being recognized as neutrality.

From a psychological perspective, people sometimes conceal their inner emotions in a state of fear, and their speech is plain and neutral, so the two are easily confused. From the two datasets of CASIA and IEMOCAP, it can be found that the highest recognition accuracy is happy emotion. Happiness is the only positive emotion in the emotional labels, so the graph features are significantly different from others. In traditional emotion recognition methods based on feature extraction, such as energy, pitch, MFCC, formant, etc., it is easy to confuse happiness and anger. From the current research results, it can be seen that there is a significant difference between spectrogram features and traditional features.

In addition, it can be seen from the confusion matrix of MELD (Figure 15) that sadness and disgust are easy to be confused. From the perspective of language, these two emotions belong to negative emotions, and the vocal performance is relatively light, slightly slow, and more low-frequency voice, which makes the two emotions closer in Mel-spectrogram and leads to more confusion in recognition.

5.2. Result Comparison and Discussion

In order to verify the performance of DRSN-BiGRU algorithm, this paper compares it with CNN-BiLSTM, DCNN+LSTM, and DRN-BIGRU's accuracy of emotion recognition in IEMOCAP and CASIA, respectively. The accuracy of these algorithms for different emotions in CASIA dataset is shown in Table 5.

Table 5. Comparison result on CASIA.

Model	Surprise	Fear	Sad	Happy	Angry	Neutral
DCNN-LSTM	78.46	82.21	84.63	80.53	75.44	75.77
CNN-BiLSTM	70.46	73.21	74.31	70.46	65.13	65.32
DRN-BiGRU	81.38	80.46	86.68	82.46	71.45	80.94
DRSN-BiGRU	86.24	78.65	88.35	92.88	86.36	83.69

Table 5 highlights the significant improvement in accuracy achieved by the DRSN-BiGRU algorithm in recognizing emotions, such as "happiness", "anger", "surprise", and "neutrality", when compared with the other three algorithms. Additionally, DBSN-BiGRU shows a slight improvement in the recognition accuracy of "sad" emotions when compared with DRN-BiGRU. DRSB-BiGRU performs slightly inferior in recognizing "fear", but exhibits a significant improvement over CNN-BILSTM, and is outperformed by DCCN+LSTM and DRN-BIGRU. Overall, the results demonstrate that DRSN-BiGRU model outperforms the other three models. To further validate the performance of DRSN-BiGRU, Table 6 presents a comparison of results obtained from IEMOCAP.

Table 6. Comparison result on IEMOCAP.

Model	Sad	Happy	Angry	Neutral
DCNN-LSTM	77.46	81.21	81.63	78.53
CNN-BiLSTM	74.31	70.54	65.13	65.32
DRN-BiGRU	80.96	79.46	85.68	81.91
DRSN-BiGRU	86.00	90.20	83.52	84.59

Table 6 illustrates that the IEMOCAP emotional dataset consists of four categories. In contrast, DRSN-BiGRU demonstrates exceptional performance for the emotions of "happy", "sad", and "neutral" significantly enhancing recognition accuracy when compared to the remaining three algorithms. Moreover, DRSN-BiGRU's accuracy for recognizing "anger" surpasses that of CNN and DCNN, albeit slightly lower than that of DRN-BiGRU. This affirms that DRSN-BiGRU remains the top-performing algorithm among the four.

In order to further verify the superiority of DRSN-BiGRU in natural situations, the algorithm was also compared with DCNN-LSTM, CNN-BiGRU, and DRN-BiGRU in MELD dataset, and the results are shown in Table 7.

Table 7. Comparison result on MELD.

Model	Neutral	Angry	Fear	Joy	Sadness	Disgust	Surprise
DCNN-LSTM	68.32	52.13	54.21	50.32	55.43	53.31	56.65
CNN-BiLSTM	71.65	47.89	45.68	61.86	55.59	48.87	59.54
DRN-BIGRU	70.23	58.56	62.78	64.45	59.35	52.31	60.34
DRSN-BIGRU	74.59	73.52	70.20	73.27	71.63	60.19	70.56

The analysis of Table 7 reveals a notable decline in the performance of the four models on MELD, as opposed to their performance on CASIA and IEMOCAP. This outcome underscores the influence of naturalistic emotional complexity on the efficacy of these models. It is worth noting that DNSN-BiGRU yields the highest recognition accuracy for the seven emotions in question, establishing it as the optimal model for MELD.

In conclusion, the addition of residual shrinkage unit has led to significant improvements in speech emotion recognition accuracy for DRSN-BiGRU compared to CNN-BiGRU, DCNN, and DRN-BiGRU models. This is attributed to the residual network's capability of transmitting shallow information to the deeper network and the automatic removal of redundant features by the shrinkage unit. These experiments confirm the effectiveness of this method in speech emotion recognition.

5.3. Complexity Analysis and Discussion

The time complexity of a Deep Residual Shrinkage Network (DRSN) includes both forward and backward propagation. During forward propagation, the computing time complexity of DRSN is similar to that of a traditional fully connected neural network, $O(n^2)$, where n is the number of input samples. This is because, in each convolution layer, convolution and pooling operations need to be performed for each input sample, and the results are then passed to the next layer. During backward propagation, the time complexity of DRSN depends on the implementation of the back propagation algorithm. If the standard back propagation algorithm is used, the time complexity is $O(nm)$, where n is the number of samples and m is the number of parameters in the network. This is because for each parameter, a gradient calculation needs to be performed, and the time complexity of computing a gradient for each sample is $O(n)$.

The time complexity of this model is mainly reflected in the training process, so we discuss the time complexity of the training process. Due to the shared weights in the convolution layer, the size of the convolution kernel directly determines the training time complexity of the model. In this paper, a 3×3 convolution kernel is used, so its time complexity will be much smaller than the model uses 5×5 or 7×7 convolution kernel. The model structure also adopts a general DRSN structure, and there is no significant increase in network depth or the number of residual shrinkage units, so there is no significant improvement in time complexity. In addition, this paper adopts BiGRU, where the forgetting gate has forgetting characteristics, which can effectively avoid the gradient explosion problem happened during the training process. Overall, the time complexity of the entire model is controllable. When deploying the model, it is necessary to consider factors, such as the real-time requirements of the application scenario and the amount of data, and provide necessary computational support according to actual needs.

5.4. Real-Life Applications Discussion

Speech emotion recognition technology can enhance the user experience in human–computer interaction by providing more natural and personalized communication with computers, and the method proposed in this paper have reached a good performance. However, the real-life application of this technology still faces many challenges. First, model training depends on a big dataset. So far the training datasets are biased and lack representation of minority groups, which can lead to algorithms with poor performance in real-life application. Developing methods to increase dataset diversity and overcome this issue is necessary. Secondly, speech emotion recognition performance is influenced by language background, culture, and accent. Future research should investigate methods for overcoming these linguistic challenges. Thirdly, speech signals are often interfered by environmental noise, which can affect the accuracy of speech emotion recognition algorithms. Developing methods for enhancing speech signals in the presence of noise is crucial. Finally, computational complexity also needs to be considered. Our current experimental research does not require real-time calculation results. In human–computer interaction applications, real-time performance greatly enhances user experience rather than emotional understanding. So in real-life applications, it is necessary to research on deploying the model and the computational power requirements.

6. Conclusions

The utilization of speech emotion recognition technology plays a crucial role in the field of human–computer interaction. It offers a significant enhancement in user experience during human–computer interaction processes and serves a vital function in product recommendation, public opinion monitoring, human–computer dialogue, and other related areas. This paper focuses on exploring speech emotion recognition technology. Firstly, it suggests the use of Mel-spectrogram of speech as a feature, and elaborates on the extraction process of Mel-spectrogram meticulously. Then, a deep residual shrinkage network (DRSN) is proposed for speech emotion recognition. This network incorporates residual shrinkage building units, attention mechanisms, and bi-directional gated recurrent unit on top of the deep residual network (DRN), effectively allowing the network to disregard signal noise and enhance its ability to extract relevant features from noisy data. Finally, the paper performed network optimization and comparative verification on open emotional speech datasets, such as CASIA, IEMOCAP, and MELD. Experimental results indicate that the network performs best with five residual shrinkage units. Deploying the DRSN-BiGRU model resulted in an accuracy of 86.03% in CASIA, 86.07% in IEMOCAP, and 70.57% in MELD datasets. The model's performance in MELD is relatively poorer than the other datasets due to variations in human emotion's voice expressions in natural situations. Lastly, the paper compared DBSN-BiGRU to other state-of-the-art models, such as DCNN+LSTM, CNN+BiLSTM, and DRN-BiGRU on three datasets, and our proposed method proved to be superior.

Author Contributions: Methodology, T.H.; writing—original draft preparation, Z.Z.; conceptualization, C.D.; data curation, M.R.; supervision, X.J.; software, Q.Z.; project administration, T.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Jinhua Science and Technology Bureau, grant number 2022-1-046 and Jinhua Advanced Research Institute, grant number G202209 and G202207.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: http://www.chineseldc.org/resource_info.php?rid=76 and https://sail.usc.edu/iemocap/iemocap_release.htm.

Acknowledgments: This research was supported by Jinhua Science and Technology Bureau and Jinhua Advanced Research Institute.

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

References

1. Wani, T.M.; Gunawan, T.S.; Qadri SA, A.; Kartiwi, M.; Ambikairajah, E. A comprehensive review of speech emotion recognition systems. *IEEE Access* **2021**, *9*, 47795–47814. [CrossRef]
2. 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]
3. Zvarevashe, K.; Olugbara, O. Ensemble learning of hybrid acoustic features for speech emotion recognition. *Inf. Process. Manag.* **2020**, *13*, 70. [CrossRef]
4. Zhao, Z.; Bao, Z.; Zhao, Y.; Zhang, Z.; Cummins, N.; Ren, Z.; Schuller, B. Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition. *IEEE Access* **2019**, *7*, 97515–97525. [CrossRef]
5. Bhavan, A.; Chauhan, P.; Shah, R.R. Bagged support vector machines for emotion recognition from speech. *Knowl.-Based Syst.* **2019**, *184*, 104886. [CrossRef]
6. Fahad, M.S.; Deepak, A.; Pradhan, G.; Yadav, J. DNN-HMM-Based Speaker-Adaptive Emotion Recognition Using MFCC and Epoch-Based Features. *Circuits Syst Signal Process.* **2021**, *40*, 466–489. [CrossRef]
7. Shahin, I.; Nassif, A.B.; Hamsa, S. Emotion recognition using hybrid Gaussian mixture model and deep neural network. *IEEE Access* **2019**, *58*, 26777–26787. [CrossRef]
8. Liu, Z.T.; Wu, M.; Cao, W.H.; Mao, J.W.; Xu, J.P.; Tan, G.Z. Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing* **2018**, *273*, 271–280. [CrossRef]
9. Ke, X.; Zhu, Y.; Wen, L.; Zhang, W. Speech emotion recognition based on SVM and ANN. *Int. J. Mach. Learn. Comput.* **2018**, *8*, 198–202. [CrossRef]
10. Daneshfar, F.; Kabudian, S.J.; Neekabadi, A. Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier. *Appl. Acoust.* **2020**, *166*, 107360. [CrossRef]
11. Alex, S.B.; Mary, L.; Babu, B.P. Attention and feature selection for automatic speech emotion recognition using utterance and syllable-level prosodic features. *Circuits Syst. Signal Process.* **2020**, *39*, 5681–5709. [CrossRef]
12. Patnaik, S. Speech emotion recognition by using complex MFCC and deep sequential model. *Multimed. Tools Appl.* **2023**, *82*, 11897–11922. [CrossRef]
13. Bhangale, K.; Kothandaraman, M. Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network. *Electronics* **2023**, *12*, 839. [CrossRef]
14. Patil, S.; Kharate, G.K. PCA-Based Random Forest Classifier for Speech Emotion Recognition Using FFTF Features, Jitter, and Shimmer. *Proc. ICEEE* **2022**, *2*, 194–205.
15. Gumelar, A.B.; Yuniarno, E.M.; Adi, D.P.; Setiawan, R.; Sugiarto, I.; Purnomo, M.H. Transformer-CNN Automatic Hyperparameter Tuning for Speech Emotion Recognition. In Proceedings of the 2022 IEEE International Conference on Imaging Systems and Techniques, Kaohsiung Taiwan, China, 21 June 2022.
16. Kaya, H.; Fedotov, D.; Yesilkanat, A.; Verkholyak, O.; Zhang, Y.; Karpov, A. LSTM Based Cross-corpus and Cross-task Acoustic Emotion Recognition. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018.
17. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control.* **2019**, *47*, 312–323.
18. Zhang, S.; Zhang, S.; Huang, T.; Gao, W. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. *IEEE Trans. Multimed.* **2018**, *20*, 1576–1590. [CrossRef]
19. Sun, L.; Zou, B.; Fu, S.; Chen, J.; Wang, F. Speech emotion recognition based on DNN-decision tree SVM model. *Speech Commun.* **2019**, *115*, 29–37. [CrossRef]
20. Huang, J.; Tao, J.; Liu, B.; Lian, Z. Learning Utterance-Level Representations with Label Smoothing for Speech Emotion Recognition. In Proceedings of the INTERSPEECH, Shanghai, China, 25–29 October 2020.
21. Atmaja, B.T.; Akagi, M. Evaluation of error-and correlation-based loss functions for multitask learning dimensional speech emotion recognition. *J. Physics Conf. Ser. IOP Publ.* **2021**, *1896*, 012004. [CrossRef]
22. Cai, X.; Yuan, J.; Zheng, R.; Huang, L.; Church, K. Speech Emotion Recognition with Multi-Task Learning. In Proceeding of the Interspeech, Brno, Czechia, 30 August–3 September 2021.
23. Yeh, S.L.; Lin, Y.S.; Lee, C.C. Speech Representation Learning for Emotion Recognition Using End-to-End ASR with Factorized Adaptation. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020.
24. Bakhshi, A.; Wong, A.S.W.; Chalup, S. End-to-end speech emotion recognition based on time and frequency information using deep neural networks. In Proceedings of the ECAI 2020, Santiago de Compostela, Spain, 29 August–8 September 2020; IOS Press: Amsterdam, The Netherlands, 2020; pp. 969–975.
25. Sun, T.W. End-to-end speech emotion recognition with gender information. *IEEE Access* **2020**, *8*, 152423–152438. [CrossRef]

26. Sajjad, M.; Kwon, S. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access* **2020**, *8*, 79861–79875.
27. Issa, D.; Demirci, M.F.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control.* **2020**, *59*, 101894. [CrossRef]
28. Wang, Y.; Shen, G.; Xu, Y.; Li, J.; Zhao, Z. Learning Mutual Correlation in Multimodal Transformer for Speech Emotion Recognition. In Proceedings of the Interspeech, Brno, Czechia, 30 August–3 September 2021.
29. Zou, H.; Si, Y.; Chen, C.; Rajan, D.; Chng, E.S. Speech emotion recognition with co-attention based multi-level acoustic information. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022.
30. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
31. Li, Y.; Tao, J.; Chao, L.; Bao, W.; Liu, Y. CHEAVD: A Chinese natural emotional audio–visual database. *J. Ambient. Intell. Humaniz. Comput.* **2017**, *8*, 913–924. [CrossRef]
32. Yu, Y.; Kim, Y.J. Attention-LSTM-attention model for speech emotion recognition and analysis of IEMOCAP database. *Electronics* **2020**, *9*, 713. [CrossRef]
33. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019.