entropy

MDPI

Article

Strong Generalized Speech Emotion Recognition Based on Effective Data Augmentation

Huawei Tao [1,2,*], Shuai Shan [1], Ziyi Hu [1], Chunhua Zhu [1,2] and Hongyi Ge [1,2]

[1] Key Laboratory of Food Information Processing and Control, Ministry of Education, Henan University of Technology, Zhengzhou 450001, China
[2] Henan Engineering Laboratory of Grain IOT Technology, Henan University of Technology, Zhengzhou 450001, China
* Correspondence: thw@haut.edu.cn

Abstract: The absence of labeled samples limits the development of speech emotion recognition (SER). Data augmentation is an effective way to address sample sparsity. However, there is a lack of research on data augmentation algorithms in the field of SER. In this paper, the effectiveness of classical acoustic data augmentation methods in SER is analyzed, based on which a strong generalized speech emotion recognition model based on effective data augmentation is proposed. The model uses a multi-channel feature extractor consisting of multiple sub-networks to extract emotional representations. Different kinds of augmented data that can effectively improve SER performance are fed into the sub-networks, and the emotional representations are obtained by the weighted fusion of the output feature maps of each sub-network. And in order to make the model robust to unseen speakers, we employ adversarial training to generalize emotion representations. A discriminator is used to estimate the Wasserstein distance between the feature distributions of different speakers and to force the feature extractor to learn the speaker-invariant emotional representations by adversarial training. The simulation experimental results on the IEMOCAP corpus show that the performance of the proposed method is 2–9% ahead of the related SER algorithm, which proves the effectiveness of the proposed method.

Keywords: speech emotion recognition; data augmentation; multi-channel feature extractor; Wasserstein distance; feature distributions; speaker-invariant emotional representations

Citation: Tao, H.; Shan, S.; Hu, Z.; Zhu, C.; Ge, H. Strong Generalized Speech Emotion Recognition Based on Effective Data Augmentation. *Entropy* **2023**, *25*, 68. https://doi.org/10.3390/e25010068

Academic Editor: Yuan Zong

Received: 10 December 2022
Revised: 24 December 2022
Accepted: 26 December 2022
Published: 30 December 2022

1. Introduction

Speech emotion recognition (SER) plays an important role in Human-Computer Interaction (HCI) systems, and it has become increasingly involved in a wide variety of industrial applications. SER, for instance, can be used to detect the presence and severity of a patient's distress without requiring any intervention from a human [1]. An intelligent customer service system in the call center will transfer a call to a human customer service representative if it recognizes that the customer expresses a negative emotion [2]. In the field of education, the use of SER can greatly improve teaching and learning outcomes [3]. It is of great practical importance to conduct research for SER to make HCI more intelligent and humane.

Deep learning has become a viable technical solution for SER, and SER methods based on deep learning have achieved better performance in a variety of scenarios. The training of high-performing models requires a large number of samples. Manually labeling emotion labels is, however, time-consuming and costly, limiting the size of the existing emotion corpus. SER is limited by the lack of a large-scale labeled emotion corpus. Researchers have attempted to solve the problem of small emotional corpus samples using data augmentation methods in recent years, with some success. Aldeneh et al. [4] varied the speed of the speech, creating two additional speech copies of 0.9× and 1.1× speed to increase the

size of the training data and achieved a 2.8% recognition rate improvement on the MSP-IMPROV [5] corpus. Li et al. [6] also used variable speed speech to expand the corpus and achieved better performance in the speaker-independent case. Braunschweiler et al. [7] used speed augmentation and volume perturbation, and the experimental results showed that data augmentation is still effective in improving performance in the case of cross-corpus recognition. The authors of [8] also reported significant performance gains using data augmentation which is noise injection and volume perturbation in the EMODB [9] corpus. Liu et al. [10] used SpecAugment, a data augmentation algorithm proposed by Google [11], to enrich a few classes of emotional speech samples to balance the corpus. To some extent, these works fill the research gap regarding data augmentation in the SER field. However, fewer data augmentation methods have been explored, and no scholars have yet made detailed and specific studies on acoustic-based data enhancement methods in the field of SER.

Besides directly altering the acoustic properties of the original speech to create new speech to augment the training data, some scholars have also used GANs [12] to address the problem of data scarcity in SER. In [13], the authors used Balancing GAN [14] to generate speech spectrograms of target categories to increase the number of training samples. Since it is difficult for generators to generate high-dimensional samples directly, Yi et al. [15] proposed ADAN (Adversarial Data Augmentation Network), which combines Autoencoder techniques with GAN to generate low-dimensional emotional vectors in latent space. However, it is difficult to train a generator that is able to generate accurate emotional samples from the target categories due to the confusion among some specific emotions [10] and the possibility of mode collapse of the generators [16].

Log–Mel spectrograms have the advantage of high correlation and are widely used in various speech tasks [17–20]. We also use log–Mel spectrograms as the input to the proposed end-to-end SER model. In this paper, we first investigate the impact of acoustic-based data augmentation methods on SER through a simple model, and, based on that, we propose a strong generalized speech emotion recognition model based on effective data augmentation. There are three components to the model: a feature extractor, an emotion classifier, and a discriminator. We utilize a multi-channel feature extractor consisting of multiple sub-networks to extract emotional representations under multiple data augmentations. The discriminators were used to estimate the Wasserstein distance [21] between different speakers' emotional representations. By reducing this distance through adversarial training, the feature extractor can learn the speaker-invariant emotional representations. The main contributions of this paper are three points:

- The effectiveness of acoustic-based data augmentation methods is evaluated in SER.
- There is a feature extractor architecture proposed that can make better use of data augmentation methods, which consists of multiple sub-networks, and a model weight parameter sharing strategy is applied among the sub-networks. The output feature maps of each sub-network are fused to generate emotional representations.
- In order to generalize the emotional representations, the Wasserstein distance is used to measure the distribution of emotional representations among speakers. The distribution of representations is approximated in hidden space by adversarial training. In this way, the feature extractor learns the speaker-invariant emotional representations.

2. Proposed Method

2.1. Strong Generalized Speech Emotion Recognition Model Based on Effective Data Augmentation

The structure of the proposed model is shown in Figure 1. There are three modules in the model: a feature extractor, an emotion classifier, and a discriminator. Log–Mel spectrograms of the original and augmented speech are fed into a multi-channel feature extractor, which is responsible for extracting the emotional representations in the spectrograms. Emotional representations are input to the emotion classifier for classification and to the discriminator for estimating the Wasserstein distance. In the emotion classifier, there are two fully connected layers (256:64, 64:4) as well as a softmax layer and dropout

set to 0.5. In the discriminator, there are two fully connected layers (265:64, 64:16, 16:1). The activation function used in these two modules is the Rectified Linear Unit (ReLU). A detailed description of the feature extractor is provided in the following Section 2.3. After the model has been trained, the feature extractor and emotion classifier can form a complete SER system.

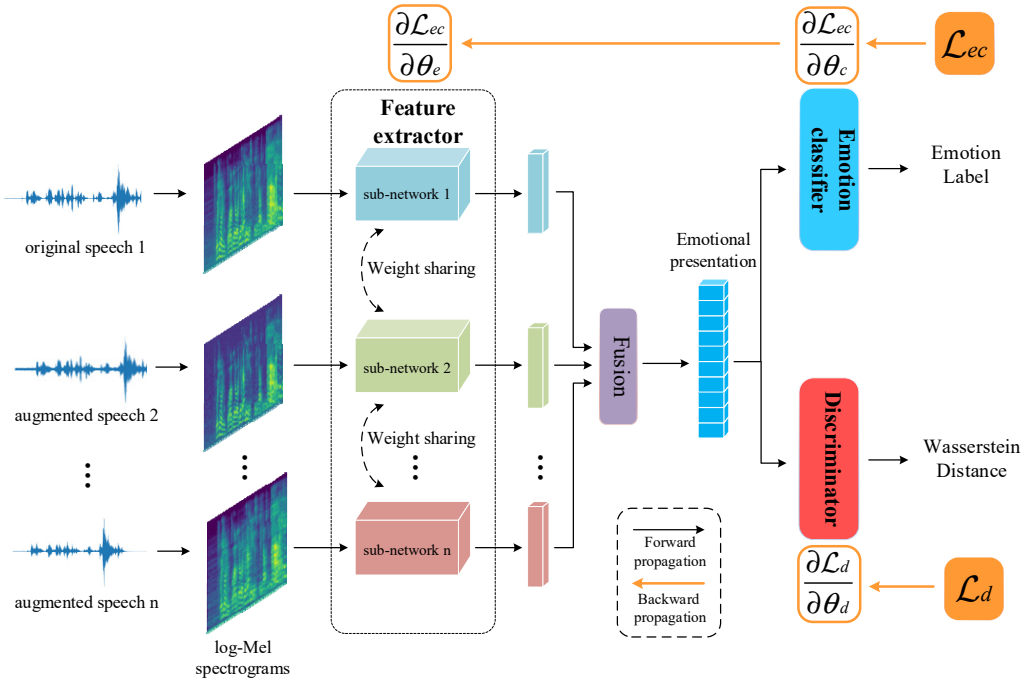


Figure 1. The structure of the proposed strong generalized speech emotion recognition model based on effective data augmentation.

2.2. Data Augmentation

Following a summary of acoustic data augmentation methods previously applied to the SER and other speech tasks, six acoustic data augmentation methods were selected for analysis: speed augmentation [4,6,7], noise injection [8], time shifting [22], resampling [23], pitch shifting [24], and reverberation augmentation [25]. Figure 2 shows speech waveforms and spectrograms with different data augmentation methods. It should be noted that speed augmentation and reverberation augmentation change the length of the speech, but we unify the length of their waveforms and spectrograms in Figure 2 for the sake of comparison. A detailed description and implementation of each type of data augmentation are provided in the following Section 3.2.

2.3. Feature Extractor

The feature extractor consists of n sub-networks with the same parameter settings, and each sub-network receives log–Mel spectrograms for different augmented speech as inputs. The Residual Network [26], which is commonly used in SER, is selected as the main part of the sub-network in this study based on a literature review [27–30]. Figure 3 shows a specific sub-network setup where log–Mel spectrograms are fed into two parallel convolution layers with convolutional kernels of (10, 2) and (2, 8), respectively. Such a convolution kernel setup can fully extract the time and frequency domain information of log–Mel spectrograms [31]. Five consecutive residual blocks [26] are used to extract deep emotional information from the concatenated outputs of the two convolutional layers described above. The feature map size is eventually compressed using adaptive average pooling to retain only relevant information. In order to facilitate sharing of some learned knowledge, such as low-level acoustic features, between channels of the feature extractor while accelerating the convergence speed, a model weight parameter sharing

strategy is applied in Conv2D_A, Conv2D_B and Residual_Block_1 between each sub-network. The output feature maps of each sub-network are fused into the emotional representations needed for the subsequent classification and metric tasks based on their weighting coefficient. The emotional representation *EmoRep* is calculated by:

$$EmoRep = \alpha_1 * f_{\theta 1}(x_1) + \alpha_2 * f_{\theta 2}(x_2) + \ldots\ldots + \alpha_n * f_{\theta n}(x_n) \quad (1)$$

where α_k is the weighting coefficient, and $f_{\theta k}$ denotes the function of a sub-network.

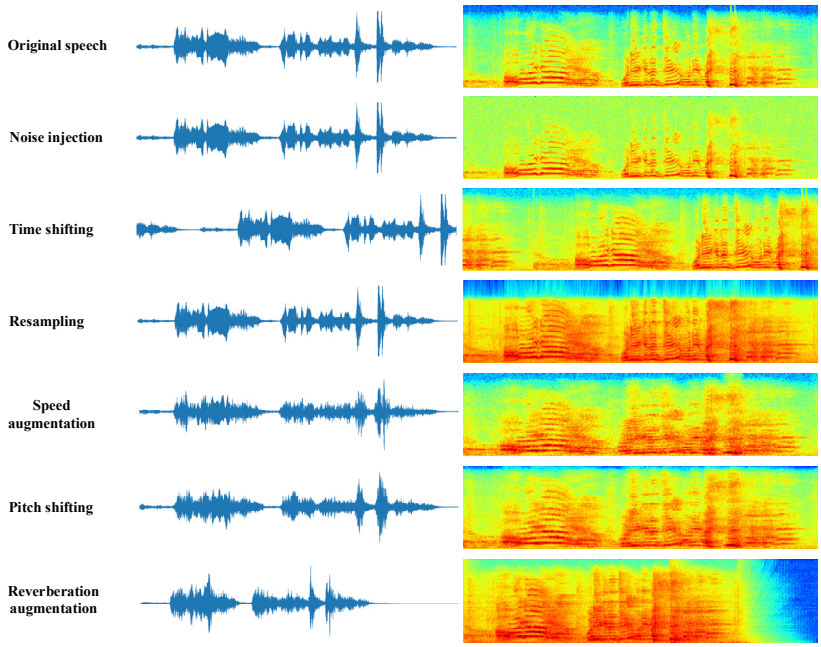


Figure 2. Speech waveforms and spectrograms with different data augmentation methods.

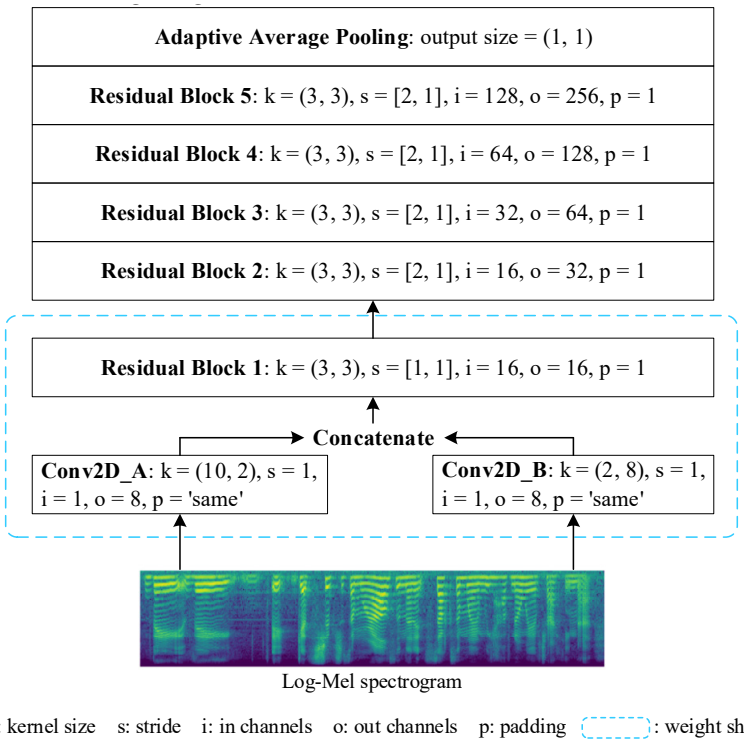


Figure 3. The sub-network structure in the feature extractor used by the proposed model.

2.4. Measuring Distance of Emotional Representation Distribution

The Wasserstein distance was used to measure the distance between the distributions of emotional representations among the speakers. In high-dimensional space, if two distributions do not overlap or the overlap can be ignored, the KL and JS divergence do not reflect the distance between distributions or provide the gradient. In [21], the authors solved this problem by using the Wasserstein distance rather than the KL and JS divergences in the original GAN. Due to the superiority of the Wasserstein distance as a distribution measure, it is used in this paper to measure the distance between emotional representations. Given probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, the Wasserstein distance between them is defined as follows:

$$W(\mathbb{P}_1, \mathbb{P}_2) = \inf_{\gamma \sim \prod(\mathbb{P}_1, \mathbb{P}_2)} \mathbb{E}_{(x,y)\sim\gamma}[||x - y||] \tag{2}$$

where γ denotes the joint distribution of samples x and y, $\prod(\mathbb{P}_1, \mathbb{P}_2)$ denotes the set of all possible joint distributions of $\mathbb{P}_1$ and $\mathbb{P}_2$ combined, and $||x - y||$ is the inter-sample distance. Based on the joint distribution γ, the expectation value of the distance between the sample pair is $\mathbb{E}_{(x,y)\sim\gamma}[||x - y||]$. Wasserstein distance between $\mathbb{P}_1$ and $\mathbb{P}_2$ is the infimum of the expectation value in all possible joint distributions. Using the Kantorovich-Rubinstein duality, Equation (2) can be transformed into:

$$W(\mathbb{P}_1, \mathbb{P}_2) = \sup_{||f||L \leq 1} \mathbb{E}_{x\sim\mathbb{P}_1}[f(x)] - \mathbb{E}x_{\sim\mathbb{P}_2}[f(x)] \tag{3}$$

where $||f||L \leq 1$ indicates that f is 1-Lipschitz continuous.

2.5. Generalization of Emotional Representation in Adversarial Training

There are individual differences among speakers, such as timbre, expressive habits, etc., which make it difficult for the model to learn robust emotional representations that can cover all speakers. In order to generalize the emotional representations, we use adversarial training that forces the feature extractor to learn speaker-invariant emotional representations. We train the modules in the model alternatively, which consists of two steps: (1) training the discriminator; (2) training the feature extractor and emotion classifier. The details of these two steps are described in the following.

2.5.1. Training of Discriminator

The discriminator is primarily responsible for estimating the Wasserstein distance between the distribution of the emotional representations of the source domain and the target domain speaker. We can represent all possible functions f in Equation (3) with discriminators since neural networks can fit various functions. Then the Wasserstein distance between the source domain and the target domain speakers can be calculated by:

$$W(\mathbb{P}_s, \mathbb{P}_t) = \sup_{||f_D||L \leq 1} \mathbb{E}_{x\sim\mathbb{P}_s}[f_d(f_e(x))] - \mathbb{E}_{x\sim\mathbb{P}_t}[f_d(f_e(x))] \tag{4}$$

where f_d denotes the discriminator and f_e denotes the feature extractor. To ensure that the discriminator function is 1-Lipschitz continuous, in [21], the authors propose to clip the weights of the discriminator within a compact space $[-c, c]$ after each gradient update. However, Gulrajani et al. [32] found that weight clipping would result in gradient explosion or vanishing. In order to enhance the stability of gradients, Gulrajani et al. proposed the use of gradient penalties instead of weight clipping. As suggested in [32], gradient penalties are used to make the discriminator function 1-Lipschitz continuous in this paper. The gradient penalty term is defined as follows:

$$GP = \lambda \mathbb{E}_{x\sim\chi}[||\nabla f_d(f_e(x))||_2 - 1]^2 \tag{5}$$

where λ is the penalty factor, χ denotes the sample space distribution, and ∇f_d is the gradient of the discriminator.

Then the loss function $\mathcal{L}_d$ of the discriminator is shown as follows:

$$\mathcal{L}_d = \mathbb{E}_{x\sim\mathbb{P}_t}[f_d(f_e(x^t))] - \mathbb{E}_{x\sim\mathbb{P}_s}[f_d(f_e(x^s))] + GP \tag{6}$$

The discriminator weight parameters are updated by minimizing Equation (6). In this training step, the weight parameters of the feature extractor and the emotion classifier are frozen.

2.5.2. Training of Feature Extractor and Emotion Classifier

The feature extractor is responsible for extracting emotional representations under multiple data augmentations, and the emotion classifier gives the labels to which the representations belong. In this step, we force the feature extractor to learn the speaker-invariant emotional representations, and significant classifiable information is retained in those representations. The loss function $\mathcal{L}_e$ of the feature extractor and the loss function $\mathcal{L}_c$ of the emotion classifier are as follows:

$$\mathcal{L}_e = \mathbb{E}_{x\sim\mathbb{P}_s}[f_d(f_e(x^s))] - \mathbb{E}_{x\sim\mathbb{P}_t}[f_d(f_e(x^t))] \tag{7}$$

$$\mathcal{L}_c = -\sum_i \sum_{c=1, x\in X^s}^{M} y_{ic} \log[f_c(f_e(x^s))] \tag{8}$$

In Equation (8), f_c denotes the emotion classifier and y_{ic} denotes the sample label. Then the joint loss function $\mathcal{L}_{ec}$ of the feature extractor and the emotion classifier is as follows:

$$\mathcal{L}_{ec} = \beta\mathcal{L}_e + \mathcal{L}_c \tag{9}$$

where β is the coefficient that controls the balance between discriminative and generalized representation learning. By minimizing Equation (9), the weight parameters of the feature extractor and emotion classifier are updated. In this step, the weight parameters of the discriminator are frozen.

Following adversarial training, a feature extractor that generalizes emotional representations while retaining classifiable information is developed, as well as an emotion classifier with superior classification performance.

3. Experiments

We evaluated the effectiveness of the proposed data augmentation methods in SER in Section 3.2. In Section 3.3, data augmentation methods that can significantly improve SER performance are applied to the proposed strong generalized speech emotion recognition model based on effective data augmentation.

3.1. Speech Emotion Corpus

To evaluate the proposed model, we conducted our experiments on Interactive Emotional Dyadic Motion Capture (IEMOCAP) [33]. The IEMOCAP contains 10,039 utterances annotated by at least three expert evaluators with a total length of approximately 12 h, which are divided into 9 emotions. There are five sessions in IEMOCAP, each with two speakers interacting (one male and one female).

To be consistent with previous studies, our experiment considered four emotions: happy, angry, sad, and neutral, and merged excitement into the happy class. The total number of utterances was 5531. Due to the variable length of each utterance, we segmented each utterance into two-second segments for extraction of log–Mel spectrograms. Log–Mel spectrogram features of each speech segment were extracted using 64 sets of Mel filters, 25 ms Hamming windows, and 10 ms window shifts. The experiments all adopted a speaker-independent strategy, i.e., a 5-fold Leave-One-Session-Out cross-validation strategy. For each fold, four sessions were selected for training and one session for testing. Weighted

accuracy (WA) is used as a performance evaluation metric, which is commonly used in the SER field.

3.2. Experiment of Data Augmentation

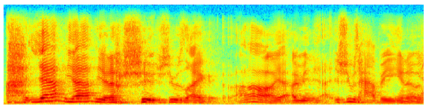
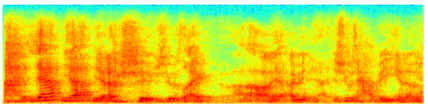
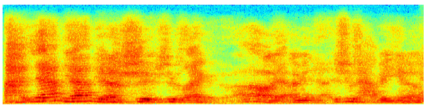
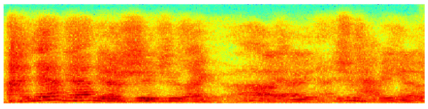
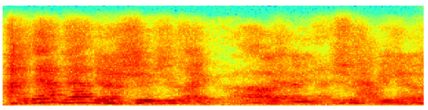
In this section, a simple model is used to test the effects of the six data augmentation methods on SER under a variety of parameter settings.

The model used in this section consists of a feature extraction component and a classification component. The feature extraction component is the same as the sub-network setup in Section 2.3, and the classification component consists of fully connected layers and a softmax layer. A 1:1 ratio of original and augmented data amounts are used in the training set, and the test set data are not augmented. Adam optimizer is chosen with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-6}. The batch size is 64 and the number of training epochs is 50 in each fold.

3.2.1. Speed Augmentation

We change the speed of the original speech to produce new speech to increase the training set. Table 1 shows the experimental results for the four speed factors. The first line of Table 1 shows that the WA of the model without data augmentation is 60.43%, and we use this result as the baseline. We first verified the impact of slow speech augmented in the training set on performance. Slowing down the speech rate has a small or no effect on performance improvement, according to the results. In subsequent experiments, two acceleration strategies are used. Performance is improved more significantly when speech speed is accelerated in addition to the training set. Compared to the baseline, WA improved by 2.01% and 0.93% at 1.5× and 2.0× speed, respectively.

Table 1. Experimental results of speed augmentation. Speed factor represents the speed of augmented speech.

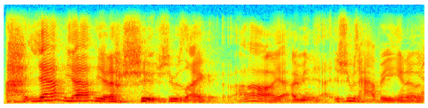
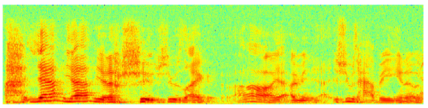
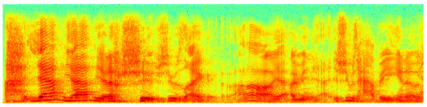
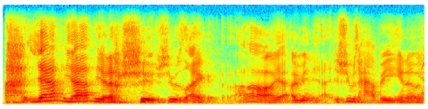
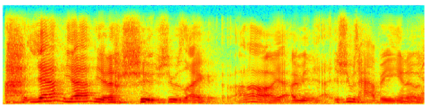
Speed Factor	WA (%)	Spectrogram
-	60.43	
0.5	60.57	
0.8	60.38	
1.5	62.44	
2.0	61.36	

3.2.2. Noise Injection

To create new speech, we added White Gaussian Noise (WGN) with a mean of 0 and standard deviation of 1 to the original speech. We generated augmented speech by controlling the ratio of speech signal to WGN. Four signal-to-noise ratio (SNR) strategies were employed, and the experimental results are shown in Table 2. It should be noted that SNR here refers to the ratio of speech signal to WGN added. When the SNR was set to 30 dB, the WA decreased by 0.17% and improved by 0.99% when it was set to 60 dB. When the SNR was set to 90 dB and 120 dB, the WA also improved compared with the baseline.

Noise interference will fade the spectrogram pattern, as shown in the spectrograms. As a result of the noise injection, the sentiment details are obscured to some extent, and it may appear that the performance improvement is not significant after augmentation at different SNR.

Table 2. Experimental results of noise injection. The SNR represents the ratio of speech signal to WGN added.

SNR	WA (%)	Spectrogram
-	60.43	
30 dB	60.26	
60 dB	61.42	
90 dB	60.44	
120 dB	61.25	

3.2.3. Time Shifting

Time shifting refers to rolling the speech signal in the time domain. A total of four different time shifting strategies were compared, and the experimental results are shown in Table 3. With strategies of 60% and 80% shifting ratios, respectively, WA improved by 0.96% and 1.68% over the baseline. However, WA decreases at shifting ratios of 20% and 40%. As a result of rolling the speech signal, the overall coherence of speech is disrupted, which causes confusion for the classifier.

3.2.4. Resampling

Resampling means changing the sampling rate of speech and creating re-sampled speech to augment the training set. We change the sampling rate of speech from 16,000 Hz to an intermediate sampling rate, and then back to 16,000 Hz. The experimental results are shown in Table 4 for four intermediate sampling rates. When the intermediate sampling rate is 11,000 Hz and 13,000 Hz, that is, when the intermediate sampling rate is smaller than the original sampling rate, WA performs better by 0.5% and 0.8%, respectively. As can be seen from the spectrogram, high-frequency details of speech are lost in these cases. A larger intermediate sampling rate was used in subsequent experiments. When the intermediate sampling rate was 18,000 Hz, WA improved by 1.4%. However, when the intermediate sampling rate was 20,000 Hz, WA decreased by 0.25%. As a result of changing the sampling rate of speech, some information is lost, which is a disadvantage for SER.

Table 3. Experimental results of time shifting. Shifting ratio represents the proportion of speech signal rolling in the time domain.

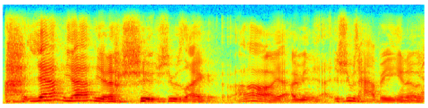
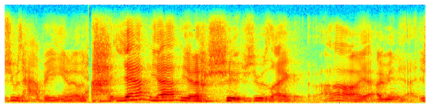
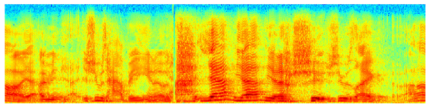
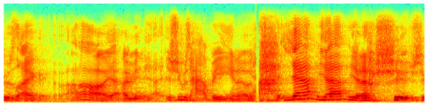
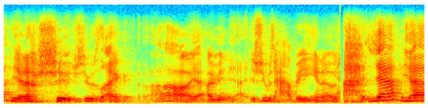
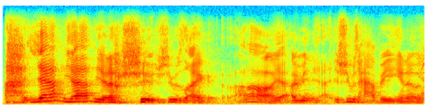
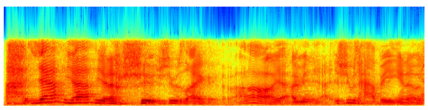
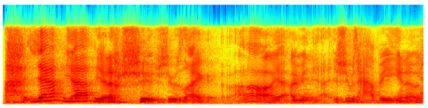
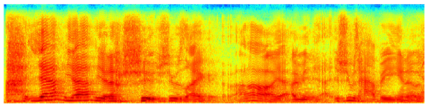
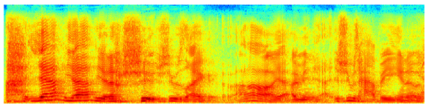
Shifting Ratio	WA (%)	Spectrogram
-	60.43	
20%	59.73	
40%	60.29	
60%	61.39	
80%	62.11	

Table 4. Experimental results of resampling.

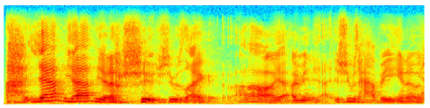
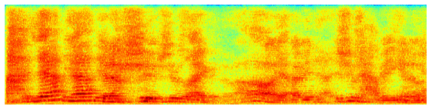
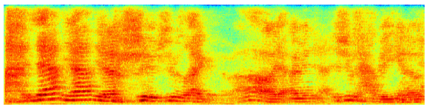
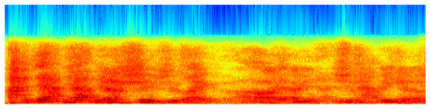
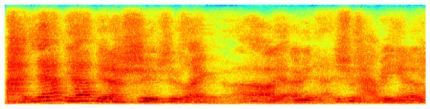
Intermediate Sampling Rate	WA (%)	Spectrogram
-	60.43	
11,000 Hz	60.93	
13,000 Hz	61.23	
18,000 Hz	61.81	
20,000 Hz	60.18	

3.2.5. Pitch Shifting

We use Python's Librosa toolkit to change the pitch of the original speech. Pitch is altered by setting the parameters n_steps and bins_per_octave, where n_steps is how many steps to shift and bins_per_octave is how many steps per octave. A total of four different parameters were set, and the experimental results are presented in Table 5. Bins_per_octave was fixed at 12 in the first three experiments. As n_steps is set to 4 and 8, the WA is boosted by 2.42% and 0.56% respectively. When the pitch was adjusted downward, i.e., when n_steps was set to 6, the WA reached 63.60%, which is 3.17% higher than the baseline. A WA increase of 2.78% was observed when n_steps was set to 3 and the bins_per_octave

was adjusted to 24 for the last experiment. Pitch shifting only changes the pitch without affecting the speed of speech, which can improve the generalizability of the data and the model to a certain extent. After changing the pitch, the waveform frequency increases, and the amplitude decreases. The corresponding spectrogram is more separable in frequency, i.e., the harmonics of speech can be separated more clearly, thus improving classification accuracy.

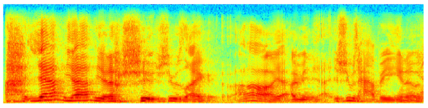
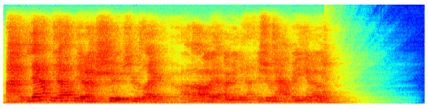
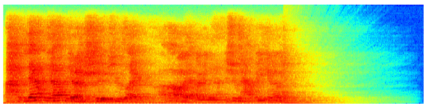
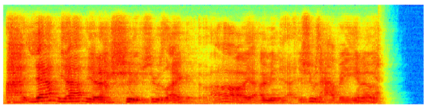
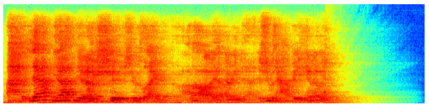
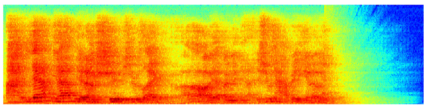
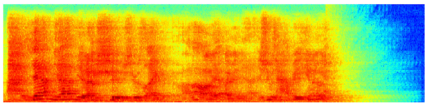
Table 5. Experimental results of pitch shifting.

Parameter 1 (bins_per_octave)	Parameter 2 (n_steps)	WA (%)	Spectrogram
-	-	60.43	
12	4	62.85	
12	8	60.99	
12	−6	63.60	
24	3	63.21	

3.2.6. Reverberation Augmentation

Reverberation is an effect that simulates the impulse response of a room to speech. We used the Pyroomacoustics toolkit in Python to add a reverberation effect to the original speech. It was decided that the spatial dimensions would be fixed at (10, 8, 3.5), which is similar to the dimensions of a real room. The reverberation time, the sound source location, and the microphone location were changed to generate six different reverberation effects. Table 6 presents the experimental results. During the first three experiments, we fixed the sound source and microphone position and only changed the reverberation time. Results indicated that higher or lower reverberation times could improve performance. The WA reached 63.05% when the reverberation time was set to 0.5 s, which is a 2.62% improvement over the baseline. Clearly, setting the reverberation time to 0.5 s is more appropriate. Following this, we alter the position of the sound source and microphone in space and fix the reverberation time. A WA of 61.26% was obtained when the sound source was placed at the edge of space and the microphone position was fixed. We then placed the microphone at the edge of space with the source at the previous position (3, 5, 1.75), and the WA reached 63.85%, an improvement of 3.42% compared to the baseline. We have achieved not only the highest level in reverberation augmentation, but the highest level in six of our proposed data augmentation methods. In the last experiment, we positioned both the sound source and microphone at the edge of space at a much farther distance from each other. The performance improvement was only 1.06% over the baseline, which was not satisfactory. As can be seen from spectrograms, after adding the reverberation effect to speech, there is a certain spread of frequencies. The more reverberation time there is, the more obvious the spread becomes. Moreover, the reverberation effect blurs the texture boundary of the spectrogram, making the correlation between adjacent frames stronger. This indicates that the superposition between signals after adding the reverberation effect makes emotional information in speech more apparent to some extent.

Table 6. Experimental results of reverberation augmentation. Time represents the reverberation time of speech, and Source and Microphone represent the positions of the sound source and microphone in the simulated space, respectively.

Time	Source	Microphone	WA (%)	Spectrogram
—	—	—	60.43	
0.5	(3, 5, 1.75)	(7.5, 5.8, 1.2)	63.05	
0.8	(3, 5, 1.75)	(7.5, 5.8, 1.2)	62.25	
0.2	(3, 5, 1.75)	(7.5, 5.8, 1.2)	61.19	
0.5	(1, 1, 1.75)	(7.5, 5.8, 1.2)	61.26	
0.5	(3, 5, 1.75)	(9, 7, 1.2)	63.85	
0.5	(1, 1, 1.75)	(9, 7, 1.2)	61.49	

3.3. Experiment of the Strong Generalized Speech Emotion Recognition Model Based on Effective Data Augmentation

In this section, we apply the two best-performing data augmentation strategies from the experiments in the previous section to the strong generalized speech emotion recognition model based on effective data augmentation proposed in this paper.

We used pitch shifting and reverberation augmentation, and the most successful setting from the previous experiments was used. During training, the ratio of augmented data obtained by each data augmentation strategy and original data is 1:1. The feature extractor in the model contains three sub-networks, and the Log–Mel spectrogram of original speech, pitch shifting speech, and reverberation speech is input into each of the three sub-networks. The weighting coefficients of the sub-networks are set to 0.6, 0.2, and 0.2, respectively, and β is 0.6. The Adam optimizer is chosen with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-6}, and the learning rate is dynamically reduced during the training process according to model performance. The batch size is 32, and the number of training epochs in each fold is 50.

3.3.1. Ablation Experiments

In Table 7 we report the performance of the strong generalized speech emotion recognition model based on effective data augmentation proposed in this paper. The WA is 66.51% under the speaker-independent experimental strategy. In order to verify the effectiveness of each module of the proposed model, three ablation experimental strategies are designed. (1) Without augmentation: no data augmentation is performed, and only the Log–Mel spectrogram features of the original speech are input. (2) Without multi-channel: the feature extractor has only one sub-network. (3) Without discriminator: remove the discriminator

and do not generalize emotional representations. Table 7 presents the results of the ablation experiments. In the absence of data augmentation, WA decreases by 1.43%. With only one sub-network in the feature extractor, WA is also lower than that of the proposed model. This illustrates that the multi-channel feature extractor presented in this paper can effectively take advantage of multiple data augmentations to improve SER performance. When no discriminator is applied, WA decreases by 4.54%. It is evident from this that robust emotional representations are essential for SER. The results of ablation experiments show that the proposed model can further improve performance by using data augmentation on the basis of aligning the distribution of emotional representations.

Table 7. Results of the ablation experiment.

Methods	WA (%)
Our proposed	66.51
Without augmentation	65.08
Without multi-channel	66.01
Without discriminator	61.97

3.3.2. Comparison with Mainstream SER Algorithms

Additionally, the proposed model was compared with mainstream SER algorithms. Table 8 shows the results of the comparison of the WA of the algorithms obtained using a speaker-independent experimental strategy on the IEMOCAP corpus. By comparison with traditional algorithms such as SVM, HMM, and ELM, the proposed model in this paper is superior by 9.76%, 7.05%, and 2.31%, respectively. Furthermore, we compare algorithms that use deep learning. The model proposed in this paper leads by 3.01% when compared to the RNN algorithm incorporating the attention mechanism. Finally, we compare an algorithm that utilizes adversarial training to learn speaker-invariant emotional representations, and the algorithm also employs speed augmentation. Our proposed model still outperforms this algorithm by 7.89%, which is a significant improvement. It further demonstrates the superiority of the proposed model in terms of effective data augmentation and generalization of emotional representations.

Table 8. Performance of the proposed model and mainstream SER algorithms on the IEMOCAP corpus.

Methods	WA (%)
Our proposed	66.51
Lexical-Norm + SVM [34]	56.75
MEnAN + Speed augmentation [6]	58.62
MFCC + HMM [35]	59.46
RNN + Attention [36]	63.50
Region-switching + ELM [37]	64.20

3.3.3. T-SNE Visualization of Emotional Representations

We compared the learned emotional representations from the model proposed in this paper and the baseline model in Section 3.2. Both emotional representations were visualized as T-SNE [38] in Figure 4. The emotional representations from the baseline model have a low degree of inter-class separation; particularly, "happy" and "angry" are entangled in hidden space. Furthermore, there is a domain shift between the representations of the training set and test set speakers. In contrast, the emotional representations from the proposed model form four distinguishable clusters, i.e., there is a high degree of separation between classes. Even the representations of speakers from the test set show excellent inter-class discrimination. Additionally, the representation distributions are effectively aligned, allowing for the mixing of representations from different speakers. Accordingly, the proposed model is capable of generalizing emotional representations while maintaining the validity of the category information in the representations.

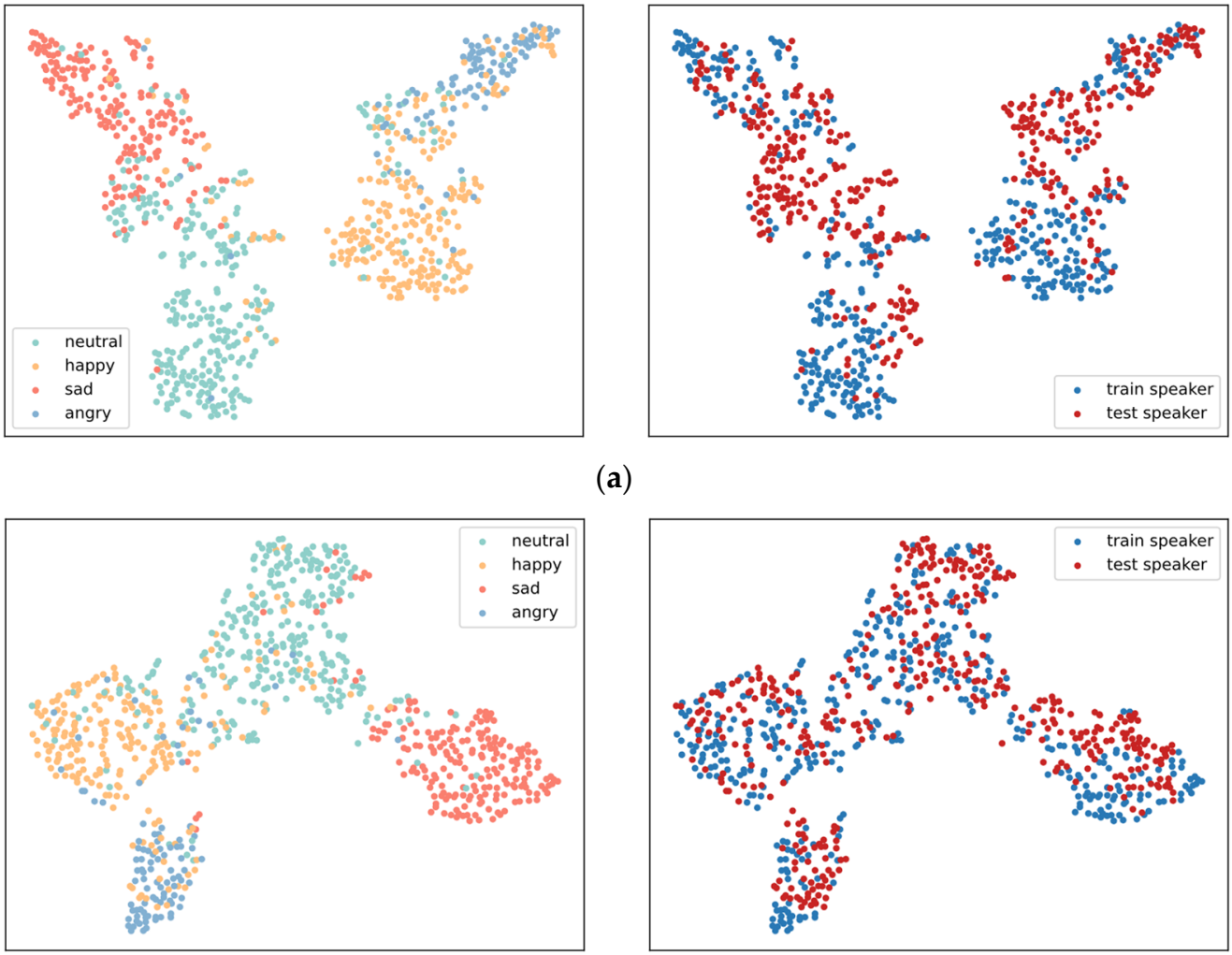


Figure 4. T-SNE visualization of emotional representations. (**a**) representations from the baseline model; (**b**) emotional representations from the proposed model.

4. Conclusions

In this paper, we investigated the problem of data augmentation in SER and proposed a strong generalized speech emotion recognition model based on effective data augmentation. First, we evaluated the effectiveness of the six proposed data augmentation methods: speed augmentation, noise injection, time shifting, resampling, pitch shifting, and reverberation augmentation. The experimental results of data augmentation show that some attributes of speech can be detrimental to emotion recognition when they are changed. Injection of noise obscures the emotional details of speech to a certain extent, resampling causes the loss of information in speech, and time shifting disrupts the overall coherence of speech, all of which are detrimental to speech recognition. The experimental results indicate that pitch shifting and reverberation augmentation are the two most effective methods for improving SER performance. In the pitch shifting experiment, when bins_per_octave and n_steps were set to 12 and -6, respectively, WA was improved by 3.17% compared to the baseline results. When the pitch of speech is changed, the spectrogram is more separable in frequency, i.e., the harmonics of speech can be separated more clearly, thus improving the classification accuracy. In the reverberation augmentation experiment, WA was improved by 3.42% when the reverberation time, source location and microphone location were set to 0.5, (3, 5, 1.75) and (9, 7, 1.2), respectively. The superposition between the signals

after adding the reverberation effect somehow makes the emotional information in speech more obvious.

Then, these two data augmentation strategies were applied to the model proposed in this paper. We conducted ablation experiments on the proposed model, and the results show that performance degradation is the greatest when the discriminator is not used. This indicates that individual differences among speakers are responsible for the performance degradation of SER and that the development of robust emotional representations is important for SER. The results of mainstream SER algorithms on the IEMOCAP corpus were compared with the proposed model. The WA of the proposed model is 2–9% higher than those of the relevant algorithms. According to the T-SNE visualization results, the representations from the proposed model exhibit better inter-class separability as well as generalization, which further proves the superiority of this work.

In future work, we may study the problem of representation generalization on cross-corpus SER. Performance tends to drop significantly in the case of cross-corpus, which is a challenging task. Moreover, we may consider modalities such as video and text to further improve the performance of emotion recognition.

Author Contributions: Conceptualization and methodology, H.T. and S.S.; software, S.S.; validation, S.S.; formal analysis, H.T., C.Z. and H.G.; investigation, Z.H.; data curation, S.S.; writing—original draft preparation, H.T. and S.S.; writing—review and editing, H.T. and S.S.; visualization, S.S.; funding acquisition H.T. All authors have read and agreed to the published version of the manuscript.

Funding: The research was supported in part by the Henan Province Key Scientific Research Projects Plan of Colleges and Universities (Grant No. 22A520004, Grant No. 22A510001 and Grant No. 22A510013) and in part by the National Natural Science Foundation of China (Grant No. 61975053).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: [IEMOCAP] [https://sail.usc.edu/iemocap/] (accessed on 29 December 2022).

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

References

1. Rana, R.; Latif, S.; Gururajan, R.; Gray, A.; Mackenzie, G.; Humphris, G.; Dunn, J. Automated screening for distress: A perspective for the future. *Eur. J. Cancer Care* **2019**, *28*, 13. [CrossRef]
2. Zhou, Y.; Liang, X.F.; Gu, Y.; Yin, Y.F.; Yao, L.S. Multi-Classifier Interactive Learning for Ambiguous Speech Emotion Recognition. *IEEE-ACM Trans. Audio Speech Lang.* **2022**, *30*, 695–705. [CrossRef]
3. Yadegaridehkordi, E.; Noor, N.; Bin Ayub, M.N.; Affal, H.B.; Hussin, N.B. Affective computing in education: A systematic review and future research. *Comput. Educ.* **2019**, *142*, 19. [CrossRef]
4. Aldeneh, Z.; Provost, E.M. Using regional saliency for speech emotion recognition. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2741–2745.
5. Busso, C.; Parthasarathy, S.; Burmania, A.; AbdelWahab, M.; Sadoughi, N.; Provost, E.M. MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception. *IEEE Trans. Affect. Comput.* **2017**, *8*, 67–80. [CrossRef]
6. Li, H.; Tu, M.; Huang, J.; Narayanan, S.; Georgiou, P. Speaker-Invariant Affective Representation Learning via Adversarial Training. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7144–7148.
7. Braunschweiler, N.; Doddipatla, R.; Keizer, S.; Stoyanchev, S. A Study on Cross-Corpus Speech Emotion Recognition and Data Augmentation. In Proceedings of the 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 13–17 December 2021; pp. 24–30.
8. Mujaddidurrahman, A.; Ernawan, F.; Wibowo, A.; Sarwoko, E.A.; Sugiharto, A.; Wahyudi, M.D.R. Speech Emotion Recognition Using 2D-CNN with Data Augmentation. In Proceedings of the 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), Pekan, Malaysia, 24–26 August 2021; pp. 685–689.
9. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the Interspeech, Lisbon, Portugal, 4–8 September 2005; pp. 1517–1520.
10. Liu, J.; Wang, H. A Speech Emotion Recognition Framework for Better Discrimination of Confusions. In Proceedings of the Interspeech, Toronto, ON, Canada, 6–11 June 2021; pp. 4483–4487.

11. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.-C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *arXiv* **2019**, arXiv:1904.08779.
12. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *arXiv* **2014**, arXiv:1406.2661. [CrossRef]
13. Chatziagapi, A.; Paraskevopoulos, G.; Sgouropoulos, D.; Pantazopoulos, G.; Nikandrou, M.; Giannakopoulos, T.; Katsamanis, A.; Potamianos, A.; Narayanan, S. Data Augmentation Using GANs for Speech Emotion Recognition. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 171–175.
14. Mariani, G.; Scheidegger, F.; Istrate, R.; Bekas, C.; Malossi, C. BAGAN: Data Augmentation with Balancing GAN. *arXiv* **2018**, arXiv:1803.09655.
15. Yi, L.; Mak, M.W. Adversarial Data Augmentation Network for Speech Emotion Recognition. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 529–534.
16. Li, W.; Fan, L.; Wang, Z.Y.; Ma, C.; Cui, X.H. Tackling mode collapse in multi-generator GANs with orthogonal vectors. *Pattern Recognit.* **2021**, *110*, 107646. [CrossRef]
17. Meng, H.; Yan, T.H.; Yuan, F.; Wei, H.W. Speech Emotion Recognition from 3D Log-Mel Spectrograms with Deep Learning Network. *IEEE Access* **2019**, *7*, 125868–125881. [CrossRef]
18. Fan, W.Q.; Xu, X.M.; Cai, B.L.; Xing, X.F. ISNet: Individual Standardization Network for Speech Emotion Recognition. *IEEE-ACM Trans. Audio Speech Lang.* **2022**, *30*, 1803–1814. [CrossRef]
19. Xu, Y.; Kong, Q.; Wang, W.; Plumbley, M.D. Large-Scale Weakly Supervised Audio Classification Using Gated Convolutional Neural Network. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 121–125.
20. Gui, J.; Li, Y.; Chen, K.; Siebert, J.; Chen, Q. End-to-End ASR-Enhanced Neural Network for Alzheimer's Disease Diagnosis. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 8562–8566.
21. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein GAN. *arXiv* **2017**, arXiv:1701.07875.
22. Lounnas, K.; Lichouri, M.; Abbas, M. Analysis of the Effect of Audio Data Augmentation Techniques on Phone Digit Recognition for Algerian Arabic Dialect. In Proceedings of the 2022 International Conference on Advanced Aspects of Software Engineering (ICAASE), Constantine, Algeria, 17–18 September 2022; pp. 1–5.
23. Hailu, N.; Siegert, I.; Nürnberger, A. Improving Automatic Speech Recognition Utilizing Audio-codecs for Data Augmentation. In Proceedings of the 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland, 21–24 September 2020; pp. 1–5.
24. Zhao, W.; Yin, B. Environmental sound classification based on pitch shifting. In Proceedings of the 2022 International Seminar on Computer Science and Engineering Technology (SCSET), Indianapolis, IN, USA, 8–9 January 2022; pp. 275–280.
25. Lin, W.; Mak, M.W. Robust Speaker Verification Using Population-Based Data Augmentation. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 7642–7646.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
27. Han, S.; Leng, F.; Jin, Z. Speech Emotion Recognition with a ResNet-CNN-Transformer Parallel Neural Network. In Proceedings of the 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), Beijing, China, 14–16 May 2021; pp. 803–807.
28. Hsu, J.H.; Su, M.H.; Wu, C.H.; Chen, Y.H. Speech Emotion Recognition Considering Nonverbal Vocalization in Affective Conversations. *IEEE-ACM Trans. Audio Speech Lang.* **2021**, *29*, 1675–1686. [CrossRef]
29. Jiang, X.; Guo, Y.; Xiong, X.; Tian, H. A Speech Emotion Recognition Method Based on Improved Residual Network. In Proceedings of the 2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST), Guangzhou, China, 10–12 December 2021; pp. 539–542.
30. Luo, D.; Zou, Y.; Huang, D. Speech emotion recognition via ensembling neural networks. In Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, 12–15 December 2017; pp. 1351–1355.
31. Xu, M.; Zhang, F.; Khan, S.U. Improve Accuracy of Speech Emotion Recognition with Attention Head Fusion. In Proceedings of the 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 6–8 January 2020; pp. 1058–1064.
32. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. Improved Training of Wasserstein GANs. *arXiv* **2017**, arXiv:1704.00028.
33. Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
34. Mariooryad, S.; Busso, C. Compensating for speaker or lexical variabilities in speech for emotion recognition. *Speech Commun.* **2014**, *57*, 1–12. [CrossRef]

35. Chenchah, F.; Lachiri, Z. Impact of emotion type on emotion recognition through vocal channel. In Proceedings of the 2019 International Conference on Signal, Control and Communication (SCC), Hammamet, Tunisia, 16–18 December 2019; pp. 274–277.
36. Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic speech emotion recognition using recurrent neural networks with local attention. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2227–2231.
37. Deb, S.; Dandapat, S. Emotion Classification Using Segmentation of Vowel-Like and Non-Vowel-Like Regions. *IEEE Trans. Affect. Comput.* **2019**, *10*, 360–373. [CrossRef]
38. Van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.