electronics

Article

Research on Speech Emotion Recognition Based on AA-CBGRU Network

Yu Yan and Xizhong Shen *

School of Electrical and Electronic Engineering, Shanghai Institute of Technology, Shanghai 201418, China; yuyan_sit@163.com
* Correspondence: xzshen@yeah.net

Abstract: Speech emotion recognition is an emerging research field in the 21st century, which is of great significance to human–computer interaction. In order to enable various smart devices to better recognize and understand the emotions contained in human speech, in view of the problems of gradient disappearance and poor learning ability of the time series information in the current speech emotion classification model, an AA-CBGRU network model is proposed for speech emotion recognition. The model first extracts the spectrogram and its first and second order derivative features of the speech signal, then extracts the spatial features of the inputs through the convolutional neural network with residual blocks, then uses the BGRU network with an attention layer to mine deep time series information, and finally uses the full connection layer to achieve the final emotion recognition. The experimental results on the IEMOCAP sentiment corpus show that the model in this paper improves both the weighted accuracy (WA) and the unweighted accuracy (UA).

Keywords: speech emotion recognition; attention mechanism; residual block; bidirectional gated recurrent unit; human–computer interaction

Citation: Yan, Y.; Shen, X. Research on Speech Emotion Recognition Based on AA-CBGRU Network. *Electronics* **2022**, *11*, 1409. https://doi.org/10.3390/electronics11091409

Academic Editor: Ahmad Taher Azar

Received: 17 March 2022
Accepted: 25 April 2022
Published: 28 April 2022

1. Introduction

The human voice contains rich information, and voice interaction is the most direct and convenient way of communication between people. In the era of artificial intelligence, the voice is of great significance for human–computer interaction, and people hope to give machines the same voice interaction capabilities as humans. In the early 21st century, with the rapid development of voice technology, voice interaction is no longer limited to the text content that the speaker needs to express. How to make various smart devices understand and judge the emotional state of the speaker has become a new problem and a research hotspot. Speech emotion recognition (SER) [1–4] refers to the technology that the computer infers the emotional state of the speaker by analyzing and processing the speech signal collected from the sensor. Speech emotion recognition can enable various smart devices to better understand the user's intention, so as to better serve people.

In recent years, researchers have mainly studied SER in two aspects: the construction of a speech emotion database and the extraction of speech emotion features. Among them, feature extraction methods can be divided into two categories. One is the traditional manual features. The traditional static model obtains the emotional key element features of each sample through two links: the first link is to accurately collect the most basic descriptive information in the voice content, such as harmonic-to-noise ratio, fundamental frequency, zero-crossing rate, etc. In the second link, each piece of basic descriptive information is expressed as a feature vector by different statistical aggregation functions, which expresses the temporal variation and contours of different basic descriptive information at the sentence level, including mean, variance, skewness, kurtosis, linear regression coefficient, quartile, etc. [5]. However, traditional machine learning methods based on advanced statistical function features cannot effectively utilize the contextual information of the original speech signal, resulting in a low recognition rate.

In recent years, the deep feature extraction methods based on deep learning have shown outstanding performance in SER tasks [6–8]. Compared with traditional hand-crafted feature extraction, deep neural networks (DNNs) are able to extract task-specific hierarchical feature representations from a large number of training samples through supervised learning. The authors of [9] mentioned that by designing a speech emotion recognition system that can use the convolutional neural network (CNN) and recurrent neural network (RNN) technology, and using the sound spectrum as the input of the model, the final recognition rate on the EMO-DB dataset can reach 88.01%, and the experiment proves the effectiveness of the CRNN network model in the speech emotion classification task. Zhong et al. [10] introduced an attention mechanism into the CRNN model, which improved the weighted accuracy by 7% compared with the original CRNN network. The experiments proved that the attention mechanism could effectively improve the accuracy of speech emotion recognition. Li et al. [11] mentioned that in the process of their research, the one-dimensional convolutional network model was used as the construction carrier of the speech emotion recognition system, and the Mel spectrogram and the Log-Mel spectrogram were combined to the input end, and finally the reliability of this method was demonstrated and affirmed. Zhang et al. [12] proposed a speech emotion recognition method based on a fully convolutional network and an attention mechanism, which achieved an unweighted accuracy of 63.9% on the IEMOCAP dataset.

In the process of this research, through a large number of collections and the analysis of existing research conclusions and research cases, we proposed a model that was based on the AA-CBGRU network for SER. In the design process of the model in this paper, we firstly extracted the spectrogram and its first and second order derivatives from the original speech signal, and then we fed these features into the advanced convolutional neural network to extract the spatial features. The residual network was introduced to avoid the gradient disappearance problem caused by the deep network structure. Finally, the bidirectional gated recurrent unit network (BGRU) with attention mechanism was designed to strengthen the learning of time series information to improve the classification performance of the model.

2. Speech Emotion Recognition Model Based on AA-CBGRU Network

The specific architecture of speech emotion recognition based on the AA-CBGRU network designed and constructed in this paper is shown in Figure 1 below. In this model, it mainly includes four parts: data input, spatial feature collection, time series feature collection and classification.

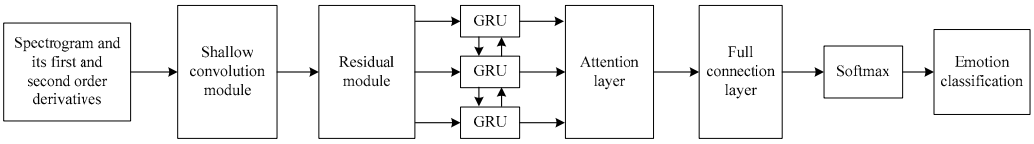

Figure 1. Speech emotion recognition model based on AA-CBGRU network.

The data input part first preprocesses the original speech signal, and then extracts the spectrogram and its first and second order derivatives as the data input for the advanced convolutional neural network. The spatial feature extraction part extracts spatial features through the convolutional neural network with residual structure, and the time series feature extraction part then inputs the extracted spatial features into the BGRU layer along the time axis to capture the time series features of the speech signal. The classification part is responsible for receiving the output of the attention layer and then inputting it to the fully connected layer, and finally the model realizes sentiment classification through the SoftMax layer.

2.1. Feature Selection

Since static features such as the logarithmic Mel-spectrogram cannot reflect the process of emotional change [13], the consideration of the dynamic features can reflect the changing process of emotion, retain effective emotional information, and reduce the content, environment, speaker and other emotionally irrelevant factors. Therefore, the model in this paper selects the spectrogram and its first and second order derivatives which contain dynamic features such as the input of the improved convolutional neural network.

For a given speech signal, the power spectrum of each frame is obtained by discrete Fourier transform, and the output p_i is obtained by Mel filter bank. Then the Log-Mel static feature m_i, the first-order difference feature m_i^d, and the second-order difference feature m_i^{dd} are calculated as follows:

$$m_i = \log(p_i) \tag{1}$$

$$m_i^d = \frac{\sum_{n=1}^{N} n(m_{i+n} - m_{i-n})}{2\sum_{n=1}^{N} n^2} \tag{2}$$

$$m_i^{dd} = \frac{\sum_{n=1}^{N} n\left(m_{i+n}^d - m_{i-n}^d\right)}{2\sum_{n=1}^{N} n^2} \tag{3}$$

where the i describes the number of frames, N is set to 2, based on the popular experience, and n is a variable.

After the spectrogram and its first and second order derivatives are calculated, the 3D feature $X \in \mathrm{R}^{t \times f \times c}$ is used as the input of the convolutional neural network, where t represents the time (frame) length and f represents the number of Mel filter banks, c represents the number of feature channels.

2.2. Advanced Convolutional Neural Network (Advanced CNN, A-CNN)

In recent years, many researchers have used convolutional neural networks (CNN) to compute spectrogram features in speech emotion recognition tasks [14–16]. In the deep neural network model, in general, there are many convolution layers under the front end of the network system structure, and the collected relevant information is described through partial or all input features. The deep neural network structure can make the receptive field wider, so as to obtain better recognition effect. However, practice has proved that a network structure that is too deep will increase the possibility of gradient disappearance.

In order to extract more effective features and avoid the risk of vanishing gradients, this paper adds a residual network [17] to the network architecture. As shown in the feature representation module in Figure 1, a convolutional neural network with residual blocks is constructed in the feature representation layer. The advanced convolutional neural network consists of a shallow convolution module and a residual module. Specifically, the shallow convolution module consists of one convolutional layer, one max-pooling layer, three convolutional layers, and one max-pooling layer successively. The first convolutional layer has 128 feature maps, while the remaining convolutional layers have 256 feature maps, and the filter size of each convolutional layer is 5×3. In the two max-pooling layers, the pooling size is 2×2. The shallow convolution module takes the spectrogram and its first and second order derivatives as inputs, extracting the low-level features of the input data, and then inputs the obtained low-level features to the residual module. The residual module includes a total of five identical residual blocks. The residual block structure is shown in Figure 2.

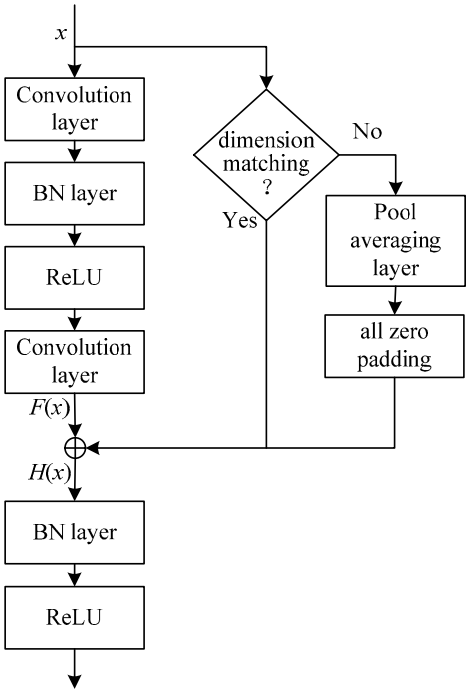

Figure 2. Residual block structure.

Residual networks add skipping structures outside the convolutional layers. Let the input of the network be x, the result of the operation after the convolution layer is $F(x)$, and the output of the residual network is $H(x)$. If there are x and $F(x)$ dimensions matching, then

$$H(x) = f(F(x) + x) \tag{4}$$

If there is a difference in the dimensions between x and $F(x)$, then some padding layers and average pooling layers will be added at the jump structure, and the padding layer will be used to enrich the dimension of the input information, so that the output dimension will be the same as that of $F(x)$ dimensions match.

2.3. Bidirectional Gated Recurrent Unit (BGRU)

GRU [18] (gated recurrent unit) is a variant of the LSTM [19] network that works well. Three gate functions are introduced in LSTM: input gate, forget gate and output gate. The GRU only includes two gate functions: update gate and reset gate. The purpose of the former is to continue the valid information of the content of the last moment, and to control the input of candidate information. The purpose of the latter is to control whether the calculation of the candidate state depends on the state of the previous moment. GRU has fewer tensor operations, so it is faster to train than LSTM. The network structure is shown in Figure 3.

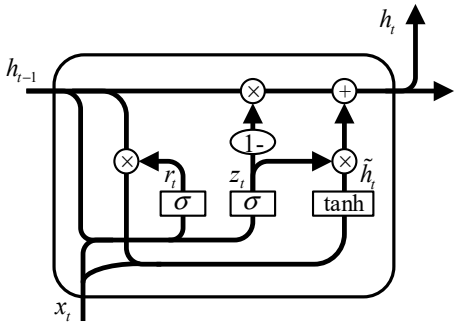

Figure 3. The structure of GRU.

At time t, let the current input be x_t, the hidden state output of GRU is h_t, and the hidden state at the previous time is h_{t-1}. The calculation method is as follows:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{5}$$

$$r_t = \sigma(W_r[h_{t-1}, x_t]) \tag{6}$$

$$\widetilde{h}_t = \tan \mathrm{h}(W \cdot [r_t^* h_{t-1}, x_t]) \tag{7}$$

$$h_t = (1 - z_t) \times h_{t-1} + z_t \times \widetilde{h}_t \tag{8}$$

Among them, z_t is the update gate, r_t is the reset gate, W is the weight matrix connecting the two layers, and σ and tanh are the activation functions.

However, the standard GRU can only transfer data in one direction in time when dealing with sequence problems, that is, data can only flow from the past to the present, without considering the influence of the later data on the previous data. The BRNN (bidirectional recurrent neural network) model improves the unidirectional recurrent neural network, and processes the input sequence from both positive and negative directions at the same time. The bidirectional gated recurrent unit (BGRU) can be regarded as the parallel connection of two unidirectional GRUs in opposite directions through the output layer, that is, a unidirectional GRU that processes sequence data from front to back and a unidirectional GRU that processes sequence data from back to front. Unidirectional GRU is connected to the same output layer [20]. The BGRU structure can pay attention to the context information at the same time to make more accurate judgments. In this article, we use one BGRU layer to collect the temporal information, and each direction contains 128 cells, then we can obtain a sequence of 256-dimensional high-level feature representations. The network structure is shown in Figure 4.

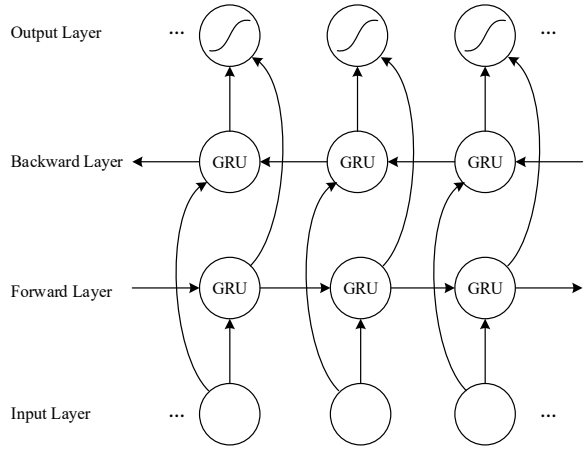

Figure 4. The structure of BGRU.

For a given n-dimensional input $(x_1, x_2, \cdots, x_n)$, at time t, the hidden layer of BGRU outputs h_t, and the calculation method is as follows:

$$\overrightarrow{h}_t = \sigma\left(W_{\underset{xh}{\rightarrow}} x_t + W_{\underset{hh}{\rightarrow\rightarrow}} \overrightarrow{h}_{t-1} + b_{\underset{h}{\rightarrow}}\right) \tag{9}$$

$$\overleftarrow{h}_t = \sigma\left(W_{\underset{xh}{\leftarrow}} x_t + W_{\underset{hh}{\leftarrow\leftarrow}} \overleftarrow{h}_{t-1} + b_{\underset{h}{\leftarrow}}\right) \tag{10}$$

where b is the bias vector, and $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ are the outputs of the forward and reverse GRUs, respectively.

2.4. Attention Mechanism

When humans process information, they selectively focus on some of the most important information, while ignoring other visible information to a certain extent. This mechanism is called the attention mechanism. In the speech emotion recognition task, not all speech features in a speech play an important role in the judgment of emotional state. In this paper, the attention mechanism is applied to the features extracted from emotional speech, so that the model can give different features to the output of the BGRU network of attention. The specific implementation method is as follows:

$$e_i = u^{\mathrm{T}} \tanh(W a_i + b) \tag{11}$$

$$\alpha_i = (\exp(\lambda e_i)) / \left(\sum_{k=1}^{L} \exp(\lambda e_i) \right) \tag{12}$$

$$k = \sum_{i=1}^{L} \alpha_i a_i \tag{13}$$

The new representation e_i of the input sequence a_i is obtained by using Equation (11) through the multilayer perceptron layer (MLP) with tanh as the nonlinear activation function. Equation (12) normalizes the attention score e_i to an attention weight α_i between 0 and 1 through the SoftMax function. Formula (13) uses the obtained attention weight α_i to weight the feature vector a_i in the input L frame features, and finally obtains the weighted feature representation k.

3. Experimental Results and Analysis

3.1. Sentiment Corpus and Network Parameter Settings

This paper used the IEMOCAP emotional speech database recorded by the University of Southern California for experiments. The corpus contains approximately 12 h of audiovisual data. It consists of five sessions, and each session is displayed by a man and a woman. The two professional actors performed scripted and improvised scenarios that elicit expressions of specific emotions. The IEMOCAP database is annotated into categorical labels by multiple annotators and contains a total of nine sentiment categories. According to previous work [21], this paper selected four types of emotional data that researchers are generally concerned about and the most representative, with a total of 2280 sentences, including 289 angry sentences, 284 happy sentences, 608 sad sentences, and 1099 neutral sentences. We implemented a five-fold cross validation. In each fold, the data from four sessions was used for model training, and the data from the remaining session was split: one actor for validation and the other one as the testing set.

The model in this paper used the PyTorch deep learning framework, and the parameters were set as follows: the learning rate was 0.0001, the batch size was 40, and the number of iterations (epoch) was 300. Adam was used as the network optimizer.

3.2. Evaluation Indicators

This paper used the evaluation indicators commonly used in the field of speech emotion recognition: the value of weighted accuracy (WA) and unweighted accuracy (UA) was used to achieve an effective interpretation of performance, and among them, WA represents the overall accuracy of the sample information, and UA stands for the mean of sentiment accuracy.

3.3. Experimental Results and Analysis

In order to verify the effectiveness of the model proposed in this paper, experiments were carried out on the IEMOCAP emotion corpus, and we provide a comparison of the accuracy of existing research and our model in Table 1, all of the experiments in the table used the improvised part of IEMOCAP as data set.

Table 1. Comparison of proposed model with existing model.

Model	WA	UA
Siddique et al. [22]	-	60.23
Etienne et al. [23]	64.50	61.70
Zhang et al. [12]	70.40	63.90
Ma et al. [24]	71.45	64.22
Our Method	72.83	67.75

As can be seen from Table 1, compared with the research methods in the above literature, our proposed model is state-of-the-art and improved on both WA and UA on IEMOCAP dataset.

To verify the effectiveness of each module in the proposed model, we conducted ablation experiments, and the experimental results are shown in Table 2.

Table 2. Ablation experiments.

Model	WA	UA
CNN + BGRU + attention	69.90	64.29
CNN + residual block + BGRU	69.37	64.72
CNN + residual block + GRU + attention	71.18	66.58
CNN + residual block + BGRU + attention	72.83	67.75

The experiment ablated the residual block, attention and bidirectional network structure in the model, respectively. After comparing the experimental results with the model in this paper, it can be seen that when the residual block, attention and bidirectional network structure were not used, both WA and UA of the model had different degrees of decline. Among them, the ablation of residual block and attention had the most obvious impact on the classification performance of the model. The experimental results verified the effectiveness of the proposed model and the importance of each component to the classification performance of the model.

Through the above demonstration, in order to better realize the in-depth research and analysis of the recognition model, the confusion matrix of the model in this paper under the IEMOCAP data set is given, as shown in Figure 5.

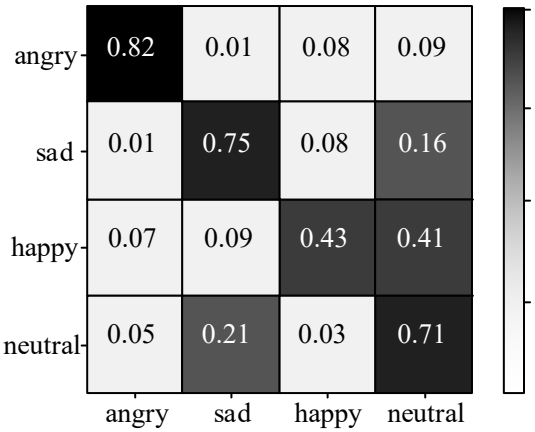

Figure 5. The confusion matrix of our model.

It can be seen from the confusion matrix in Figure 5 that the three types of emotions, anger, sadness and neutral, have obtained high recognition accuracy rates of 82, 75 and 71%, respectively, while the recognition rate of happy is only 43%; 41% of the happy samples were wrongly classified as neutral. There are two possible reasons for the low recognition rate of happiness in the IEMOCAP dataset: (1) from the dimension space of emotion, the

activation level of happiness is higher, while neutral is at the center of the activation valence space. (2) The IEMOCAP dataset has the problem of sample imbalance, while happy is less in the number of samples. Subsequently, the number of samples can be balanced by overlapping methods.

4. Conclusions

This paper designs a reasonable and feasible speech emotion recognition system with the help of an advanced convolutional neural network and BGRU with attention mechanism. The model uses the spectrogram and its first and second order derivatives as the input of the advanced convolutional neural network, and uses the advanced convolutional neural network to extract the spatial features of the inputs. The participation of the residual network not only helps to improve the depth of the network architecture, but also can avoid unnecessary gradient vanishing risks. The experimental results on the IEMOCAP dataset showed that the proposed method had more advantages than similar models in terms of recognition accuracy. In future research, we will continue to explore deep neural network models with different structures to further improve the accuracy of speech emotion recognition.

Author Contributions: Methodology, Y.Y.; software, Y.Y.; writing—original draft preparation, Y.Y.; writing—review and editing, X.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available in article.

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

References

1. Khalil, R.A.; Babar, M.I.; Jan, T. *Speech Emotion Recognition Using Deep Learning Techniques: A Review*; IEEE Access: Piscataway Township, NJ, USA, 2019; Volume 7, pp. 117327–117345.
2. Issa, D.; Demirci, M.F.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control* **2020**, *59*, 101894. [CrossRef]
3. Abbaschian, B.J.; Sierra-Sosa, D.; Elmaghraby, A. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. *Sensors* **2021**, *21*, 1249. [CrossRef] [PubMed]
4. Kwon, S. Optimal Feature Selection Based Speech Emotion Recognition Using Two-Stream Deep Convolutional Neural Network. *Int. J. Intell. Syst.* **2021**, *36*, 5116–5135.
5. Kim, J.; Saurous, R.A. *Emotion Recognition from Human Speech Using Temporal Information and Deep Learning*; Interspeech: Hyderabad, India, 2018.
6. Tzirakis, P.; Zhang, J.; Schuller, B.W. End-to-end speech emotion recognition using deep neural networks. In Proceedings of the 2018 IEEE International Conference On Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: Piscataway Township, NJ, USA, 2018.
7. Li, P.; Song, Y.; McLoughlin, I.; Guo, W.; Dai, L. *An Attention Pooling Based Representation Learning Method for Speech Emotion Recognition*; Interspeech: Hyderabad, India, 2018.
8. Zhao, Z.; Zheng, Y.; Zhang, Z.; Wang, H.; Zhao, Y.; Li, C. *Exploring Spatio-Temporal Representations by Integrating Attention-Based Bidirectional-LSTM-RNNs and FCNs for Speech Emotion Recognition*; Harvard University: Cambridge, MA, USA, 2018.
9. Lim, W.; Jang, D.; Lee, T. Speech Emotion Recognition Using Convolutional And Recurrent Neural Networks. In Proceedings of the 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, Korea, 31 May 2016; IEEE: Piscataway Township, NJ, USA, 2016.
10. Zhong, Y.; Hu, Y.; Huang, H.; Silamu, W. *A Lightweight Model Based on Separable Convolution for Speech Emotion Recognition*; Interspeech: Shanghai, China, 2020.
11. Li, Y.; Baidoo, C.; Cai, T.; Kusi, G.A. Speech Emotion Recognition Using 1d Cnn with No Attention. In Proceedings of the 2019 23rd International Computer Science and Engineering Conference (ICSEC), Phuket, Thailand, 30 October–1 November 2019; IEEE: Piscataway Township, NJ, USA, 2019.
12. Zhang, Y.; Li, H.; Hashimoto, K.; Patil, H.A.; Nankaku, Y.; Oura, K. Attention based fully convolutional network for speech emotion recognition. In Proceedings of the 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 12–15 November 2018; IEEE: Piscataway Township, NJ, USA, 2018.

13. Abdel-Hamid, O.; Mohamed, A.R.; Jiang, H.; Deng, L.; Penn, G.; Yu, D. *Convolutional Neural Networks for Speech Recognition*; IEEE/ACM Transactions on Audio, Speech, and Language Processing: Piscataway Township, NJ, USA, 2014; Volume 22, pp. 1533–1545.
14. Cummins, N.; Liu, Q.; Lienhart, R. An Image-Based Deep Spectrum Feature Representation for the Recognition of Emotional Speech. In Proceedings of the 25th ACM International Conference on Multimedia, New York, NY, USA, 23–27 October 2017.
15. Huang, C.-W.; Narayanan, S. *Characterizing Types of Convolution in Deep Convolutional Recurrent Neural Networks for Robust Speech Emotion Recognition*; IEEE: Piscataway Township, NJ, USA, 2017.
16. Neumann, M.; Vu, N.T. *Attentive Convolutional Neural Network Based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech*; Interspeech: Hyderabad, India, 2017.
17. Xi, Y.; Li, H.; Dang, J.; Tao, J.; Yi, J.; Akagi, M. Speaker to emotion: Domain adaptation for speech emotion recognition with residual adapters. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Beijing, China, 18–21 November 2019; IEEE: Piscataway Township, NJ, USA, 2019.
18. Dey, R.; Salem, F.M. Gate-variants of gated recurrent unit (GRU) neural networks. In Proceedings of the 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), Medford, MA, USA, 6–9 August 2017; IEEE: Piscataway Township, NJ, USA, 2019.
19. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
20. Mu, Y.; Gómez, L.H.; Montes, A.C.; Martínez, C.A.; Wang, X.; Gao, H. Speech emotion recognition using convolutional-recurrent neural networks with attention model. *DEStech Trans. Comput. Sci. Eng.* **2017**, 341–350. [CrossRef]
21. Satt, A.; Rozenberg, S.; Hoory, R. *Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms*; Interspeech: Hyderabad, India, 2017.
22. Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Epps, J. *Direct Modelling of Speech Emotion from Raw Speech*; Interspeech: Graz, Austria, 2019.
23. 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. [CrossRef]
24. Ma, X.; Wu, Z.; Jia, J.; Xu, M.; Meng, H.; Cai, L. *Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms*; Interspeech: Hyderabad, India, 2018.