Article

Unsupervised Feature Learning for Speech Emotion Recognition Based on Autoencoder

Yangwei Ying , Yuanwu Tu and Hong Zhou *

Key Laboratory for Biomedical Engineering of Ministry of Education, Zhejiang University, Hangzhou 310027, China; ywyang@zju.edu.cn (Y.Y.); 3170103731@zju.edu.cn (Y.T.)

* Correspondence: zhouh@mail.bme.zju.edu.cn

Abstract: Speech signals contain abundant information on personal emotions, which plays an important part in the representation of human potential characteristics and expressions. However, the deficiency of emotion speech data affects the development of speech emotion recognition (SER), which also limits the promotion of recognition accuracy. Currently, the most effective approach is to make use of unsupervised feature learning techniques to extract speech features from available speech data and generate emotion classifiers with these features. In this paper, we proposed to implement autoencoders such as a denoising autoencoder (DAE) and an adversarial autoencoder (AAE) to extract the features from LibriSpeech for model pre-training, and then conducted experiments on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) datasets for classification. Considering the imbalance of data distribution in IEMOCAP, we developed a novel data augmentation approach to optimize the overlap shift between consecutive segments and redesigned the data division. The best classification accuracy reached 78.67% (weighted accuracy, WA) and 76.89% (unweighted accuracy, UA) with AAE. Compared with state-of-the-art results to our knowledge (76.18% of WA and 76.36% of UA with the supervised learning method), we achieved a slight advantage. This suggests that using unsupervised learning benefits the development of SER and provides a new approach to eliminate the problem of data scarcity.

Keywords: unsupervised learning; deep learning; speech emotion recognition; deep neural network; autoencoders

Citation: Ying, Y.; Tu, Y.; Zhou, H. Unsupervised Feature Learning for Speech Emotion Recognition Based on Autoencoder. *Electronics* **2021**, *10*, 2086. <https://doi.org/10.3390/electronics10172086>

Academic Editors: George A. Papakostas and Gwanggil Jeon

Received: 1 July 2021
Accepted: 26 August 2021
Published: 28 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Language is the most fundamental mode of human emotional expression and speech is the main way humans communicate. Speech carries a speaker's semantic information and contains the emotional information the speaker wants to convey [1]. The same text may express opposite meanings in different emotional contexts. Emotional interaction is an indispensable part of human social activities, while emotional intelligence plays an important part in human computer interaction (HCI) [2]. The emotional states affect human interaction, such as facial expression [3,4], body posture [5], communication content [6], and speech mannerisms [7]. Influenced by multimodal features, recognizing emotions is vital to develop an automatic speech emotion recognition system for understanding of these interactions, and various speech recognition for pronunciation training and learning are developed [8,9]. Currently, most emotional recognition tasks are accomplished with hand-crafted features, which requires the guarantee of data validity and quantity. To avoid the lack of data and make feature extraction objective, developing an unsupervised learning method in emotion recognition is worthwhile. In this paper, we attempt to design a robust and generic mechanism to recognize emotions accurately with acoustic features extracted from widely used public speech data by unsupervised learning approaches of autoencoder model.

In the past decades, research has mainly focused on using traditional methods, e.g., Hidden Markov Model (HMM) and Gaussian Mixed Model (GMM) [10,11] to recognize

emotions. However, GMM cannot make use of the frame context information nor learn the deep nonlinear feature transformation. With the rapid development of deep learning techniques [12], various neural networks have been applied in speech emotion recognition (SER) [13]. For SER, the most important factor is the extraction of speech features. In supervised learning methods, scholars always considered acoustic features like Mel-Frequency Cepstral Coefficients (MFCCs), and paralinguistic features including pitch, harmonics, speech energy, pause and central moments [14]. Besides, SER mainly relies on the labeled speech data and requires quantitative training data for model building. In addition, the manual selection of acoustic features can be too randomized, which can limit the generalization of deep learning architectures when there is a lack of enough data sources and hinders further research.

SER is an emerging field where quantities of researchers developed various techniques and approaches to strengthen this area [15]. Most research pays attention to making use of certain common speech signals for emotional identification. In the recent years, with the development of deep learning, many approaches are developed and employed for feature detection and emotional classification. However, challenges still exist, especially when lacking enough speech data, it is difficult to make a fair judgement with limited datasets. Moreover, the usage of hand-crafted features may lead to loss of important information. Furthermore, when applying to new dataset, the model has to be re-designed, which is of poor generalization. Some researchers have noticed and focused on identifying features through various approaches especially with unsupervised learning techniques [16,17]. Unsupervised feature learning aims to use a large number of unlabeled data to train and extract features. A common method is the autoencoder method [18,19], which usually reduces the dimensions of the input information and reconstructs the information according to the reduced features, which tries to restore the input information. The process of acquiring dimension reduction features is called encoder, and the process of restoring information is called decoder. A convolutional neural network is widely used in the field of image recognition and classification to build an autoencoder model.

Therefore, in this paper, to eliminate the influence of manually selected features and expand the generalization of a pre-trained model, we attempt to transform the speech information into the corresponding spectrogram, use a convolutional neural network to extract features widely and effectively, then transfer the pre-trained autoencoder model to target datasets for further studies. The main contributions of this paper can be summarized as follows:

- A new speech data pre-processing approach is proposed with different time shifts to balance data distribution and realize data augmentation.
- Three convolutional neural network-based autoencoders, namely simple autoencoder, denoising autoencoder (DAE), and adversarial autoencoder (AAE), are proposed to extract speech features from LibriSpeech.
- The pre-trained autoencoder models are transferred to IEMOCAP for feature extraction and emotion classification, which leads to a comparable weighted accuracy (WA) and unweighted accuracy (UA) with state-of-the-art results.

The organization of this paper is as follows. Section 2 describes the related works in speech emotion recognition. Section 3 describes the data sets we used in this paper. In Section 4, the methodologies and proposed model architecture are performed. Section 5 details the experimental process and results. Section 6 is the conclusion and expected future work.

2. Related Works

Speech features are most important for recognizing speech emotions. Conventional spectrogram features usually pay attention to MFCC, linear frequency cepstral coefficient (LFCC), etc. Generally speaking, the overall SER procedure is as follows. The speech signal is cut into frame-level after pre-processing, and the frame feature is extracted respectively. In the meantime, the frame level features are integrated to obtain statistically significant

utterance-level features, and then they are used to complete the task of emotion recognition and classification. With numerous neural networks applied in signal processing, more researchers put forth effort toward studying emotion recognition. For instance, Han et al. [20] selected frame-level features as inputs of deep neural network (DNN) to obtain utterance-level features, then utilized an extreme learning machine (ELM) for classification and performed 48.2% unweighted accuracy on IEMOCAP dataset. Chen et al. [21] proposed to use three dimensions—MFCC, deltas, and delta-deltas of MFCC—as input, and adopted RCNN to extract emotional features from 3-D dimensional input. The attention mechanism was applied to get more robust utterance-level features, achieving 64.7% unweighted average recall on IEMOCAP and 82.8% on Emo-DB. Zhang et al. [22] combined a fully convolutional network (FCN) and attention mechanism to realize 63.79% unweighted accuracy on IEMOCAP with raw spectrogram as the input. Mustaqeem et al. [23] proposed a simple and lightweight deep learning-based self-attention module (SAM) for SER system.

The academic achievements mentioned above mainly focused on using well-established hand-craft speech features. However, sometimes these features cannot represent the best and robust speech characteristics. As a result, some researchers began to attempt using deep learning methods to find the most appropriate features, especially unsupervised learning methods like autoencoder. A pre-trained model is built by an autoencoder to extract emotional features and the input data are reduced to low-dimensional feature vectors which represent the input data. The autoencoder can be divided into two parts: encoder and decoder, and the network parameters are optimized by minimizing the loss of input and output reconstruction, which makes the vector of bottleneck layer representative.

Xia et al. [24] applied DAE to map standardized static features into two hidden layer representations, one for reconstruction information, and the other for emotional features. Finally, an SVM model was used for emotion classification, and achieved 61.42% unweighted accuracy on IEMOCAP. Mustaqeem et al. [25] proposed a unique artificial intelligence (AI)-based system structure for the SER to address the limitations of the existing SER systems, and obtained a 75% and an 80% recognition rate in IEMOCAP and RAVDESS. Ghosh et al. [26] used stacked DAE and BLSTM autoencoder to extract features, and obtained 49.09% unweighted accuracy on IEMOCAP. Eskimez et al. [27] made a comprehensive analysis of the application effect of four autoencoders in speech recognition, and concluded that VAE, AAE, and AVB with inferential properties have certain advantages in speech emotion recognition task with limited identification data. Neumann et al. [28] trained a recurrent sequence-to-sequence AE on a large number of unlabeled data, and used convolutional network with attention mechanism to synthesize speech features. Finally, an average unweighted recognition rate of 59.54% was obtained on IEMOCAP.

Although these works have proved the efficiency and validity of using unsupervised learning method to extract speech data, the recognition accuracy is not satisfactory and still needs to be improved. In our work, we proposed to rearrange and optimize data division and augmentation, and compare multiple different autoencoders to extract speech features. Finally, network models, e.g., a CNN model and CNN with LSTM, are compared to achieve better recognition results.

3. Data

3.1. LibriSpeech

The LibriSpeech corpus [29] is derived from audiobooks that are part of the LibriVox project. It is a large-scale corpus of read English speech, which contains 1000 h data for analysis. Moreover, reading different books can reflect certain emotion and it is more objective. Based on the quantity of data size and data quality, it is chosen for pre-training. Here, it is used as training data for autoencoders to extract features and output the pre-trained model. There are about 1000 h of speech data from around 2500 speakers. Its purpose is to enable the training and testing of automatic speech recognition(ASR) systems.

LipriSpeech is split into several parts for users to selectively download according to their needs, for example, training data contain 100 and 360 h of “clean” speech, and the sample rate is 16 kHz. The subsets with “clean” in their name are supposedly “cleaner” (at least on average), than the rest of the audio and US English accented subsets. The subset labelled “other” should be more challenging as that is possibly misclassified.

For data accuracy and purity, we use clean subsets for generating the pre-trained model. Specifically, the pre-trained encoder network is used to extract acoustic features from the log-Mel spectrogram of the input speech.

3.2. IEMOCAP

The Interaction Emotional Dyadic Motion Capture (IEMOCAP) [30] dataset released from the University of Southern California (USC) contains about 12 h of talking speech data from five sessions. Each session consists of scripted and improvised interactions between male and female actors. In total, there are ten categorical emotions with 10,038 utterances, including angry (1103 samples), happy (595), sad (1084), neutral (1708), frustrated (1849), excited (1041), fearful (40), surprised (107), disgusted (2), and other (3) by at least three different annotators.

For the improvised part, actors emote more real emotions than that in the scripted transcriptions. Therefore, to simulate and match natural speech as much as possible, we only consider using the improvised speech data in our research. While most research work on IEMOCAP only considered anger, sad, happy, and neutral, these four emotions are the most common in real life, and thus we chose these four representative emotions in our work.

4. Methodology

The overall network framework is divided into two parts to complete the speech emotion recognition task. One part is a pre-trained autoencoder based on LibriSpeech data, and the other part is a fully connected classification network based on IEMOCAP data. The overall architecture of the proposed research mechanism is shown in Figure 1.

Figure 1. Architecture of the proposed research mechanism. The pre-train model derived from LibriSpeech and the features extracted from pre-train model are utilized for classification, which conducted the certain emotion of “Happy”, “Angry”, “Sad”, and “Neutral”.

4.1. Data Pre-Processing

The speech data of LibriSpeech and IEMOCAP are dealt with the same pre-processing procedure. First, the minimum maximum normalization is carried out for each utterance according to Equation (1) to eliminate value differences, where x' , x , x_{min} , x_{max} are normalized data, original data, the minimal data, and the maximal data of acoustic features' values in each speech segment, respectively. Second, the speech data are truncated into

500 ms segments, and the speech segment less than 500 ms are padded with 0 to generate the Mel-spectrogram of each segment.

Librosa [31], which is commonly used in the speech recognition domain, is used to extract Mel-spectrogram with a sampling rate of 16 kHz. We use a Hanning window with a length of 512 and a hop length of 64 to perform a short time Fourier transform (STFT), and the spectrogram is mapped into 128 Mel-scale. Afterwards, the Mel-spectrogram is taken as a logarithm to obtain the log Mel-spectrogram. Finally, in order to keep the features and speed up the training, z-score standardization (zero mean normalization) is adopted in each discrete time. The normalized data have the characteristics of zero mean and standard deviation of one. Equation (2) describes the normalization process, where x' , x , μ , σ are normalized data, original data, the mean value, and the variance respectively. Figure 2 shows the data-processing procedure.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

$$x' = \frac{x - \mu}{\sigma} \quad (2)$$

Figure 2. Speech data pre-processing procedure.

Consider that there is a significant difference among different emotional categories in quantity. In order to balance the emotion distribution of the divided data set, part of the linguistic data would be discarded in conventional processing. This further reduces the amount of training data, which is an important factor that affects the model recognition results. Therefore, we attempt to approximately divide four kinds of emotional speech data of IEMOCAP of improvised happy, angry, neutral, and sad emotional speech into ten parts with newly designed approach, which also augments the data size and makes full use of the original data.

Taking “happy” as an example, we traverse all five sessions of improvised speech data. For each “happy” utterance we find, it is put into part 1 to part 10 in turn, so that the data amount of “happy” emotion in each part is close to other 9 parts. Then, it turns to “angry”, afterwards “neutral”, and finally the “sad” utterance. The remaining three emotions are dealt with the same way. Besides, due to the difference of total amount and utterance duration of each emotion, in order to make the number of segments similar in each part after pre-processing, each emotion uses a different time shift to obtain the 500 ms segment. Then the time duration of each part is counted, and the “neutral” is the longest, so the “neutral” emotion is segmented without overlapping, while the rest emotion segments are partially overlapped. According to Equation (3), set the “neutral” emotion time shift as 500 ms, and get the number of segments after segmentation, then the time shifts of other three emotion are inferred. Then, the pre-processing of each part is completed through the process shown in Figure 2.

$$N = \frac{T_{all} - (T_f - T_s)}{T_s} \quad (3)$$

where N represents the segment number and make it close among the four emotion utterances, T_{all} is the total duration of certain emotions, T_f is the segment duration after segment (500 ms in this paper), and T_s is the time shift.

4.2. Pre-Train Model

We use four-layer CNN with leaky RELU activation function and a fully connected layer to generate bottleneck. Then a fully connected layer with four-transposed convolutional layers was used to map the bottleneck to reconstructed spectrogram. A 3×3 kernel is adopted in the convolutional layers.

A multi-layer convolutional network is used to extract the feature based on the spectrogram, and the maximum pooling is used to compress the data and reduce the dimension [32]. Finally, the hidden layer feature is obtained through the fully connected network to complete the feature-extraction task, which is called encoding. According to the characteristics of the hidden layer, the network structure symmetrical with the encoder is used for deconvolution and unpooling, and the data are recovered step-by-step. Finally, the reconstructed spectrogram is the output.

Figure 3 shows the autoencoder network framework. Denoising autoencoder (DAE) and adversarial autoencoder (AAE) are built based on it.

Figure 3. General framework of the autoencoder network, which consists of encoder and decoder networks.

The basic structure of the DAE is consistent with the block diagram in Figure 3, except that the input is replaced with the logarithmic spectrogram with Gaussian noise. Moreover, the reconstructed spectrogram output is required to be as close as possible to the original spectrogram without noise. Here, x is used to represent the input logarithmic spectrogram, $\tilde{x}$ is the spectrogram with noise, z is the output features of the bottleneck layer, and $\hat{x}$ is the reconstructed spectrogram. Equation (4) adopts the mean squared error (MSE) as a loss function to evaluate the quality of the reconstructed spectrogram, where L , x , $\hat{x}$, and N are loss value, input log-Mel spectrogram, reconstructed spectrogram, and number of spectrograms, respectively.

$$L = \frac{1}{N} \sum_{i=1}^N (x - \hat{x})^2 \quad (4)$$

AAE employs generating adversarial training to constrain z , and takes the encoder part as a generator. Then a discriminator is added to judge whether the input z matches the prior distribution.

Figure 4 shows the discriminator network framework. The output score of the discriminator is limited between zero to one by the sigmoid function. In the process of training, the network parameters of encoder and decoder are updated through Equation (4) for each batch. The discriminator updates the parameters every n batches, which makes the discrimination stricter. Each batch of the generator updates its parameters, so that the z generated by itself can reach a higher score in the discriminator. During the training process, Equation (5) is used to update the discriminator, and Equation (6) is used to update the generator parameters, where $D(z)$ is the discriminator of AAE to judge if input z matches the priori distribution. Finally, the fully connected network is used for emotion classification. The cross entropy loss function is used to update the model parameters.

$$L = -\frac{1}{N} \sum_{i=1}^N \log(D(z_{gauss})) + \log(1 - D(z)) \quad (5)$$

$$L = -\frac{1}{N} \sum_{i=1}^N \log(D(z)) \quad (6)$$

Figure 4. Framework of the discriminator network, where z is the input and the discriminator is used to judge if z matches the priori distribution.

Figure 5 shows the structure and parameters of the encoder network, decoder network, and discriminator model. In encoder model, a four-layer convolution network and maximum pooling are used to extract features and compress data. Batch normalization is used in each layer to improve the model performance and training speed. Each layer of the CNN consists of a 3×3 kernel to map the fine structures of the time-frequency representation into different feature maps. In decoder network, the interpolation method is used for unpooling, and a deconvolution network is used to reconstruct data.

Figure 5. Scheme of the encoder, decoder and discriminator models. FC: fully connected layer. (a) The architecture of encoder and decoder model; (b) the discriminator in pre-train model; (c) the discriminator in classification model.

Three autoencoders are compared in this section. The first is a simple encoding-decoding network, which adopts the network structure shown in Figure 5a, using the log

spectrogram after LibriSpeech pre-processing as the input, and Equation (4) is used as the loss function. The second one is the DAE network, which has the same structure as the first one, but uses the pollution spectrogram with Gaussian noise as the input. The last one is the AAE, which adds the discriminator as shown in Figure 5b. In order to make adversarial training more effective, the updating frequency of the discriminator parameters is lower than that of the generator. During training, the parameters of the discriminator are updated with Equation (4) every six batches. When the minimum reconstruction error of 10 epochs in validation sets remains the same, the training process is considered to be finished and feature extraction is completed.

4.3. Classification Model

The IEMOCAP data set is employed for classification network training, and the specific model and parameters are shown in Figure 5c.

At present, the commonly used emotion recognition classifiers include support vector machine (SVM), hidden Markov model (HMM), extreme learning machine (ELM), deep learning network classifier, etc. With the continuous development of deep learning framework and the improvement of training speed, using deep learning network for emotion classification is a faster and more convenient approach. In addition, it is able to extract and integrate all the previous useful information. Coupled with the nonlinear mapping of the activation function, the multilayer fully connected layer can simulate any nonlinear transformation in theory. Thus, the multi-layer full connection network is employed for the classification task in this paper. It is worth noting that the low dimensional features are processed to increase dimension first in the network, which is used to prevent the network from directly transferring the input data to the next layer. The dropout layer with high discard rate is used here, where the drop rate is set as 0.7 between the fully connected layers to avoid over-fitting. Then, batch normalization is used to improve the speed of the model. Finally, softmax function is used to output the normalized emotion recognition results. The validation process for the classifiers follows a 10-fold cross-validation strategy. We consider eight parts for training, one for validation and the remaining one for test.

As the speech data after pre-processing come in 500 ms segments, the source of 500 ms segments is recorded. These 500 ms segments are used for classification and recognition, which are divided into four types of emotions. The highest score of each emotion is recorded by synthesizing the 500 ms segments from the same source, and the output result of the original speech is composed, among which the highest one is the emotion tag of the utterance. The classification process can be seen in Figure 6.

The classifier was trained with an ADAM optimizer [33] with learning rate at 0.0002.

4.4. Evaluation Metrics

Unweighted accuracy (UA) and weighted accuracy (WA) are selected for evaluation. UA is the average accuracy of each emotion category, which is the sum of all class accuracies divided by the number of emotion classes. WA refers to the accuracy of all samples, which is the total number of correctly classified samples averaged by the total testing samples. These two types of accuracies are calculated according to Equations (7) and (8).

$$UA = \frac{\sum_1^n \frac{x_i}{X_i}}{k} \quad (7)$$

$$WA = \frac{\sum_1^n x_i}{\sum_1^n X_i} \quad (8)$$

where k represents the emotion classes, X_i is the number of utterances in emotion i , and x_i is the number of correctly classified samples in emotion i .

Figure 6. Overall architecture of decision process of emotion prediction, the highest score of all the utterance is decided as the final output emotion class.

5. Experiments and Results

5.1. Analysis of Data Pre-Processing

All the models mentioned in this paper are implemented with PyTorch [34].

The average duration of speech utterances was about 4.5 s. Table 1 shows the improved utterance distribution, which indicates that huge differences exist in different emotions. The “neutral” accounts for the majority while “happy” and “anger” are minority, which leads to severe data imbalance. This immediately follows the speech segment since the original data are too long, the speech data are truncated into 500 ms segment, the detailed emotional segments are shown in Table 2. As we can see that “Happy” and “Anger” only represent a small part. All the speech data processed with the method mentioned in Figure 1 will come to a Mel-spectrogram. Figure 7 shows a typical spectrogram of the speech segment with x-axis of time series and y-axis of frequency.

Table 1. Categorical emotion distribution of IEMOCAP with four emotions in training, validation, and test set.

	Happy	Sad	Anger	Neutral
Training set	176	473	258	805
Validation set	49	82	24	142
Test set	57	51	7	141

Table 2. Categorical emotion distribution of IEMOCAP after segments of four emotions.

Emotion Category	Number
Happy	1270
Sad	4052
Anger	1944
Neutral	5437

As different emotion speech data vary considerably, for data that are uniformly distributed and to avoid the deficiency of data imbalance, according to Equation (3), four classes of emotion segment time shifts were calculated. Since the original data have been separated into ten parts, here we present the detailed data information of Part 1 in Table 3. As we can see from the result, all the emotions should be comparable and this is a new attempt at data augmentation. With different time shift, four emotions are well organized and the output almost looks the same with around 770 segments. Thus, we can simply regard the new derived data to be balanced.

Figure 7. Spectrogram of one utterance sample. The x-axis represents the time series and y-axis represents the frequency.

Table 3. Detailed data distribution of IEMOCAP Part1 after uniformly processing.

	Happy	Sad	Anger	Neutral
Original utterance	29	61	29	110
Time shift	0.176 s	0.323 s	0.121 s	0.5 s
Segments	779	779	743	793

5.2. Pre-Training on LibriSpeech

This section demonstrates the results of data reconstruction with three different autoencoders, which illustrates the effectiveness of autoencoder training. All the three models encode the features into 512 dimensions, and then utilize the decoder model to reconstruct the original Mel-spectrogram. For visual check in this section, a randomly selected Mel-spectrogram in the training set is input into autoencoder networks to accomplish the pre-training procedure, and the corresponding reconstructed output is obtained, as shown in Figure 8, where Figure 8a–c represents the reconstruction result with simple AE, DAE, and AAE respectively. According to the reconstructed images, we find that all the autoencoders obtain certain satisfactory results, which can be well adopted in the further classification. For the comparability of results, the same loss function of MSE is adopted in the training process of three autoencoders, and the total loss of the current epoch is obtained by adding the loss of each iteration. Afterwards, the model with the lowest loss is saved in the validation dataset. Based on the above conditions, for simple autoencoder, the overall reconstruction loss is 16.54 and the average loss is 0.117; for DAE, the overall reconstruction loss is 18.83 and the average loss is 0.134; for AAE, the overall reconstruction loss is 19.57 and the average loss is 0.139. Therefore, the simple autoencoder can achieve a smaller loss on the training set than the others. However, by observing the comparison in Figure 8 and repeating five times randomly selected pictures for comparison, AAE is superior to the former two in the reconstruction of feature details.

Figure 8. Reconstruction results on LibriSpeech with three autoencoders. The x -axis represents the time table and y -axis represents the frequency. (a) Input log Mel-spectrogram, (b) Simple AE, (c) DAE, (d) AAE.

Immediately, we transferred three autoencoders to IEMOCAP data for reconstruction, and the results can be seen in Figure 9. The reconstruction Mel-spectrogram is similar to the original one to a certain extent, which proves that transferring autoencoder pre-train model to target datasets is available. However, compared with the original input data, the reconstructed data still loses a lot of characteristic information. This section attributes it to the difference in the content of the two databases. LibriSpeech is mainly the corpus of speaker reading the book, which is relatively stable with little emotional fluctuation. The IEMOCAP data are conversations in the natural context and have strong volatility.

5.3. Classification of IEMOCAP Emotions

The IEMOCAP dataset is used for classification based on three pre-trained autoencoder models, and classification accuracy is compared with different dimensions of 128, 256, and 512 of extracted features. The intact speech segments and emotion categories are used for classification and recognition. First, we preprocess the speech data into segments according to Section 4.1, and the source emotional category of 500 ms segments after segmentation are recorded. Then, these 500 ms segments are employed for subsequent training and testing, which are divided into four types of emotions. For each utterance, the highest score of each emotion is recorded by analyzing the 500 ms segments from the same source, which is regarded as the output result of the original speech. Among the segments, the emotion with highest score is regarded as the category.

To investigate the consequence of different feature dimensions, three dimensions are considered, 128, 256, and 512 respectively, which is also known as the hidden layer feature dimension. Table 4 shows the result of unweighted accuracy under different hidden layer feature dimensions, which are used to compare the classification accuracy. It can be seen from the result that when the dimension increased, UA followed by an increase. Hence, we chose 512 dimensions for further studies as it achieves the best UA of 50.73%. It also makes sense since the high dimension features contain more information, which is advantageous to build a robust model, though it may increase the computational complexity.

Figure 9. Reconstruction results on IEMOCAP with three autoencoders. The x-axis represents the time table and y-axis represents the frequency. (a) Input log Mel-spectrogram, (b) Simple AE, (c) DAE, (d) AAE.

Table 4. Unweighted accuracy of classification with different feature dimensions in AAE.

Dimensions	128	256	512
UA	45.41%	48.05%	50.73%

For a more comparable result, the results derived from three autoencoders and literature are analyzed. Table 5 demonstrates the classification results with unweighted accuracy as an evaluation metric, while the proposed model added weighted accuracy as reference. Compared with the result derived from Eskimez et al. [27], we concluded a slightly better result, which increased by around two percent. However, further improvement is still needed since current accuracy cannot be suitable for practical applications.

Table 5. Comparison of three autoencoder models in classification with unweighted accuracy (UA).

Model	Simple AE	DAE	AAE	-
Eskimez et al. [27]	44.82%	46.02%	48.18%	UA
Proposed model	45.12%	48.72%	50.73%	UA
	47.66%	47.66%	51.17%	WA

Further analysis revealed that all the above three methods exist misjudging “happy” as “neutral” and “neutral” as “sad”. In addition, the difference of “anger” in the three kinds of recognition is attributed to the fact that the number of “anger” emotions in the test set is too small, which is far from the number of other emotions, as shown in Table 1.

To explore whether optimizing data pre-processing could improve the classification accuracy, we pre-processed the IEMOCAP data into ten parts and performed data augmentation as mentioned in Section 4.1 for optimization, with eight parts for training, one for validation, and the last for testing. In training process, the training epoch was set as 50, and the model with highest unweighted accuracy is saved in the validation set. Finally, the model was tested on the reserved test set, to obtain the weighted accuracy and unweighted

accuracy as the model performance. As a result, the effectiveness of the pre-trained model was further proved.

Table 6 shows the results of the three pre-trained models on the new augmented and balanced data set, with comparisons to results in other references. Among them, the pre-trained model using AAE method with balanced dataset achieved the best performance, and the unweighted accuracy reached 76.89% while the weighted accuracy reached 78.67%. Even though we introduced LSTM into pre-training process, it is still not as good as the model without LSTM, especially on DAE and AAE. It may be caused by the reason that DAE and AAE influenced the sequence of speech data, which led to an unsatisfying result. Comparing with the results of Eskimez et al. [27], Dissanayake et al. [35], and Xia et al. [24], they only reached a UA with 44%~61.50%, which is strongly due to the uneven data, while the autoencoder models are quite similar with ours. Therefore, the newly proposed data pre-processing approach indeed resolves the problem.

Table 6. Comparison of three models in classification with unweighted accuracy.

Model	Simple AE	DAE	AAE	
Eskimez et al. [27]	44.82%	46.02%	48.18%	UA
Xu et al. [36]	76.18%			UA
	76.36%			WA
Dissanayake et al. [35]	46.79%	-	-	UA
Xia et al. [24]	-	61.50%	-	UA
AE + CNN	45.12%	48.72%	50.73%	UA
	47.66%	47.66%	51.17%	WA
AE + CNN + balanced data	71.92%	71.26%	76.89%	UA
	73.78%	75.11%	78.67%	WA
AE + CNN + LSTM + balanced data	74.38%	65.20%	60.73%	UA
	76.00%	71.11%	59.11%	WA

The confusion matrix of 76.89% unweighted accuracy result is shown in Figure 10. As we can see that the happy emotion got the lowest classification accuracy of 64.29%, far below that of anger 85.71%. It is still caused by the data distribution mentioned before that the total number of happy is smallest.

	Anger	Sad	Happy	Neutral
Anger	85.71%		7.14%	7.14%
Sad	1.67%	75.00%	8.33%	15.00%
Happy	7.14%	3.57%	64.29%	25.00%
Neutral	6.42%	1.83%	9.17%	82.57%

Figure 10. Confusion matrix of unweighted accuracy 76.89% of AAE.

6. Discussion and Conclusions

In this paper, we explored the practicability of applying the unsupervised learning method into speech features and implemented speech emotion recognition. We proposed to adapt multiple autoencoders for feature extraction and utilize a convolutional neural network for classification, which were combined to analyze the influence on speech emotion recognition result. The autoencoders were trained with a corpus of read English speech in order to have a robust model that could serve as a reference for a posterior extraction of features related with the presence of emotional speech. The experiments conducted on the IEMOCAP data set achieved state-of-the-art accuracy. Moreover, with the application of unsupervised learning approaches, the influence of hand-crafted features is eliminated and the result is more objective.

However, we discovered that for the IEMOCAP dataset, there were only ten actors producing thousands of utterance-level speeches, which would limit the generalization of the training model for speaker-independent classification. Though we achieved a two percent higher accuracy than the literature mentioned (around 50%), this is still far from practical application. Therefore, an augmented and balanced training set was designed, which greatly improved the training effect of the classifier. Furthermore, the results obtained in this study are comparable or better than those observed in the literature when the same data are considered. Using the pre-trained AAE model with CNN discriminator, the best unweighted accuracy and weighted accuracy reached 76.89% and 78.67%, respectively. The out-performed results proved that it is feasible to complete the pre-trained model on a large number of available databases and apply it to the target database to implement the emotion recognition tasks.

Though a satisfactory recognition accuracy is obtained with our proposed approaches, the difficulty and limitation in applying the approaches still exist. The pre-trained model derived from a certain corpus, here LibriSpeech, the features extracted will be influenced by the quality of pre-train model to a certain extent. As we only check the encoder and decoder performance, if the reconstructed output is similar with the input, then the model is employed. In the future, the quality of features should be taken into consideration and well investigated.

Moreover, autoencoder model in this paper is designed based on convolutional neural network, which does not make good use of the time information contained in Mel-spectrogram. The advantages of attention mechanism and recurrent neural network in distance-dependent feature extraction should be considered in the future study. Finally, adversarial training can be used to constrain the potential representation and make it more interpretable. Such work has been applied in the research of generating countermeasure network, but it is still less used in the field of speech emotion recognition.

Aiming at the problem of misjudgment in references, we designed and implemented the data set augmentation and equalization for IEMOCAP, and achieved satisfactory results in the end. Due to the limitation of the experimental platform, only LibriSpeech was considered for pre-training. In the future work, we could use a larger amount of data to improve the autoencoder reconstruction effect to obtain better feature-extraction effects. Moreover, work is needed to improve the model and algorithm using a variety of feature data, controlling the distribution of hidden layers, and constraining the dimensions of various features in the hidden layer. We will continue to optimize autoencoder models and generate more representative features to provide a more reliable unsupervised learning-feature-extraction architecture for further research.

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

Funding: This research was funded by the National Key R & D Program of China, grant number 2019YFC0118202.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://sail.usc.edu/iemocap/iemocap_release.htm, <http://www.openslr.org/12/> (accessed on 1 June 2021).

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

References

1. Gangamohan, P.; Kadiri, S.R.; Yegnanarayana, B. Analysis of Emotional Speech—A Review. In *Toward Robotic Socially Believable Behaving Systems—Volume I: Modeling Emotions*; Springer International Publishing: Cham, Switzerland, 2016; pp. 205–238.
2. Duric, Z.; Gray, W.D.; Heishman, R.; Li, F.; Rosenfeld, A.; Schoelles, M.J.; Schunn, C.; Wechsler, H. Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. *Proc. IEEE* **2002**, *90*, 1272–1289. [CrossRef]
3. Salmam, F.Z.; Madani, A.; Kissi, M. Emotion Recognition from Facial Expression Based on Fiducial Points Detection and using Neural Network. *Int. J. Electr. Comput. Eng.* **2018**, *8*, 52–59. [CrossRef]
4. Yang, J.; Zhang, F.; Chen, B.; Khan, S.U. Facial Expression Recognition Based on Facial Action Unit. In Proceedings of the 2019 Tenth International Green and Sustainable Computing Conference (IGSC), Alexandria, VA, USA, 21–24 October 2019; pp. 1–6.
5. Gentile, V.; Milazzo, F.; Sorce, S.; Gentile, A.; Augello, A.; Pilato, G. Body Gestures and Spoken Sentences: A Novel Approach for Revealing User's Emotions. In Proceedings of the 2017 IEEE 11th International Conference on Semantic Computing (ICSC), San Diego, CA, USA, 30 January–1 February 2017; pp. 69–72.
6. Xiong, H.; Lv, S. Factors Affecting Social Media Users' Emotions Regarding Food Safety Issues: Content Analysis of a Debate among Chinese Weibo Users on Genetically Modified Food Security. *Healthcare* **2021**, *9*, 113. [CrossRef]
7. Deng, J.; Frühholz, S.; Zhang, Z.; Schuller, B. Recognizing Emotions from Whispered Speech Based on Acoustic Feature Transfer Learning. *IEEE Access* **2017**, *5*, 1. [CrossRef]
8. Brien, M.G.; Derwing, T.M.; Cucchiari, C.; Hardison, D.M.; Mixdorff, H.; Thomson, R.I.; Strik, H.; Levis, J.M.; Munro, M.J.; Foote, J.A.; et al. Directions for the future of technology in pronunciation research and teaching. *J. Second Lang. Pronunciation* **2018**, *4*, 182–207.
9. Tejedor-Garcia, C.; Escudero-Mancebo, D.; Camara-Arenas, E.; Gonzalez-Ferreras, C.; Cardenoso-Payo, V. Assessing Pronunciation Improvement in Students of English Using a Controlled Computer-Assisted Pronunciation Tool. *IEEE Trans. Learn. Technol.* **2020**, *13*, 269–282. [CrossRef]
10. Khelifa, M.O.M.; Elhadj, Y.M.; Abdellah, Y.; Belkasmi, M. Constructing accurate and robust HMM/GMM models for an Arabic speech recognition system. *Int. J. Speech Technol.* **2017**, *20*, 937–949. [CrossRef]
11. Wang, D.; Wang, X.; Lv, S. An Overview of End-to-End Automatic Speech Recognition. *Symmetry* **2019**, *11*, 1018. [CrossRef]
12. Fayek, H.M.; Lech, M.; Cavedon, L. Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Netw.* **2017**, *92*, 60–68. [CrossRef]
13. Lieskovská, E.; Jakubec, M.; Jarina, R.; Chmulík, M. A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism. *Electronics* **2021**, *10*, 1163. [CrossRef]
14. Dahake, P.P.; Shaw, K.; Malathi, P. Speaker dependent speech emotion recognition using MFCC and Support Vector Machine. In Proceedings of the 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), Pune, India, 9–10 September 2016; pp. 1080–1084. [CrossRef]
15. Mustaqeem; Kwon, S. A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition. *Sensors* **2019**, *20*, 183. [CrossRef]
16. Barlow, H.B. Unsupervised Learning. *Neural Comput.* **1989**, *1*, 295–311. [CrossRef]
17. Hsu, W.-N.; Glass, J. Extracting Domain Invariant Features by Unsupervised Learning for Robust Automatic Speech Recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5614–5618.
18. Deng, J.; Zhang, Z.; Marchi, E.; Schuller, B. Sparse Autoencoder-Based Feature Transfer Learning for Speech Emotion Recognition. In Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland, 2–5 September 2013; pp. 511–516.
19. Deng, J.; Zhang, Z.; Eyben, F.; Schuller, B. Autoencoder-based Unsupervised Domain Adaptation for Speech Emotion Recognition. *IEEE Signal Process. Lett.* **2014**, *21*, 1068–1072. [CrossRef]
20. Han, K.; Yu, D.; Tashev, I. Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. In Proceedings of the 15th Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014; pp. 223–227.
21. Chen, M.; He, X.; Yang, J.; Zhang, H. 3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition. *IEEE Signal Process. Lett.* **2018**, *25*, 1440–1444. [CrossRef]
22. Zhang, Y.; Du, J.; Wang, Z.; Zhang, J.; Tu, Y. Attention Based Fully Convolutional Network for Speech Emotion Recognition. In Proceedings of the 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 12–15 November 2018; pp. 1771–1775. [CrossRef]

-
23. Mustaqeem, K.S. Att-Net: Enhanced emotion recognition system using lightweight self-attention module. *Appl. Soft. Comput.* **2021**, *102*, 107101. [[CrossRef](#)]
 24. Xia, R.; Liu, Y. Using Denoising Autoencoder for Emotion Recognition. *Interspeech* **2013**, *2013*, 2886–2889.
 25. Mustaqeem, K.S. CLSTM: Deep Feature-Based Speech Emotion Recognition Using the Hierarchical ConvLSTM Network. *Mathematics* **2020**, *8*, 2133. [[CrossRef](#)]
 26. Ghosh, S.; Laksana, E.; Morency, L.; Scherer, S. Learning Representations of Affect from Speech. *arXiv* **2015**, arXiv:1511.04747.
 27. Eskimez, S.E.; Duan, Z.; Heinzelman, W. Unsupervised Learning Approach to Feature Analysis for Automatic Speech Emotion Recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5099–5103. [[CrossRef](#)]
 28. Michael, N.; Vu, N.T. Improving speech emotion recognition with unsupervised representation learning on unlabeled speech. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 7390–7394.
 29. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 5206–5210.
 30. 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](#)]
 31. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.W.; McVicar, M.; Eric, B.; Oriol, N. Librosa: Audio and music signal analysis in python. In Proceedings of the 14th Python in Science Conference, Austin, TX, USA, 6–12 July 2015; pp. 18–25.
 32. Mustaqeem; Kwon, S. 1D-CNN: Speech Emotion Recognition System Using a Stacked Network with Dilated CNN Features. *Comput. Mater. Contin.* **2021**, *67*, 4039–4059. [[CrossRef](#)]
 33. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
 34. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv* **2019**, arXiv:1912.01703.
 35. Dissanayake, V.; Zhang, H.; Billingham, M.; Nanayakkara, S. Speech Emotion Recognition in the Wild Using an Autoencoder. In Proceedings of the INTERSPEECH 2020, Shanghai, China, 25–29 October 2020; pp. 526–530.
 36. Xu, M.; Zhang, F.; Zhang, W. Head Fusion: Improving the Accuracy and Robustness of Speech Emotion Recognition on the IEMOCAP and RAVDESS Dataset. *IEEE Access* **2021**, *9*, 74539–74549. [[CrossRef](#)]