applied sciences

MDPI

Article

Research on Speech Emotion Recognition Method Based A-CapsNet

Yingmei Qi [1,2], Heming Huang [1,2,*] and Huiyun Zhang [1,2]

[1] School of Computer Science, Qinghai Normal University, No. 38, Wusixilu Road, Xining 810000, China
[2] The State Key Laboratory of Tibetan Intelligent Information Processing and Application, No. 38, Wusixilu Road, Xining 810000, China
* Correspondence: huanghm@qhnu.edu.cn

Abstract: Speech emotion recognition is a crucial work direction in speech recognition. To increase the performance of speech emotion detection, researchers have worked relentlessly to improve data augmentation, feature extraction, and pattern formation. To address the concerns of limited speech data resources and model training overfitting, A-CapsNet, a neural network model based on data augmentation methodologies, is proposed in this research. In order to solve the issue of data scarcity and achieve the goal of data augmentation, the noise from the Noisex-92 database is first combined with four different data division methods (emotion-independent random-division, emotion-dependent random-division, emotion-independent cross-validation and emotion-dependent cross-validation methods, abbreviated as EIRD, EDRD, EICV and EDCV, respectively). The database EMODB is then used to analyze and compare the performance of the model proposed in this paper under different signal-to-noise ratios, and the results show that the proposed model and data augmentation are effective.

Keywords: speech emotion recognition; data augmentation; data division; network model

Citation: Qi, Y.; Huang, H.; Zhang, H. Research on Speech Emotion Recognition Method Based A-CapsNet. *Appl. Sci.* **2022**, *12*, 12983. https://doi.org/10.3390/app122412983

Academic Editors: Shengzong Zhou and Jingsha He

Received: 21 October 2022
Accepted: 1 December 2022
Published: 17 December 2022

1. Introduction

As science and technology have progressed, computers have invaded all parts of daily life, and they now serve a key role in learning, entertainment, and work [1]. Human–computer interaction (HCI) is a significant area of study, industry, and application as the need of people for computer functions grows. The purpose of speech emotion processing is to extract emotion parameters from speech signals so that the emotion information buried in the sentence can be recognized [2]. It is an important part of information processing, and it is used in artificial intelligence (AI), aided psychotherapy, and emotional translation. The goal of emotional computing is to enable computers to observe, interpret, and think in the same way as human do. The goal of affective computing is to provide computers with the ability to observe anthropomorphically, to comprehend and generate various emotional traits, to give computers human-like emotional intelligence, and finally, to realize natural, friendly, and vivid HCI [3]. Traditional speech signal study focuses mostly on semantics, but the emotional information embedded in the signal cannot be overlooked. The same statement with varied emotional elements might cause significant perception variations in the listener.

Speech signals can help a listener to form a more accurate assessment of the emotional state of a speaker [4]. Human communication and emotion transmission rely heavily on speech. Emotion recognition of speech signals is of great practical significance for accurately judging human emotion and realizing man–machine natural communication and computer intelligence [5]. Speech emotion recognition (SER) has become an active research area, and it provides users with a smoother interface and gives them appropriate feedback or recommendations [6]. Most of the SER systems mainly consist of the following two stages:

feature extraction and emotion classification. Therefore, constructing an effective acoustic model is a key factor to the success of SER systems [7].

In order for the HCI technique to be timely changed in accordance with the different sorts of emotions, the computer must be able to precisely identify the operator's feelings. The recognition rate was calculated using the hidden Markov model (HMM) in the literature [8]. Literatures [9–13] use artificial neural networks (ANN), recurrent neural networks (RNN) and convolution neural networks (CNN) to recognize speech emotion. Increasing numbers of models are studied by scholars and the accuracy of recognition can be gradually improved. Among the mentioned models, the HMM is a parametric representation of time-varying features that simulate the process of human language processing, and it needs many samples for time-consuming training [14]. A Gaussian mixture model (GMM) is a probability density estimation model that can fit all probability distribution functions; however, it depends heavily on data distribution or volume, and it is sensitive to data noise [15]. Support vector machines (SVM) map the feature vectors from the input space to a high-dimensional Hilbert space by using kernel tricks at first and then seeking an optimal hyper-plane in the high-dimensional space to classify samples. However, it cannot solve the problem of large-scale training samples that lead to a large or prohibitively large kernel matrix [10].

Deep neural networks have better performance regarding their learning capabilities when handling large-scale data. However, different models have their own pros and cons. For example, an RNN is good at dealing with time series information [16] and a CNN effectively collects spatial data [17]. The most well-known deep learning models are CNN-based, and they perform exceptionally well when learning representations [18]. However, CNNs have two significant flaws, which are as follows: they do not account for significant spatial hierarchies between features, and they do not have rotational invariance [19]. Convolutional layers and pooling layers make up the typical CNN structure. When this happens, the Max-pooling layer, one of the most popular pooling layers, unavoidably removes the information that is not "max-information," which is also important in SER. The convolutional layer takes into account the data inside its receptive field and extracts characteristics in this local region, but it disregards the spatial connections and orientation data within the global region of the emotional speech [20]. To overcome these disadvantages, Sabour et al. propose capsule networks (CapsNet) [19], which contain a set of neurons that maintain activity vectors, whose lengths represent presence probabilities and orientations represent entity attributes. However, the computational complexity of CapsNet is high. Therefore, to overcome this problem, this paper proposes an average-pooling capsule network (A-CapsNet), a speech emotion recognition method based on average-pooled ensemble capsule networks. In addition, because existing databases such as the Chinese language database (CASIA) [21], Berlin Emotional Database (EMODB) [22], and Surrey Audio-Visual Expressed Emotion (SAVEE) database [23] have small data volumes, a data division method is used to increase the data volume in order to address the problem of model over-fitting brought on by data scarcity.

The main contributions of this study may be summed up as follows: (1) the data augmentation approach is proposed to supplement baseline databases and further reduce the overfitting issue; (2) the model A-CapsNet is proposed to identify emotions and reduce the overfitting danger.

The structure of this paper is as follows. In Section 2, the used databases are described and the data augmentation approach is shown. The preprocessing and feature extraction processes are described in Section 3. In Section 4, the suggested model A-CapsNet is introduced. The experimental and results analyses are described in Section 5. The suggested model's advantages and disadvantages are examined in Section 6, along with suggestions for future research.

2. Our Database

To query the effectiveness of the proposed A-CapsNet in SER, the traditional baseline database EMODB [24] has been chosen. The EMODB is a German database, including 10 audio systems and seven emotions, i.e., boredom (B), anger (A), fear (F), disappointment (Sa), disgust (D), happiness (H), and impartial (N), and it consists of 535 emotional sentences in total.

To reinforce the baseline database EMODB, 15 noises from the noise database Noisex-92 [25] are used for 5 signal-to-noise ratios (−10 dB, −5 dB, 0 dB, 5 dB, and 10 dB) and 15 exceptional varieties of noise are delivered with the 5 SNRs to achieve 5 augmented unmarried SNR databases. The previously mentioned 15 types of noises are babble, buccaneer1, buccaneer2, volvo, f16, destroyerengine, destroyerops, factory1, factory2, hfchannel, leopard, ml09, red, machinegun, and white, respectively. They may be further merged to obtain a novel Multi-SNRs database (Multi-SNR-5-EMODB) with 40,125 samples, as shown in Table 1. The number and proportion of emotional samples in the augmented EMODB datasets with different SNRs are shown in Figure 1. A higher number of colored squares indicates a higher number of samples of an emotion type.

Table 1. The augmented and merged database Multi-SNR-5-EMODB.

Database	Emotion	SNR	Noises from NoiseX-92	Total Samples
EMODB	Anger/A, boredom/B, fear/F, disgust/D, happiness/H, neutral/N, and sadness/S	−10 dB	babble, white, buccaneer1, buccaneer2, destroyerengine, destroyerops, f16, volvo, factory1, factory2, hfchannel, leopard, ml09, pink, machinegun	8025
		−5 dB		8025
		0 dB		8025
		5 dB		8025
		10 dB		8025

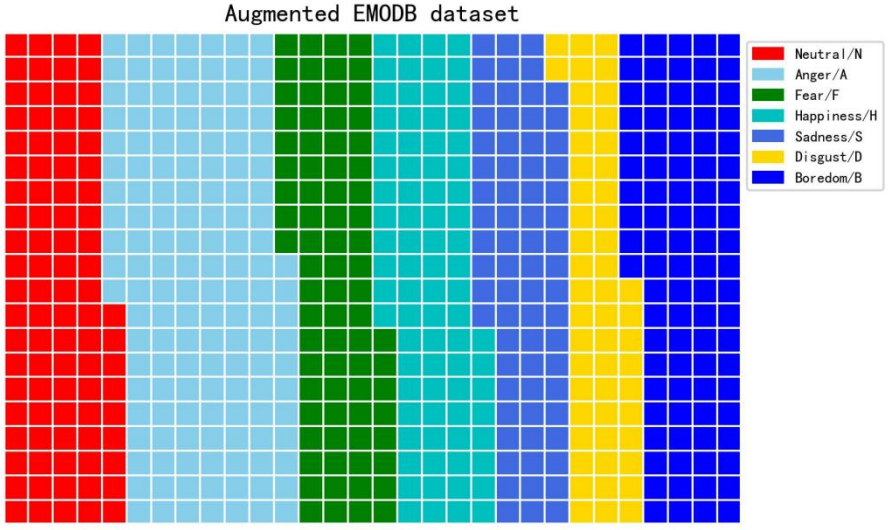

Figure 1. The number and proportion of emotional samples in the augmented EMODB datasets with different SNRs.

3. Preprocessing and Feature Extraction

3.1. Preprocessing

The feature extraction process is a very important part of speech emotion recognition, and this is followed by two basic tasks, digitization and preprocessing [26], which are also essential. The preprocessing process is the same as that for speech signals, mainly including

anti-aliasing filtering, pre-emphasis, windowing, framing, and endpoint detection. After preprocessing, the voice data are extracted. By training these voice data, we can achieve the purpose of computer recognition of speech signals. The original speech signal is an analog signal. So-called digitization is the process of converting an analog signal into a discrete digital signal, and the two most important links are sampling and quantization. The reason for pre-emphasis is that the speech signal will cause partial attenuation of high frequencies after passing through the lips and nostrils [27]. Compared with low frequency, the spectrum analysis of high frequency is harder to obtain. The purpose of pre-emphasis is to improve the high-frequency parts of speech. After pre-emphasis, the spectrum of speech signals is smoother, which is more conducive to spectrum analysis. A speech signal is a typical non-stationary time-varying signal that carries various types of information. Because of its short-time stability, it is divided into frames through the process of windowing to obtain short-time voice signals.

The framing of the speech signal is realized by the process of weighting with a window function. There are usually two methods of framing, continuous segmentation and overlapping segmentation [28]. However, to ensure a smooth transition between adjacent frames, overlapping segmentation is usually adopted. Emotion is the basis of emotion modeling, speech emotion synthesis, and speech emotion recognition. Only by establishing a large-scale and highly realistic emotional speech database can we engage in the above research. The emotional database provides a large amount of analysis data for emotional speech analysis and modeling and provides a modeling basis for emotional speech synthesis [29,30]. People can distinguish the emotions of different speakers because speech signals contain speech data that can reflect different emotions, so the differences in emotions can be reflected by the differences in speech data. Emotional acoustic features are divided into prosodic features, sound quality features, and spectrum-based correlation features. Prosodic features refer to the characteristics of pitch, duration, speed, and size of the middle tone, including duration-related features, fundamental frequency-related features, and energy-related features. Its existence will affect our emotional judgment of discourse, and many existing feature sets contain prosodic features as emotional features [31].

3.2. Feature Extraction

Feature extraction is a key component of an emotion recognition system. Whether the extracted features can correctly reflect emotions exactly or not will directly affect the recognition effect. The input of the proposed model A-CapsNet consists of 21D LLD features [32]. Each speech is segmented into frames with a 25ms window and a 10ms shifting step. For each frame, 21D LLD features, including 1D ZCR and 20D MFCCs, are extracted. Some features of the samples are quite different, so when the samples are directly inputted into the training, the contour of the loss function will be flat and long [33]. Before finding the optimal solution, the gradient descent process is not only complex, but also very time-consuming. Therefore, it is necessary to quantify each feature into a unified interval.

After data standardization, all indicators are in the same order of magnitude, and they are suitable for comprehensive evaluation. After feature normalization, the contour of the loss function will be a partial circle, the gradient descent process will be flatter [28], and the convergence will be faster; therefore, the performance will be improved. The StandardScaler [34] is used to normalize the data so that the new dataset has a mean and standard deviation of zero. It should be noted that the normalization procedure should be applied to each feature rather than each sample.

4. The Proposed Novel Model A-CapsNet

The deep neural network structure of voice emotion identification is becoming increasingly sophisticated, as deep learning continues to advance. Compared with the previous simple feedforward neural network (FNN) [24], a CNN encodes multiple low-level features into more differentiated high-level features in the way of spatial context awareness, and then reduces the dimension of the features through the pooling layer. A dense layer is used

as a classifier to detect the emotion [35]. In this paper, we refer to the Capsnet structure proposed by Sabour et al. in 2017 to solve the problem that CNN can only extract features, but cannot extract the state, direction, location, and other information of features, resulting in the poor generalization ability of the model.

Capsule networks represent a recent breakthrough in neural network architectures that introduce an alternative to translational invariance other than pooling through the use of modules, or capsules. Two key features distinguish them from CNNs, layer-based squashing and dynamic routing. Whereas CNNs have their individual neurons squashed through non-linearities, capsule networks have their output squashed as an entire vector. Capsules replace the scalar-output feature detectors of CNNs with vector-output capsules and max-pooling with routing-by-agreement. Capsnet architectures typically include several convolution layers, with a capsule layer in the final layer [36].

4.1. Network Structure

The CapsNet differs from earlier CNNs in that it employs a collection of neurons rather than a single neuron, whose activity vector reflects the instantiation parameters of a certain type of item, such as an object or object portion. The length of an activity vector reflects the chance that the entity exists, and the orientation of an activity vector represents the instantiation parameters. Active capsules at one level anticipate the instantiation parameters of high-level capsules via transformation matrices. A high-level capsule becomes activated when numerous forecasts coincide. CapsNet, in other words, produces a vector rather than a single scalar, allowing it to model global geographical information and replace pooling layers with a dynamic routing mechanism to prevent losing important data. The dynamic routing process is an information selection method that ensures that child capsule outputs are sent to the appropriate parent capsules [37,38].

Referring to the paper [19,39], a brief introduction to the structure of Capsnet is presented in this section. A simple CapsNet architecture is shown in Figure 2 [19]. It contains two halves. After three layers of processing, the first half produces a class prediction (two convolution layers and one fully connected layer). The first layer is a convolutional layer that uses a 9 × 9 kernel, 256 feature mappings, a stride of 1, and ReLU non-linearity to map a 28 × 28 image to a 6 × 6×256 volume. The second layer is a convolution capsule layer that uses a 9 × 9 kernel and stride of 2 to create a 6 × 6 × 256 volume. The volume is now divided into 32 layers of 8-dimension capsules, for a total of 6632 capsules, along its depth. A fully connected layer of 10 separate 16-dimension capsules, each of which represents an output class, makes up the third layer. It must be noted that only the PrimaryCaps and DigitCaps layers support dynamic routing. Finally, we generate a final prediction match to the embedding with the largest magnitude using the magnitude of these 16-dimension embeddings. Reconstruction as a regularization technique is implemented in the model's second half, also known as the decoder (Figure 3). Each class's 16-dimension embeddings are concatenated, with the exception of the winning class's vector components, which are all set to 0. In addition, 10-digit classes would mean a final embedding of length 160, which is then fed through 3 fully connected layers with 512, 1024, and 764 neurons, respectively. The final 784-dimension output is reshaped into a 28 × 28 reconstruction of the ground truth.

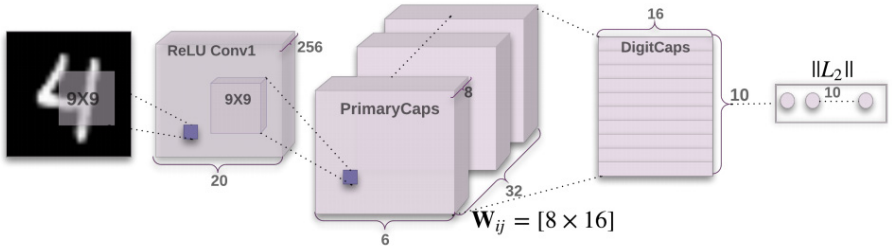

Figure 2. CapsNet architecture [19].

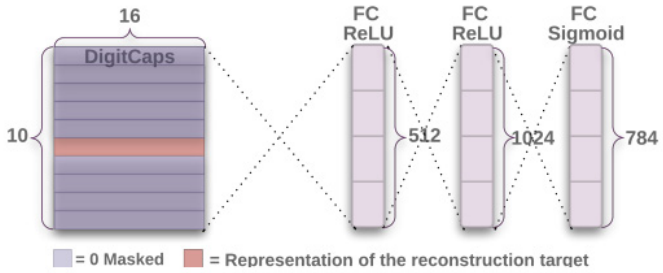

Figure 3. Reconstruction architecture [19].

The proposed model A-CapsNet, as illustrated in Figure 4, consists of three convolutional layers, two average-pooling layers, an initial capsule layer, and a digital capsule layer. The dynamic routing process of the proposed model A-CapsNet is referred to in the literature [18], and differs from Capsnet, as its average-pooling layer can reduce the computational complexity and avoid over fitting. Except for the primary capsule layer and digital capsule layer, the activation function LeakyReLU is used after each convolutional layer to enhance the non-linear mapping ability of the features.

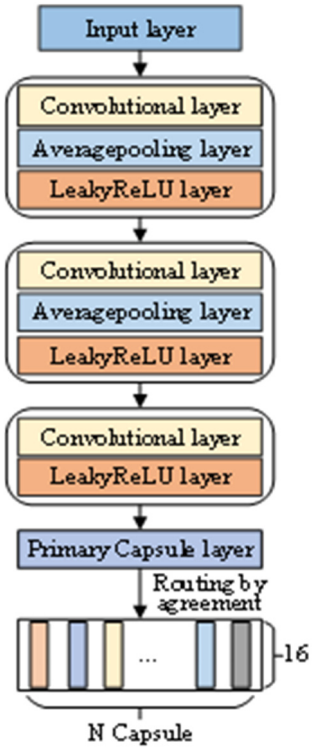

Figure 4. The structure of the proposed model A-CapsNet. It consists of three convolution layers, two pooling layers, an initial capsule layer, and a digital capsule layer.

4.2. Fundamentals

The Capsnet model is mainly composed of stacked dynamic routing layers. This section mainly explains the forward propagation principle of the dynamic routing layer.

The structure of the capsule network is similar to that of the full connection layer, as shown in Figure 5. u_1, u_2, u_m is the bottom capsule, v_1, v_2, v_N is the high-level capsule, W and C are the parameters to be adjusted, W is updated through network back propagation, and C is updated through the dynamic routing algorithm.

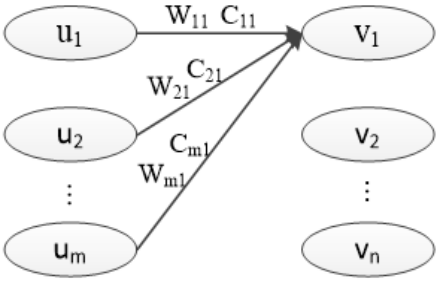

Figure 5. The structure of CapsNet.

4.2.1. Forward Propagation

The forward propagation formula is as shown in Equation (1).

$$\begin{cases} \hat{u}_j = W_j * u \\ s_j = \sum_i c_{ij} \cdot \hat{u}_j \\ v_j = squash(s_j) \end{cases} \tag{1}$$

where u_i is the output of the upper layer capsule and v_j is the output of this layer capsule. W_{ij} is not necessarily a parameter in the form of a full connection, and "*" is not necessarily a vector multiplication operation. Generally, "*" refers to a one-dimensional convolution operation, and W_{ij} refers to a convolution kernel, while c_{ij} is the coupling coefficient, which is obtained by the dynamic routing algorithm.

Squash is the activation function, as shown in Equation (2).

$$v_j = squash(s_j) = \frac{||s_j||^2}{1 + ||s_j||^2} \cdot \frac{s_j}{||s_j||} \tag{2}$$

It can be observed from the above formula that v_j and s_j have the same direction, and the module length of v_j is non-linearly mapped to (0,1) through the square () function, which increases the non-linear mapping ability of the model.

4.2.2. The Dynamic Routing Algorithm

C is obtained by the dynamic routing algorithm. In the back-propagation of the model, the value of C is not updated. The core formula of the dynamic routing algorithm is shown in Equation (3).

$$\begin{cases} c_{ij} = softmax(b_i) \\ b_{ij} \leftarrow b_{ij} + \langle \hat{u}_j, v_j \rangle \end{cases}, \tag{3}$$

where <,> represents the inner product operation. The dimension of b_i is the same and the initial value of $\hat{u}_j$ is 0. To clearly check the data flow in the capsule network, we draw the data flow diagram as shown in Figure 6.

The red line indicates the flow direction of the forward propagation data, and the green line indicates the flow direction of the dynamic routing data, and the variables with a purple circle background represent the variables that will be updated in the dynamic routing algorithm. As can be observed from Figure 6, the size of c_{ij} mainly depends on $\langle \hat{u}_j, v_j \rangle$, the similarity between v_j and $\hat{u}_j$.

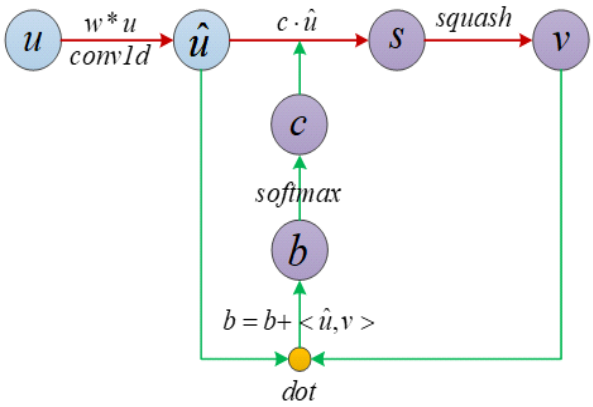

Figure 6. Data flow diagram.

The schematic diagram of the parameter updates for capsule network dynamic routing and backpropagation is shown in the Figure 7.

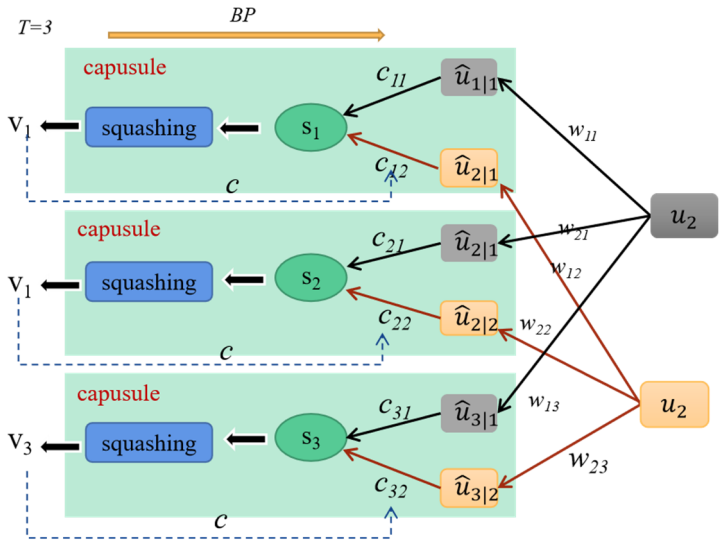

Figure 7. The dynamic routing process employed in A-CapsNet.

4.2.3. The Margin Loss of the Model A-CapsNet

The loss function formula is shown in Equation (4).

$$\begin{cases} L_j = y_j \cdot relu\left(m^+ - ||v_j||\right)^2 + \lambda(1 - \mathrm{y_j}) \cdot \mathrm{relu}(||\mathrm{v_j}|| - \mathrm{m}^-)^2 \\ m^+ = 1 - m \\ m^- = m \end{cases}. \tag{4}$$

where y is the predicted value and m and λ are the parameters to be adjusted, which are usually the following values: $m = 0.1$ and $\lambda = 0.5$.

5. Experimental Results of the Proposed Model A-CapsNet

This section includes the experimental settings, data division methods, and results analysis.

5.1. Experimental Settings

All studies are carried out on a powerful PC with 64GB of RAM with Windows 10. The processing speed is 2.10 GHz, the core is 40, and the logic processor is 80. TensorFlow [40] is used to implement all models. The following are the parameter combinations of the proposed model A-CapsNet: the convolution kernel size is set to 3 × 3, the step size is 1,

the pooling window size is set to 2 × 2, regardless of the average pooling, padding is 0, the activation function is LeakyReLU [41], and the optimizer is Adam [42]. When the emotion category is 7, the suggested model has 310,784 parameters.

5.2. Data Division Methods

It has been shown that data partitioning is a good approach for experiments with low data volumes [31]. Therefore, in order to train the proposed model A-CapsNet more adequately, four data partitioning methods, namely EIRD, EDRD, EICV and EDCV, are used, as shown in Table 2 and Figure 8.

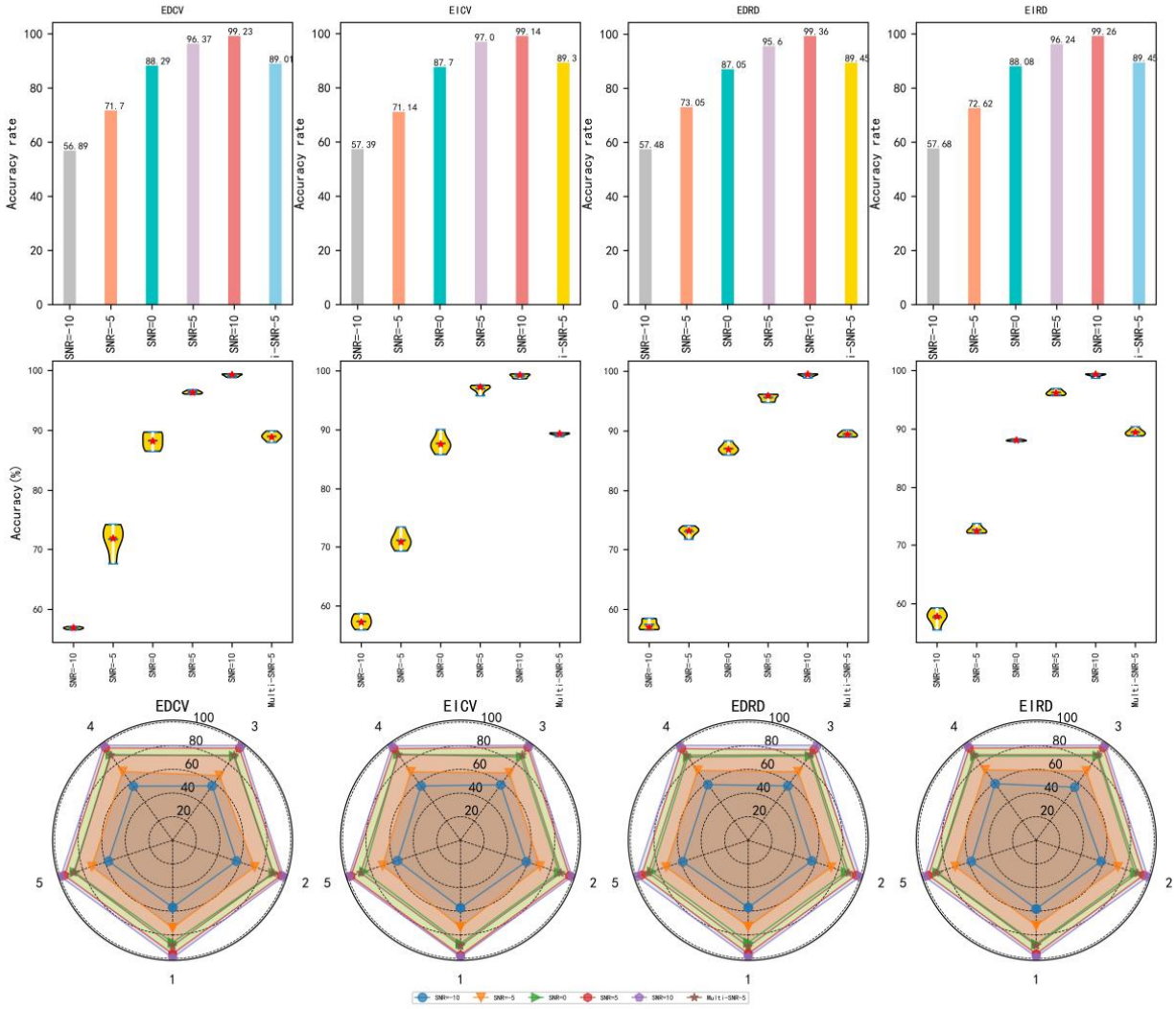

Figure 8. The experimental results of the augmented EMODB dataset with different signal-to-noise ratios.

Table 2. The performance of the proposed model A-CapsNet using the database EMODB under different SNRs with the batch size = 128.

Data Division	**Folds**	**SNR = −10**						**SNR = −5**					
		1	**2**	**3**	**4**	**5**	**Avg ± std**	**1**	**2**	**3**	**4**	**5**	**Avg ± std**
EDCV	K_folds = 5	56.76	56.95	56.95	56.64	57.13	56.89 ± 0.03	74.14	73.02	67.66	71.84	71.84	71.70 ± 4.81
EICV	K_folds = 5	57.26	58.69	58.13	56.88	56.01	57.39 ± 0.88	73.33	70.84	70.59	71.59	69.35	71.14 ± 1.72
EDRD	5 times	56.95	56.64	57.01	58.32	58.50	57.48 ± 0.59	73.08	73.96	71.78	73.15	73.27	73.05 ± 0.50
EIRD	5 times	58.07	57.82	55.58	59.25	57.69	57.68 ± 1.41	72.15	72.83	72.40	73.64	72.06	72.62 ± 0.33
Data Division	Folds	SNR = 0						SNR = 5					
		1	2	3	4	5	Avg ± std	1	2	3	4	5	Avg ± std
EDCV	K_folds = 5	86.54	89.72	88.22	89.60	87.35	88.29 ± 1.54	96.07	96.76	96.32	96.51	96.20	96.37 ± 0.06
EICV	K_folds = 5	87.60	87.73	87.29	90.03	85.86	87.70 ± 1.80	97.32	97.01	97.26	95.83	97.57	97.00 ± 0.37
EDRD	5 times	86.79	86.92	88.29	87.23	86.04	87.05 ± 0.53	96.14	96.07	94.83	95.89	95.08	95.60 ± 0.29
EIRD	5 times	87.91	87.98	88.29	88.10	88.10	88.08 ± 0.02	96.14	95.76	96.88	96.45	95.95	96.24 ± 0.16
Data Division	Folds	SNR = 10						Multi-SNR-5					
		1	2	3	4	5	Avg ± std	1	2	3	4	5	Avg ± std
EDCV	K_folds = 5	99.44	98.88	99.07	99.31	99.44	99.23 ± 0.05	89.32	89.89	88.90	88.87	88.05	89.01 ± 0.36
EICV	K_folds = 5	98.69	99.38	99.31	98.82	99.50	99.14 ± 0.10	89.37	89.46	89.26	89.50	88.91	89.30 ± 0.04
EDRD	5 times	98.94	99.56	99.50	99.25	99.56	99.36 ± 0.06	90.12	89.43	89.63	89.05	89.01	89.45 ± 0.17
EIRD	5 times	98.75	99.38	99.44	99.31	99.44	99.26 ± 0.07	89.05	90.36	89.46	89.53	88.85	89.45 ± 0.28

The EIRD is the acronym for emotion-independent random division. That is, all samples are randomly divided into five equal parts and four parts are selected as training data, while the remaining part is used as test data. The EDRD is the acronym for emotion-dependent random division. That is, the samples of each emotion category are divided into five equal halves. One sample is chosen at random as the test data for each section, while the samples are divided into five parts at random, and each part is used as test data in turn, with the remaining four parts serving as training data. The EDCV is the acronym for emotion-dependent cross-validation. That is, the samples of each emotion category are divided into five parts at random using this procedure. The test data are used for each part in turn, while the training data are used for the remaining four parts. The data division methods EICV and EDCV can ensure that all samples are used in the training and the remaining samples are used as training data. The EICV stands for emotion-independent cross-validation, which refers to all the testing processes.

5.3. Experimental Analysis

1. The performance of the model A-CapsNet improves as the SNR increases. Under the EIRD data division technique, the recognition performance with an SNR of −10 and 10 is 57.68 and 99.26, respectively, as illustrated in Figure 9. There is a rise of 40 percentage points, which is a significant improvement, and it suggests that the SNR has a significant impact on the performance of the model.

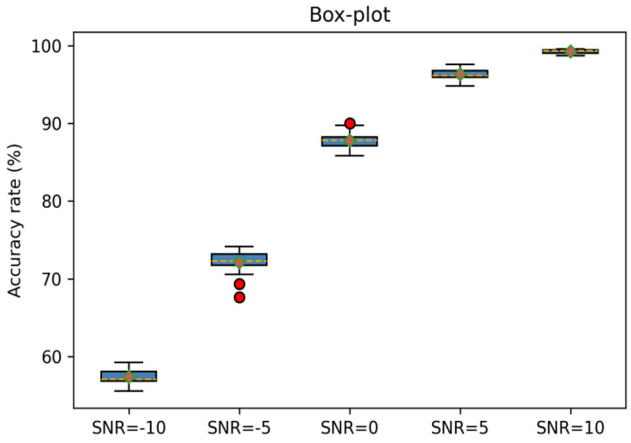

Figure 9. The boxplot of different SNRs.

2. Using the same database, the performance of the proposed model under different data division techniques varies, but no data division method outperforms the others, as illustrated in Figure 10. For example, when the SNR is −10, the model A-CapsNet performs best when using the data division technique EIRD, and when the SNR is 10, it performs best when using the data division method EDRD. With the improvement in the signal-to-noise ratio, no matter which data division method is used, the averages of the accuracies are improved, which is a relatively ideal recognition result. When the signal-to-noise ratio (SNR) is 10, there is minimal variation in the experimental performance of the four data division algorithms.

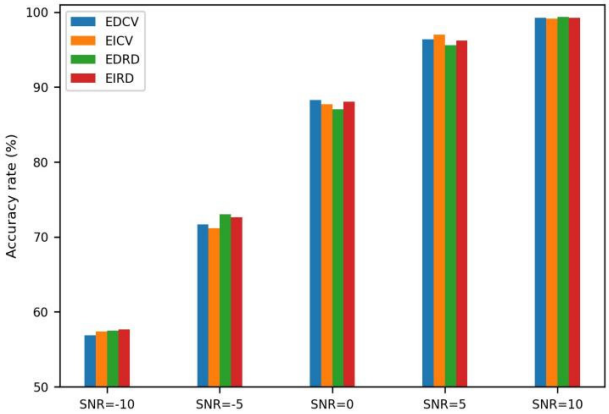

Figure 10. The results of four data division methods under different SNRs.

3. Figure 11 shows the outcomes of four data division techniques for varying model validation times using the Multi-SNR-5-EMODB database. It can be demonstrated that the suggested model A-CapsNet's performance was higher than 88.06 percent, which is an important recognition outcome. When compared to other data division techniques, the model A-CapsNet performs better under the data division approach EIRD. At the same time, it is also clear that the model's performance fluctuates significantly when cross-validated at various intervals, suggesting that data partitioning techniques may have an adverse effect on the model's performance. Therefore, it is important to assess the resilience and generalization of the model using various data partitioning techniques. However, no consistent conclusion has been obtained regarding which data partitioning method results in the best performance of the model. It can be said that it is necessary to evaluate the performance of the model by integrating the results of the four data partitioning methods.

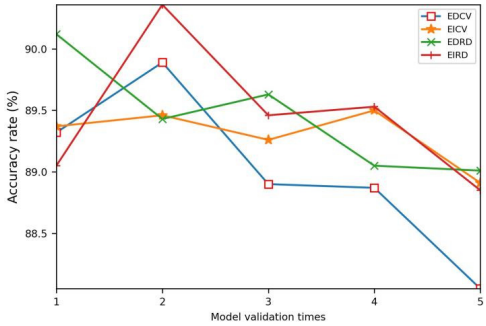

Figure 11. The results of different model validation times for the database Multi-SNR-5-EMODB using four data division methods.

In order to further highlight the advantages of the proposed model, Table 3 summarizes the performance improvement of the model A-CapsNet compared to the related peer methods with the baseline database EMODB. It shows that the suggested model A-CapsNet performs better with the EMODB databases compared to the findings of prior research, indicating that the proposed model M-CapsNet is more efficient. Additionally, in this experiment, the model parameters, feature dimensions, and training time are all relatively small. As a result, it complies with real-time system requirements.

Table 3. Performance of the proposed model A-CapsNet compared to the related models in the literature wi8th the baseline database EMODB.

Literature	CapCNN [43]	ACRNN [44]	AFSS [45]	WADAN [46]	SVM [47]	A-CapsNet
WAR	/	/	/	84.49	/	/
UAR	82.9	82.82	83.00	83.31	75.00	86.68

6. Conclusions

In this paper, A-CapsNet, an SER approach based on the average-pooling capsule network (A-CapsNet), is presented to address data scarcity and reduce over-fitting. The suggested model's operation decreases the number of parameters, while retaining more texture and background information for each feature. To increase the amount of samples and make the model more thoroughly trained, a data augmentation approach was devised. For the database EMODB, the model A-CapsNet is superior to the previous techniques.

The directions that may be further explored in the future are as follows: one may (1) conduct experiments on other corpora, and further analyze the advantages and disadvantages of the model; (2) attempt to apply the model A-CapsNet to conduct automatic speech recognition (ASR); (3) attempt to explore the optimal feature fusion technique; and (4) explore an acoustic model with better robustness and generalization for speech emotion recognition-related works.

Author Contributions: Y.Q. conducted the research study, and wrote the paper; H.Z. analyzed the data and helped to edit the paper; H.H. made suggestions to this paper and guided the research study. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant: 62066039) and Natural Science Foundation of Qinghai Province (grant: 2022-ZJ-925).

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. Jin, B.; Liu, G. Speech Emotion Recognition Based on Hyper-Prosodic Features. In Proceedings of the 2017 International Conference on Computer Technology, Electronics and Communication (ICCTEC), Dalian, China, 19–21 December 2017; pp. 82–87.
2. Li, G.; Tie, Y.; Qi, L. Multi-feature speech emotion recognition based on random forest classification and optimization. *Microelectron. Comput.* **2019**, *36*, 70–73.
3. Xu, L.; Liu, Y.; Hu, M.; Wang, X.; Reng, F. Spectrogram improves speech emotion recognition based on completely local binary patterns. *J. Electron. Meas. Instrum.* **2018**, *209*, 30–37.
4. Zhao, X.; Xu, X. Speech emotion recognition combining shallow learning and deep learning models. *Comput. Appl. Softw.* **2020**, *37*, 114–118+182.
5. Cheng, Y.; Chen, Y.; Cheng, Y.; Yang, Y. Speech emotion recognition with embedded attention mechanism combined with hierarchical context. *J. Harbin Inst. Technol.* **2019**, *51*, 100–107.
6. Ramakrishnan, S.; Emary, I. Speech emotion recognition approaches in human computer interaction. *Telecommun. Syst.* **2013**, *52*, 1467–1478. [CrossRef]
7. John, K.; Saurous, R.A. Emotion recognition from human speech using temporal information and deep learning. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 937–940.
8. Lu, G.; Cheng, X.; Li, X.; Yan, J.; Li, H. Multimodal emotional feature fusion method based on genetic algorithm. *J. Nanjing Univ. Posts Telecommun. (Nat. Sci. Ed.)* **2019**, *184*, 44–50.
9. Ma, J.; Sun, Y.; Zhang, X. Multi-modal emotion recognition based on fusion of speech signal and EEG signal. *J. Xidian Univ.* **2019**, *46*, 143–150.
10. Hu, H.; Xu, M.-X.; Wu, W. GMM supervector based SVM with spectral features for speech emotion recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Honolulu, HI, USA, 15–20 April 2007; pp. 413–416.

11. Yu, Y.; Huang, F.; Liu, Y. Speech emotion recognition based on feature dimensionality reduction and parameter optimization. *J. Yanbian Univ. (Nat. Sci. Ed.)* **2020**, *46*, 49–54.
12. Mao, X.; Chen, L.; Fu, L. Multi-level speech emotion recognition based on HMM and ANN. In Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering, Los Angeles, LA, USA, 31 March–2 April 2009; pp. 225–229.
13. Kansizoglou, I.; Misirlis, E.; Tsintotas, K.; Gasteratos, A. Continuous Emotion Recognition for Long-Term Behavior Modeling through Recurrent Neural Networks. *Technologies* **2022**, *10*, 59. [CrossRef]
14. Song, M.; Chen, C.; You, M. Audio-visual based emotion recognition using tripled hidden Markov model. In Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, QC, Canada, 17–21 May 2004; pp. 877–880.
15. Vydana, H.K.; Kumar, P.P.; Krishna, K.S.R.; Vuppala, A.K. Improved emotion recognition using GMM-UBMs. In Proceedings of the of 2015 IEEE International Conference on Signal Processing and Communication Engineering Systems, Guntur, India, 2–3 January 2015; pp. 53–57.
16. Chen, X.; Han, W.; Ruan, H.; Liu, J.; Li, H.; Jiang, D. Sequence-to-sequence modelling for categorical speech emotion recognition using recurrent neural network. In Proceedings of the 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), Beijing, China, 20–22 May 2018; pp. 1–4.
17. Bertero, D.; Fung, P. A first look into a convolutional neural network for speech emotion detection. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5115–5119.
18. Khan, N.; Ullah, A.; Haq, I.U.; Menon, V.G.; Baik, S.W. SD-Net: Understanding overcrowded scenes in real-time via an efficient dilated convolutional neural network. *J. Real-Time Image Process.* **2021**, *18*, 1729–1743. [CrossRef]
19. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic routing between capsules. In *NeurIPS Proceedings: Advances in Neural Information Processing Systems 30 (NIPS 2017)*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 3856–3866.
20. Li, R.; Wu, Z.; Jia, J.; Zhao, S.; Meng, H. Dilated residual network with multi-head self-attention for speech emotion recognition. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6675–6679.
21. Tao, J.H.; Liu, F.Z.; Zhang, M.; Jia, H.B. Design of speech corpus for mandarin text to speech. In Proceedings of the Blizzard Challenge 2008 Workshop, Brisbane, Australia, 21 September 2008; p. 1.
22. Weninger, F.; Wöllmer, M.; Schuller, B. Emotion Recognition in Naturalistic Speech and Language—A Survey. In *Emotion Recognition: A Pattern Analysis Approach*; John Wiley & Sons Inc.: Hoboken, NJ, USA, 2015; pp. 237–267.
23. Kim, Y.; Provost, E.M. ISLA: Temporal segmentation and labeling for audio-visual emotion recognition. *IEEE Trans. Affect. Comput.* **2019**, *10*, 196–208. [CrossRef]
24. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Zafeiriou, S. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016.
25. Janovi, P.; Zou, X.; Kküer, M. Speech enhancement based on Sparse Code Shrinkage employing multiple speech models. *Speech Commun.* **2012**, *54*, 108–118. [CrossRef]
26. Anagnostopoulos, C.N.; Iliou, T.; Giannoukos, I. Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. *Artif. Intell. Rev.* **2015**, *43*, 155–177. [CrossRef]
27. Lee, C.-C.; Mower, E.; Busso, C.; Lee, S.; Narayanan, S. Emotion recognition using a hierarchical binary decision tree approach. *Speech Commun.* **2011**, *53*, 1162–1171. [CrossRef]
28. Langari, S.; Marvi, H.; Zahedi, M. Efficient Speech Emotion Recognition Using Modified Feature Extraction. *Inform. Med. Unlocked* **2020**, *20*, 100424. [CrossRef]
29. Qing, G.; Zh, Z.; Da, X.; Zhi, X. Review on speech emotion recognition research. *CAAI Trans. Intell. Syst.* **2020**, *15*, 1–13.
30. Sun, Y.; Song, C. Emotional speech feature extraction and optimization of phase space reconstruction. *Xi'an Dianzi Keji Daxue Xuebao J. Xidian Univ.* **2017**, *44*, 162–168.
31. Peng, S.; Yun, J.; Cheng, Z.; Li, Z. Speech emotion recognition using sparse feature transfer. *J. Data Acquisit. Process.* **2016**, *31*, 325–330. [CrossRef]
32. Gideon, J.; Mclnnis, M.G.; Provost, E.M. Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG). *IEEE Trans. Affect. Comput.* **2021**, *12*, 1055–1068. [CrossRef]
33. Sarker, M.K.; Alam, K.M.R.; Arifuzzaman, M. Arifuzzaman Emotion recognition from speech based on relevant feature and majority voting. In Proceedings of the 2014 International Conference on Informatics, Electronics & Vision (ICIEV), Dhaka, Bangladesh, 23–24 May 2014; pp. 1–5.
34. Raju, V.N.G.; Lakshmi, K.P.; Jain, V.M.; Kalidindi, A.; Padma, V. Study the influence of normalization/transformation process on the accuracy of supervised classification. In Proceedings of the 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 20–22 August 2020; pp. 729–735.
35. Wang, L.; Dang, J.; Zhang, L.; Guan, H.; Li, X.; Guo, L. Speech emotion recognition by combining amplitude and phase information using convolutional neural network. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 1611–1615.
36. Xi, E.; Bing, S.; Yang, J. Capsule Network Performance on Complex Data. *arXiv* **2017**, arXiv:1712.03480.

37. Xiang, C.Q.; Zhang, L.; Tang, Y.; Zou, W.B.; Xu, C. MS-CapsNet: A novel multi-scale capsule network. *IEEE Signal Process. Lett.* **2018**, *25*, 1850–1854. [CrossRef]
38. Wu, X.X.; Liu, S.X.; Cao, Y.W.; Li, X.; Yu, J.W.; Dai, D.Y. Speech emotion recognition using capsule network. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6695–6699.
39. Nair, P.; Doshi, R.; Keselj, S. Pushing the Limits of Capsule Networks. *arXiv* **2021**, arXiv:2103.08074.
40. Ertam, F.; Aydın, G. Data classification with deep learning using Tensorflow. In Proceedings of the 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 5–8 October 2017; pp. 755–758.
41. Jiang, T.; Cheng, J. Target recognition based on CNN with LeakyReLU and PReLU activation functions. In Proceedings of the International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC), Beijing, China, 15–17 August 2019; pp. 718–722.
42. Chen, K.; Ding, H.; Huo, Q. Parallelizing Adam optimizer with blockwise model-update filtering. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 3027–3031.
43. Wen, X.C.; Liu, K.H.; Zhang, W.M.; Jiang, K. The application of capsule neural network based CNN for speech emotion recognition. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 9356–9362.
44. 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]
45. Cirakman, O.; Gunsel, B. Online speaker emotion tracking with a dynamic state transition model. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 307–312.
46. Yi, L.; Mak, M.-W. Improving speech emotion recognition with adversarial data augmentation network. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 172–1844. [CrossRef] [PubMed]
47. Sugan, N.; Sai Srinivas, N.S.; Kar, N.; Kumar, L.S.; Nath, M.K.; Kanhe, A. Performance comparison of different cepstral features for speech emotion recognition. In Proceedings of the 2018 International CET Conference on Control, Communication, and Computing (IC4), Thiruvananthapuram, India, 5–7 July 2018; pp. 266–271.