electronics

MDPI

Article

Combining wav2vec 2.0 Fine-Tuning and ConLearnNet for Speech Emotion Recognition

Chenjing Sun [1], Yi Zhou [2], Xin Huang [1,*], Jichen Yang [3,*] and Xianhua Hou [1]

[1] School of Electronics and Information Engineering, South China Normal University, Foshan 528234, China; 2021022411@m.scnu.edu.cn (C.S.); houxianhua@m.scnu.edu.cn (X.H.)
[2] Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583, Singapore; yi.zhou@u.nus.edu
[3] School of Cyber Security, Guangdong Polytechnic Normal University, Guangzhou 510640, China
* Correspondence: huangxin@m.scnu.edu.cn (X.H.); yangjichen@gpnu.edu.cn (J.Y.)

Abstract: Speech emotion recognition poses challenges due to the varied expression of emotions through intonation and speech rate. In order to reduce the loss of emotional information during the recognition process and to enhance the extraction and classification of speech emotions and thus improve the ability of speech emotion recognition, we propose a novel approach in two folds. Firstly, a feed-forward network with skip connections (SCFFN) is introduced to fine-tune wav2vec 2.0 and extract emotion embeddings. Subsequently, ConLearnNet is employed for emotion classification. ConLearnNet comprises three steps: feature learning, contrastive learning, and classification. Feature learning transforms the input, while contrastive learning encourages similar representations for samples from the same category and discriminative representations for different categories. Experimental results on the IEMOCAP and the EMO-DB datasets demonstrate the superiority of our proposed method compared to state-of-the-art systems. We achieve a WA and UAR of 72.86% and 72.85% on IEMOCAP, and 97.20% and 96.41% on the EMO-DB, respectively.

Keywords: speech emotion recognition (SER); wav2vec 2.0; contrastive learning

Citation: Sun, C.; Zhou, Y.; Huang, X.; Yang, J.; Hou, X. Combining wav2vec 2.0 Fine-Tuning and ConLearnNet for Speech Emotion Recognition. *Electronics* **2024**, *13*, 1103. https://doi.org/10.3390/electronics13061103

Academic Editor: Byung-Gyu Kim

Received: 26 January 2024
Revised: 11 March 2024
Accepted: 15 March 2024
Published: 17 March 2024

1. Introduction

In recent years, Speech Emotion Recognition (SER) has emerged as a pivotal component in human–computer interaction and communication systems. It offers valuable insights into the emotional states of individuals, enabling applications like voice assistants (e.g., Siri, Alexa, Google Assistant) to adapt their responses accordingly, fostering more effective and personalized interactions. SER technology, as demonstrated by Chen et al. [1], has even been successfully integrated into emotional social robots, empowering them to track and respond to various basic emotions in real time. Moreover, SER holds immense potential for applications in diverse fields, such as fraud detection and psychological testing [2,3].

However, the accurate recognition and interpretation of emotions pose a significant challenge for computers. Humans can effortlessly perceive emotions through subtle cues like changes in pitch, volume, and tempo, however, human programmers struggling to program computers have a hard time capturing and comprehending these nuanced expressions. The multifaceted nature of emotional expression further complicates the task of precise emotion recognition. Therefore, this paper is dedicated to presenting a new approach to SER that can more accurately recognize emotions in speech and better serve people.

A typical SER system comprises two key components: feature extraction and emotion classification. So, we decided to improve the performance of the SER system from these two parts, that is, to improve the feature extraction part so that the extracted features contain more emotion-related information, and to improve the emotion classification part so that it has better classification capability.

Traditional SER systems involve the careful design of appropriate spectral features and rhythmic features extracted from the speech signal [4]. In recent years, spectral features such as Mel Frequency Cepstral Coefficients (MFCCs) [5] and log Mel-spectrograms (log-Mels) [6] have gained widespread adoption as speech emotion representations. Further, the study in [7] explored the combination of multiple spectral features.

Recent advances in deep learning have promoted the utilization of Deep Neural Networks (DNNs) for SER thanks to their strength in capturing intricate patterns. However, the limited size of current emotion datasets, constrained by the costly and time-consuming manual evaluation of verbal emotions, hampers the potential of DNNs in emotion recognition. Hence, researchers have dedicated efforts to address the challenge of training effective SER models with minimal training data.

One approach to tackle the issue of limited training data is using data augmentation to expand the size of the training set [8]. This approach improves the model's robustness and generalization to some extent. However, it also introduces certain drawbacks, such as the potential inclusion of "dirty data" that do not correspond to the labels assigned to them, which can negatively impact recognition accuracy. Therefore, striking a balance between data augmentation and maintaining data quality is crucial in addressing the challenge of small training datasets in SER.

Most of the recent emotion classification models use deep learning methods such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks. Chen et al. proposed to use attention-based convolutional recurrent neural networks for learning and emotion classification of 3-D log-Mels [6]. Aftab et al. proposed a fully convolutional neural network for feature extraction and classification of MFCCs [9]. Zhong et al. used separable convolution combined with attention mechanism and focus loss, which can learn emotional information well [10]. Ye et al. proposed a temporal-aware bi-directional multi-scale network based on dilated causal convolution to mine spatiotemporal information from MFCCs [11].

Deep learning methods, including CNN, RNN and LSTM, have dominated the landscape of emotion classification models. Chen et al. introduced attention-based Convolutional Recurrent Neural Networks (CRNNs) for learning and classifying 3-D log-Mels, yielding promising results [6]. Aftab et al. proposed a fully convolutional neural network that leverages MFCCs for feature extraction and classification [9]. Zhong et al. employed a combination of separable convolution, attention mechanism, and focus loss to capture emotional information effectively [10]. Ye et al. developed a temporal-aware bi-directional multi-scale network based on dilated causal convolution, enabling the extraction of spatiotemporal information from MFCCs [11].

In addition, transfer learning has emerged as a promising technique [12]. Generally speaking, transfer learning utilizes knowledge from previously learned tasks and applies it to related or newer ones. Usually, a typical transfer learning system has two parts: (1) a pre-trained model as a starting point; (2) new training data and the pre-trained model for the related or newer task, which is also defined as fine-tuning. In this direction, a small amount of emotional training data is used to tune the pre-trained model for SER.

When it comes to pre-training techniques, one prominent candidate is wav2vec 2.0 (W2V2) [13]. W2V2 utilizes self-supervised learning on a large-scale speech dataset to acquire speech representations that prove valuable for various downstream tasks, including automatic speech recognition (ASR) [13], speaker verification [14], and SER [12]. For example, Pepino et al. [15] used pre-trained W2V2 and Dense models with unaltered weights for SER. When employing W2V2 as the pre-training network, two commonly adopted fine-tuning approaches are:

- Vanilla fine-tuning (VFT): The pre-trained model, such as W2V2 is directly updated using data specific to the SER task. For example, Yue et al. [12] applied global average pooling to the contextualized representations generated by the pre-trained W2V2 model. Subsequently, a fully connected layer (FC) is employed for emotion classification based on the pooled representations. Similarly, Morais et al. [16] used pre-

trained W2V2 for feature extraction and used a mean average pooling aggregator and a linear classifier for classification. This straightforward fine-tuning method enables the pre-trained model to adapt and specialize for the SER task using task-specific data.

- Parameter-efficient fine-tuning (PFT): This approach involves selectively updating specific parameters while keeping others frozen, resulting in a more parameter-efficient fine-tuning process. In [17], PFT was implemented by freezing all the parameters of the CNN-based feature encoder in the pre-trained model W2V2. Only the parameters of the Transformer, another component of the pre-trained model were fine-tuned on the SER task. Additionally, an average pooling layer and a linear layer were incorporated as downstream classifiers to process the fine-tuned representations and perform emotion classification. This parameter-efficient approach allows for targeted updates to specific components of the pre-trained model while keeping the majority of the parameters fixed, optimizing the fine-tuning process for the SER task.

Both VFT and PFT approaches have shown suboptimal performance in SER. One possible explanation for this is the direct tuning of the W2V2 model for speech emotion classification without considering the substantial differences between SER and ASR tasks. Based on this, Wang et al. [18] proposed a two-stage approach for SER combining fine-tuning and the k-nearest neighbor model.

In general, widely used SER models, such as CNNs, typically comprise two main components: feature learning and classification. The feature learning component aims to extract effective feature representations from the input signal by applying a series of transformations. The classification component, on the other hand is responsible for assigning labels to the input signal based on the learned feature representations. The commonly used classification component consists of an FC layer and a softmax layer. The FC layer performs feature weighting to match the dimensionality of the learned feature representations to the number of classes and obtains a score for each class, while the softmax layer converts the features into probability form.

From this analysis, it becomes evident that traditional classification models only consist of feature learning and classification parts. Although the learned feature representations are used as input for the classification part, there is a lack of direct supervision for these representations. By introducing supervised learning to guide the feature representation learning process, improved classification performance can be achieved.

In this work, we propose two strategies, which differ from VFT's single-stage approach of directly using pre-trained models for SER tasks. They can be used to enhance the extraction of emotion-related information from W2V2 and improve classification performance:

- Instead of directly applying a pre-trained W2V2 model for SER, the pre-trained W2V2 model is fine-tuned to learn the emotion representation. In specific, the pre-trained W2V2 features are used as input to train an emotion extractor. The emotion classifier is trained using emotion training data and corresponding label information, and the component parameters in W2V2 are not frozen during the training process, which is different from PFT. Once the classifier training is finished, the emotion extractor can be obtained by removing the classification part of the trained classifier and accordingly, the emotion embedding can be extracted from the trained emotion extractor.
- In order to supervise the feature representation in the process of model training, a new model is proposed in this paper which has three parts: feature learning, contrastive learning and classification. The contrastive learning part plays the role of supervising the feature representation by making samples belonging to the same category exhibit similar feature representations while those from different categories exhibit discriminative representations in the model training stage.

The contribution of the work can be summarized as:

1. Feed-forward network containing skip connections (SCFFN) is proposed to tune the pre-trained W2V2 model by learning its output with the emotion training data and

corresponding label information. Finally, the emotion embedding extractor can be obtained by combining the tuned W2V2 model and the trained SCFFN.

2. A new model, ConLearnNet for short, has a contrastive learning function which can make samples belonging to the same category exhibit similar feature representations while those from different categories exhibit discriminative representations, is proposed for SER, which can supervise the feature representation in the process of model training.
3. The proposed emotion embedding and model are evaluated on the interactive emotional dyadic motion capture (IEMOCAP) [19] and the Berlin emotional database (EMO-DB) [20], respectively. The role of contrastive learning playing is revealed by the experimental results comparison.

The rest of the paper is organized as follows. Section 2 introduces how to extract emotion embedding by fine-tuning the W2V2 model. Section 3 introduces speech emotion recognition based ConLearnNet. Section 4 reports the studies on the IEMOCAP and the EMO-DB datasets. Finally, Section 5 provides a discussion, and Section 6 concludes the paper.

2. Emotion Embedding Extraction

In this section, we will introduce how to extract emotion embedding by fine-tuning the pre-trained W2V2 model with the training data and corresponding label information. Next, we first introduce the pre-trained W2V2 model and then W2V2 fine-tuning.

2.1. Pre-Trained W2V2 Model

W2V2 is a self-supervised learning model proposed by Facebook AI Research for speech tasks deriving speech representations from raw audio data [13]. As shown in Figure 1, the pre-trained W2V2 model consists of three sub-modules: feature encoder, Transformer and quantization [13]. Transformer and quantization can output context representation and quantized representation for the same input of the feature encoder, respectively.

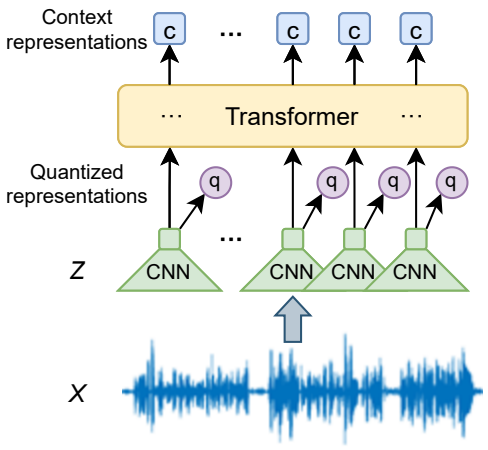

Figure 1. The architecture of pre-trained W2V2 model.

The feature encoder sub-module $f : X \rightarrow Z$ consists of a multilayer convolutional neural network that takes the original speech signal X as input and encodes the speech audio to generate potential speech representations $z_1, \ldots, z_T$ with output time steps $T = 25$ ms. The default setting of the feature encoder is constituted of seven CNN layers with 512 channels per convolutional layer, step size (5, 2, 2, 2, 2, 2, 2), and convolutional kernel width (10, 3, 3, 3, 3, 3, 2, 2). The masked feature encoder output is used as input to the Transformer sub-module to generate the contextual representation.

The transformer sub-module consists of multiple transformer encoders, in which the self-attention mechanism is able to capture the global information and can fully extract the high-dimensional contextual features.

The quantization sub-module uses product quantization to discretize the output of the feature encoder into a finite set of speech representations to perform quantization separately, making the features more robust.

2.2. W2V2 Fine-Tuning

In this paper, the pre-trained W2V2 model is fine-tuned to extract the frame-level emotional representation, and the fine-tuning framework is given in Figure 2.

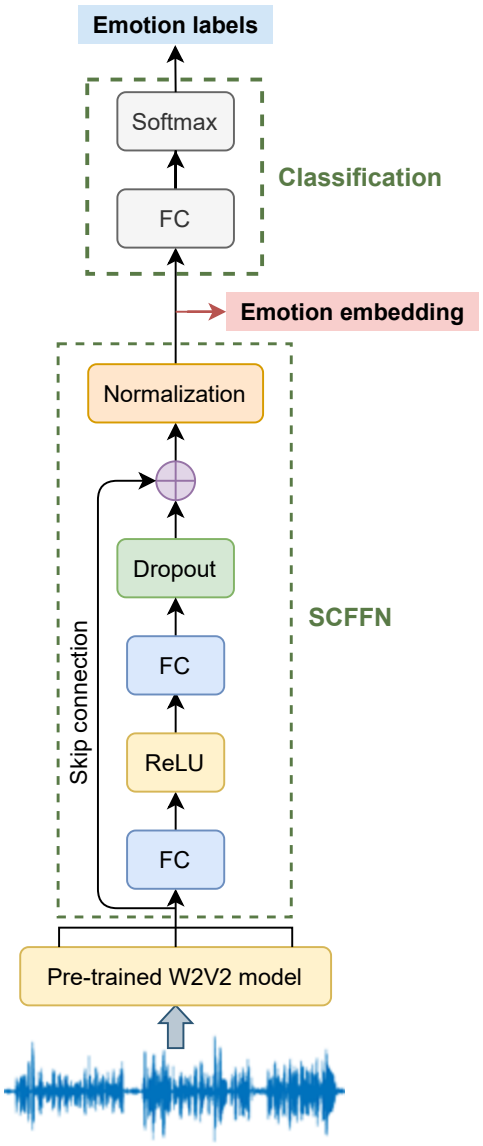

Figure 2. W2V2 fine-tuning framwork.

From Figure 2, it can be found that there are three parts in the W2V2 fine-tuning which are the pre-trained W2V2 model, SCFFN and classification. Meanwhile, it also can be found that the W2V2 fine-tuning is a classifier and the goal of the W2V2 fine-tuning is to train an emotion extractor to extract emotion embedding.

As shown in Figure 2, the SCFFN consists of five modules and a skip connection. The five modules contain two FC layers, one ReLU, one dropout, and one normalization. The role of each module is as follows:

- The FCs are used to learn the input and help enhance the model capabilities.
- The ReLU activation function is used to improve the nonlinear fitting ability of the network and accelerate the convergence of the model.

- The dropout can prevent the model from over-fitting.
- The use of skip connections can alleviate the gradient disappearance problem and prevent information loss.

The classification part is used to classify the output of SCFFN (emotion embedding), which consists of one FC and one Softmax, and the cross-entropy loss function is used as the loss function for this stage. The FC plays the role of converting the input feature dimension into the number of emotion types while the Softmax is used to obtain the corresponding probability, and then the output probabilities of the classes and the one hot form of the true classes are used to calculate the cross-entropy loss function. In order to extract emotion embeddings that are more applicable to the next stage, the classification part of this stage is the same as the classification part of the next stage, which classifies the four emotion classes of the IEMOCAP dataset and the seven emotion classes of the EMO-DB dataset, respectively.

With the emotion training data and corresponding label information, the W2V2 fine-tuning can be performed according to Figure 2. Once the W2V2 fine-tuning is finished, the emotion embedding can be extracted from the tuned W2V2 model and the trained SCFFN. Thus, we can regard the tuned W2V2 model and the trained SCFFN as an emotion extractor.

3. Speech Emotion Recognition Based on ConLearnNet

In this section, we will introduce the proposed model, i.e., ConLearnNet in detail. As shown in Figure 3, the ConLearnNet consists of a feature learning module, contrastive learning module and classification module. To be brief, the feature learning module is used to learn the input feature, and the contrastive learning module plays the role of supervising the feature representation by making samples belonging to the same category exhibit similar feature representations while those from different categories exhibit discriminative representations in the model training stage, the classification module is used to classify the learned features. Next, we will introduce them one by one.

3.1. Feature Learning

The feature learning module consists of average pooling layers, a bidirectional LSTM (BiLSTM) layer, half-step feed-forward (FFN) layers, and a Conv block. We introduce the function of the module in detail below.

To simplify the computational complexity of the network and to prevent overfitting, we add pooling layers to the network. The same as VFT in [21], average pooling is selected in the pooling layer to map the speech features to a smaller feature space, which allows for a smaller loss in the dimensional transformation of the speech features.

We introduce BiLSTM in this module that can handle feature information with a long-distance interval [22]. BiLSTM combines the advantages of a long short-term memory network and a bidirectional recurrent neural network which can better capture the bidirectional contextual information of speech data in time and is more robust and suitable for predicting time sequences.

Due to the complex network structure of our feature learning module, we propose to use skip connections around the sub-layers FFN and Conv in order to avoid the problem of gradient disappearance and at the same time to protect the information integrity during information transfer. The skip structure is inspired by the Transformer [23] model, which was proposed in [24].

In order to extract deeper features, we add a Conv block to the network. The Conv block consists of pointwise convolution layers, depthwise convolution layers and a gated linear unit (GLU), and also includes a batchNorm layer to accelerate the convergence of the model. The convolution kernel in convolution layers can capture both time and frequency domain information and has a remarkable ability to capture spatially advanced features. The number of parameters can be effectively reduced by using pointwise convolution and depthwise convolution to decompose a complete convolution operation into two steps. The GLU uses CNN and gating mechanisms to implement the RNN function, which im-

proves performance by retaining information strictly by temporal position when processing temporal data and speeds up operations through parallel processing structures [25].

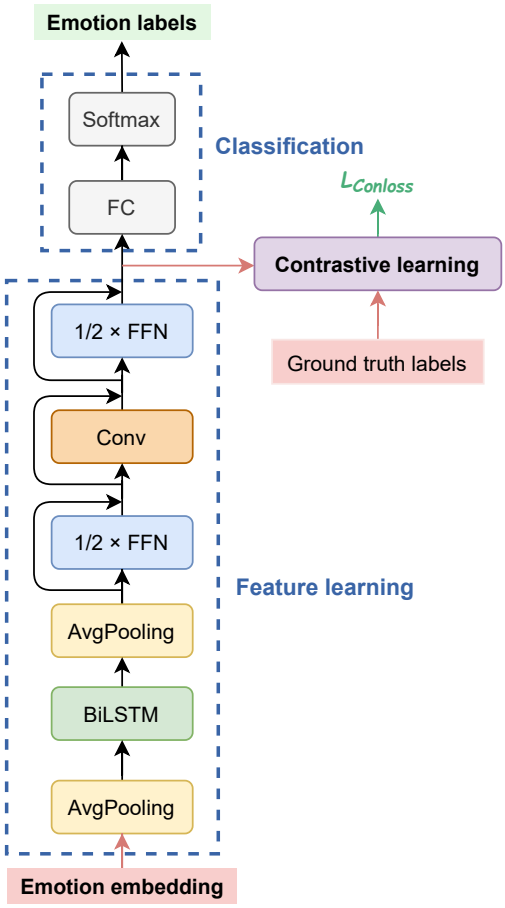

Figure 3. Model architecture of ConLearnNet.

Both Macaron-Net [26] and the Macaron-style half-step FFNs used in Conformer [27] have demonstrated that the Macaron structure can improve network performance. Based on this, we propose to use half-step skip-connected FFNs in ConLearnNet, before and after the Conv block, respectively, with the second feed-forward module followed by a final LayerNorm layer.

3.2. Contrastive Learning

Representation learning, an important process to improve the performance of deep learning models, can enhance the representation of raw speech data in the speech domain. A good speech feature representation will help to improve the performance of SER tasks. Contrastive learning is a representation learning method that optimizes the embedding representation. Through contrastive learning, samples belonging to the same category exhibit similar feature representations while those from different categories exhibit discriminative representations. Therefore, we introduce supervised contrastive learning to form a good classification model in this paper.

By contrastive learning, it is possible to achieve the clustering of points belonging to the same class pulled together in the embedding space, while separating clusters of samples from different classes [28]. In this paper, the above goal is achieved by using contrastive learning. As shown in Figure 3, the contrastive learning part has two inputs, one is the learned features obtained by the feature learning part and the other is their ground truth label, contrastive learning is performed by calculating the contrastive loss

between the learned feature and their ground truth label and then the features can be updated by back-propagation to generate more suitable features for final SER.

3.3. Classification

We use a combination of the FC layer and softmax layer as the classification part. The features obtained from the feature learning module are input into the FC layer for feature weighting to obtain the score of each category and then mapped to the probability of each category by the softmax layer for classification to obtain the emotion category corresponding to the speech data.

3.4. Loss Function

The total loss function consists of contrastive loss and cross-entropy loss commonly used for emotion classification, which is as follows:

$$\mathcal{L} = \mathcal{L}_{CE}(\hat{y}, l) + \alpha \mathcal{L}_{Conloss}(y, l) \tag{1}$$

where $\mathcal{L}_{CE}$ represents the cross-entropy loss function, $\hat{y}$ is the predicted probability distribution obtained after classification by the FC and softmax layers and l is the ground truth label corresponding to the input feature. $\mathcal{L}_{Conloss}$ represents the supervised contrastive loss function, y is the feature learned by the model before the FC layer classification, and $\alpha \in [0, 1]$ is the weight coefficient of the $L_{Conloss}$. Further,

$$\mathcal{L}_{CE} = -\sum_{k=1}^{C} t_k log(\hat{y}_k) \tag{2}$$

where C is the number of classes, $\hat{y}_k$ is the probability that the sample is predicted to be in class k, and t_k is the probability that the sample belongs to class k. t_k can only take the values 0 or 1.

$$\begin{aligned} \mathcal{L}_{Conloss} &= \sum_{i \in I} \mathcal{L}_{Conloss,i} \\ &= \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(y_i \cdot y_p / \tau)}{\sum_{a \in A(i)} \exp(y_i \cdot y_a / \tau)} \end{aligned} \tag{3}$$

In a batch, $i \in I \equiv \{1 \ldots N\}$ is the index of samples and p is the index of some other sample in the same class as i. The index i is called the *anchor*, the index p is called the *positive*, and $P(i)$ is the index set of all *positive* called *positives*. Other indices are called *negatives*. The $\cdot$ symbol denotes the dot product, $\tau \in \mathcal{R}^+$ is a scalar temperature parameter, a is the index of some sample other than i, and $A(i) \equiv I \setminus \{i\}$.

4. Experimental Results and Analysis

In this section, the proposed emotion embedding and ConLearnNet for SER are evaluated and corresponding analysis is given. Next, the details of the used databases, evaluation rule and experimental setup are introduced first.

4.1. Database

In order to evaluate the performance of the system, two of the most widely used databases in SER were used: the IEMOCAP in English and the EMO-DB in German. The use of databases in two different languages as datasets provides a better representation of the generalization performance of the method in this paper.

The IEMOCAP database was collected by the SAIL Lab at the University of Southern California and contains approximately 12 h of audio-visual recordings [19]. It contains five sessions of interactive dialogue between two people, Session 1, Session 2, Session 3, Session 4 and Session 5, performed by ten professional performers, with one male and one female performer participating in each session. The four emotions chosen in this paper are happy, neutral, angry and sad, with excited being categorized as happy. A total of 5531

voices are used in the training and test sets, including 1636 for happy, 1708 for neutral, 1103 for angry, and 1084 for sad.

The EMO-DB was developed by the Department of Technical Acoustics at the Technical University of Berlin and contains German speech presented by ten professional actors (five women and five men, labeled with the serial numbers 03, 08, 09, 10, 11, 12, 13, 14, 15 and 16) in seven emotions (neutral, anger, happiness, anxiety, sadness, disgust and boredom) [20], which were sampled at 48 kHz (later compressed to 16 kHz). In this paper, all seven emotions from the EMO-DB are used as the dataset, with a total of 535 speech items in the training and test sets, including 79 neutral, 127 anger, 71 happiness, 69 anxiety, 62 sadness, 46 disgust, and 81 boredom.

In order to compare with state-of-the-art methods, we use methods consistent with them for experiments and results analysis. For the IEMOCAP dataset, we use the speaker-independent 10-fold cross-validation method for experiments and evaluation, which can effectively avoid the possibility that the trained classifier is excellent for only a certain set of speakers, and is also more consistent with the situation that real-world speakers' speech has not been trained. For the EMO-DB dataset, we also use the 10-fold cross-validation method to experiment and evaluate the model. That is to say, we use nine pieces of the dataset for training and validation and another piece for testing in each evaluation, and the process is repeated ten times, using different pieces of training data each time.

4.2. Evaluation Rule and Experimental Setup

As mentioned above, we know that our system can be divided into two parts: the first is to extract emotion embedding for the input raw speech signal, and the second part is to classify the emotion embedding using the ConLearnNet. As the raw speech signals vary in length, the batchsize is set as 1 in the process of emotion embedding, which means that only one utterance is trained or validated at a time.

From Figure 1, we know that there are two types of representations that can be extracted from the pre-trained W2V2 models, which are context representation and quantized representation, respectively. In our study, context representation rather than the quantized representation of W2V2 is extracted from the pre-trained W2V2 models for the IEMOCAP and the EMO-DB in our study, the reason behind this is that emotion has something with its context.

In this paper, the Pytorch platform is used to conduct the experiments, and the Adam optimizer is used to optimize the classification cross-entropy. In the network parameters, the batch size of the ConLearnNet model is set to 4, the learning rate is 10^{-4}, and the dropout is 0.2. In both the emotion embedding extraction phase and the ConLearnNet emotion recognition phase, training is stopped when the training accuracy on the training set reaches 100% and the model is saved as the optimal model.

As shown in Table 1, there are two types of pre-trained wav2vec 2.0 large models used in our experiments, which are W2V2-large and W2V2-53, respectively. Wherein, W2V2-large is used for IEMOCAP while W2V2-53 is used for the EMO-DB. The reason is that W2V2-large is pre-trained on English databases and IEMOCAP is an English database, while W2V2-53 is pre-trained on multilingual databases and the EMO-DB is a German database.

Table 1. Some details about pre-trained W2V2 models.

Type	Model Type	Data for Pre-Training
W2V2-large	wav2vec 2.0 large	Librispeech
W2V2-53	wav2vec 2.0 large	Multilingual LibriSpeech, CommonVoice, BABEL

Due to the uneven distribution of labels, the use of traditional evaluation metrics such as accuracy alone may lead to over-optimism for emotion categories with large sample sizes. Therefore, both weighted average accuracy (WA) and unweighted average recall (UAR) are used as evaluation metrics [11] in this study. WA uses class probabilities to balance the recall metric across categories, while UAR treats each category equally to avoid

the model overfitting a particular category. WA is obtained by calculating the ratio between the number of correctly classified discourses in the training or testing set and the total number of discourses. UAR is computed as:

$$\text{UAR} = \frac{1}{K}\sum_{i=1}^{K}\frac{A_{ii}}{\sum_{j=1}^{K}A_{ij}} \tag{4}$$

where A is the column association matrix, A_{ii} corresponds to samples that are actually class i and are correctly classified as class i, A_{ij} corresponds to samples that are actually class i but are classified as class j, and K is the total number of emotion categories in the dataset [29]. Since we use the 10-fold cross-validation method, the WA and UAR results are averaged over all ten results.

4.3. Studies on IEMOCAP

4.3.1. Experimental Result and Analysis

Table 2 reports the experimental results on the IEMOCAP in terms of WA and UAR. From the table, we can find that our system can achieve a WA of 72.86% and a UAR of 72.85% on the IEMOCAP database.

Table 2. Experimental results on the IEMOCAP in terms of WA (%) and UAR (%).

Feature	Emotion embedding
Model	ConLearnNet
WA	72.86
UAR	72.85

Figure 4 shows the results of visualizing the obtained embedding after feature learning on the IEMOCAP test data using the t-SNE technique. The embedding is located before the FC layer after contrastive learning shown in Figure 3. This shows the feature distribution of the test data on the IEMOCAP after processing with ConLearnNet. It can be seen that the feature points of the same emotion category are each aggregated and separated from the feature points of different categories. However, there is still an overlap of feature points from different categories because the feature processing capability is still lacking.

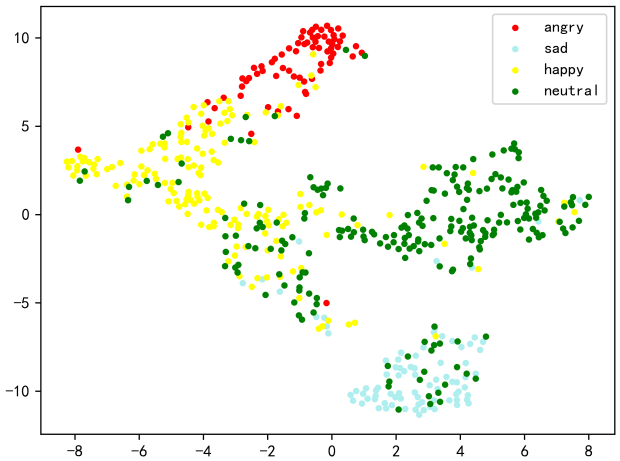

Figure 4. Visualization of t-SNE for feature distribution on the IEMOCAP dataset.

4.3.2. Ablation Experiment

Our system consists of two stages, the emotion embedding extraction stage and the ConLearnNet classification stage. In the following, we perform ablation experiments on the networks of the two stages separately.

As shown in Figure 2, the emotion extractor consists of pre-trained W2V2 and SCFFN, where SCFFN consists of a skip connection and an FFN. Here, we would like to analyze the modules in the emotion extractor from the experiments, so we perform ablation experiments on the IEMOCAP, including:

- w/o skip connection: To show the effect of the skip connection, the skip connection structure in the SCFFN is removed.
- w/o FFN: To show the efficiency of the FFN structure, the FC is used to replace the FFN in the SCFFN.
- w/o SCFFN: To show the importance of the SCFFN in fine-tuning, the SCFFN structure in the emotion extractor is removed.

From Table 3, it can be observed:

- The SCFFN using the skip connection structure improves the WA and UAR by 12.98% and 13.00%, respectively, over the SCFFN without the skip connection structure.
- The use of FFN in the SCFFN model improved the WA and UAR by 11.33% and 11.63%, respectively, over the use of the FC.
- The use of SCFFN in the emotion extractor improves the emotion recognition results WA and UAR by 13.43% and 9.70%, respectively, over that of no SCFFN.

Table 3. Ablation experiment results of the emotion extractor on the IEMOCAP dataset in terms of WA (%) and UAR (%).

Models	WA	UAR
ConLearnNet	72.86	72.85
w/o skip connection	59.88	59.85
w/o FFN	61.53	61.22
w/o SCFFN	59.43	63.15

As shown in Figure 3, ConLearnNet consists of three parts: feature learning, contrastive learning, and classification. Further, there are several modules in feature learning, which include Macaron, FFN, Conv and BiLSTM while there is one module in contrastive learning. Here, we are interested to know the modules in ConLearnNet playing from the experiments, ablation experiments are conducted on the IEMOCAP, which include:

- w/o contrastive learning: To display the role of contrastive learning playing, supposing the module of contrastive learning is removed from ConLearnNet, the obtained model can be named ConLearnNet-w/o contrastive learning.
- w/o Macaron: To display the role of FFN using the Macaron structure, normal FFN is used to replace the half-step FFN and is located after the Conv block.
- w/o FFN: To display the role of the FFN layer, the FFN layer is removed from ConLearnNet while keeping the other modules unchanged. The obtained model can be named ConLearnNet-w/o FFN.
- w/o Conv: To display the role of Conv block playing, the Conv block is removed while keeping the other modules unchanged. The obtained model can be named ConLearnNet-w/o Conv.
- w/o BiLSTM: To display the role of the BiLSTM layer playing, the BiLSTM layer is removed by adjusting the dimensionality of the pooling layer to eliminate the effect of dimensional changes due to the removal of the BiLSTM layer, keeping the other modules unchanged. In the same way, the obtained model can be named ConLearnNet-w/o BiLSTM.

From Table 4, it can be observed:

- The introduction of supervised contrastive learning on ConLearnNet improves WA and UAR by 2.09% and 3.09%, respectively, compared to without it.
- The FFN using the Macaron structure improves the WA and UAR by 1.75% and 1.36%, respectively, over the FFN without the Macaron structure.

- The addition of the FFN layer can improve WA and UAR by 1.56% and 0.69%, respectively, over those without the layer.
- Adding the Conv block can improve the WA and UAR by 10.20% and 10.46%, respectively.
- Adding a BiLSTM layer at the start of the network for contextual information extraction first, WA and UAR can be improved by 10.90% and 10.98%, respectively.

Table 4. Ablation experiment results of ConLearnNet on the IEMOCAP dataset in terms of WA (%) and UAR (%).

Models	WA	UAR
ConLearnNet	72.86	72.85
w/o contrastive learning	70.77	69.76
w/o Macaron	71.11	71.49
w/o FFN	71.30	72.16
w/o Conv	62.66	62.39
w/o BiLSTM	61.96	61.87

From the above analysis, it is known that all the modules in the emotion extractor and the ConLearnNet can contribute positively to the overall performance of the system on the IEMOCAP dataset.

4.3.3. Comparison with Commonly Used Features

To verify the effectiveness of the proposed emotion embedding, under the model of ConLearnNet, the commonly used features in the field of SER such as 3-D log-Mels and W2V2 are compared with the proposed emotion embedding. In which,

- 3-D log-Mels: First, calculate the log-Mel feature of the speech signal and its corresponding delta and delta-delta features, then, the static, delta and delta-delta are used as the first channel, second channel and third channel features to form 3-D log-Mel feature.
- W2V2: The original speech signals are directly used as the input of the pre-trained W2V2 model and W2V2 can be obtained.

The experimental results are shown in Table 5. From Table 5, it can be seen that the WA and UAR obtained from 3-D log-Mel as the input features are only 59.63% and 59.38%, respectively, which is much worse than emotion embedding. In addition, we also can observe that emotion embedding also performs better than W2V2. The reason may be that there is not so much emotion information in 3-D log-Mel and W2V2 while there is more emotion-related information in emotion embedding, which also confirms the proposed emotion embedding.

Table 5. Comparison with commonly used features on the IEMOCAP dataset in terms of WA (%) and UAR (%).

Features	WA	UAR
3-D log-Mels	59.63	59.38
W2V2	59.42	59.30
Emotion embedding	72.86	72.85

4.3.4. Confusion Matrix Analysis

To further analyze the experimental results, the confusion matrix for the IEMOCAP dataset is used to observe the actual recognition results of each type of emotion more clearly, to obtain the recognition performance of the system for each type of emotion, and then make targeted improvements to the system for the emotions with low recognition rates. Figure 5 shows the confusion matrix of the system when the SER is performed on the IEMOCAP database and the WA and UAR are 72.86% and 72.85%, respectively.

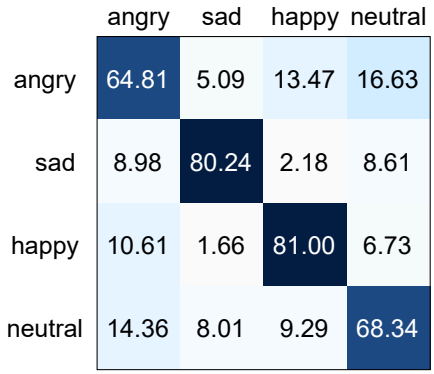

Figure 5. Confusion matrix of ConLearnNet on the IEMOCAP dataset with WA of 72.86% and UAR of 72.85%.

The confusion matrix on the IEMOCAP confirms that the system has excellent category discrimination performance. Observing the confusion matrix in Figure 5, it can be found that the system is good at recognizing happy and sad, probably because happy has more data available for training and the system can learn its emotional properties well. However, it is slightly worse at recognizing angry and neutral, which are easily recognized as each other. It may be that the system has not yet been able to capture the characteristics of the neutral category well because its own emotional factors are not prominent enough.

4.3.5. Comparison with the State-of-the-Art Systems

Table 6 shows the experimental results of our system and other state-of-the-art systems on the IEMOCAP dataset. From the table, we can observe that our method outperforms most of the systems in the table. Compared with the MFCC-TIM-Net [11] in Table 6, WA can be absolutely improved by 1.21% and UAR by 0.35%, respectively. It can confirm the effectiveness of the proposed method. In addition, by comparing the proposed W2V2 fine-tuning with VFT (systems 4) and PFT (system 5), we can say that the proposed W2V2 fine-tuning method can outperform the existing VFT and PFT. However, compared to the W2V2-SCL-kNN system [18] (system 7), our system is slightly less effective in recognition. We both used contrastive learning, while System 7 introduces the supervised contrastive learning in the feature extraction part of the fine-tuning of the W2V2, and we added it in the second stage of the classification part, which may result in our extracted feature representation being inferior to theirs. Also, System 7 uses the kNN model for label prediction and classification in the downstream inference classification task, which further improves the model performance.

Table 6. Comparison with the state-of-the-art systems on the IEMOCAP dataset in terms of WA (%) and UAR (%).

Systems	Features	Models	WA	UAR
1	log-Mel	CNN+Bi-GRU [10]	70.39	71.72
2	MFCC	SPU+MSCNN [30]	66.60	68.40
3	MFCC	Light-SERNet [9]	70.23	70.76
4	Raw Speech	VFT W2V2 [12]	62.68	65.11
5	Raw Speech	PFT W2V2 [17]	70.99	-
6	MFCC	TIM-Net [11]	71.65	72.50
7	Extracted features by W2V2-SCL	kNN [18]	74.13	75.14
Proposed	Emotion embedding	ConLearnNet	72.86	72.85

4.4. Studies on the EMO-DB

4.4.1. Experimental Result and Analysis

Table 7 reports the experimental results on the EMO-DB in terms of WA and UAR. From the table, we find that our system can achieve WA of 97.20% and UAR of 96.41% on the EMO-DB database, which means that our system can nearly classify all the emotion signals correctly.

Table 7. Experimental results on the EMO-DB in terms of WA (%) and UAR (%).

Feature	Emotion embedding
Model	ConLearnNet
WA	97.20
UAR	96.41

Figure 6 shows the results of visualizing the obtained embedding after feature learning on the EMO-DB test data using the t-SNE technique. The embedding is located before the FC layer after contrastive learning shown in Figure 3. This shows the feature distribution of the test data on the EMO-DB after processing with ConLearnNet. It can be seen that the feature points of each of the seven emotion categories are clustered separately and do not intersect with the feature points of different categories with clear classification boundaries. From the results, it can be seen that our system can obtain good results on two different language datasets, which proves that the system has strong robustness.

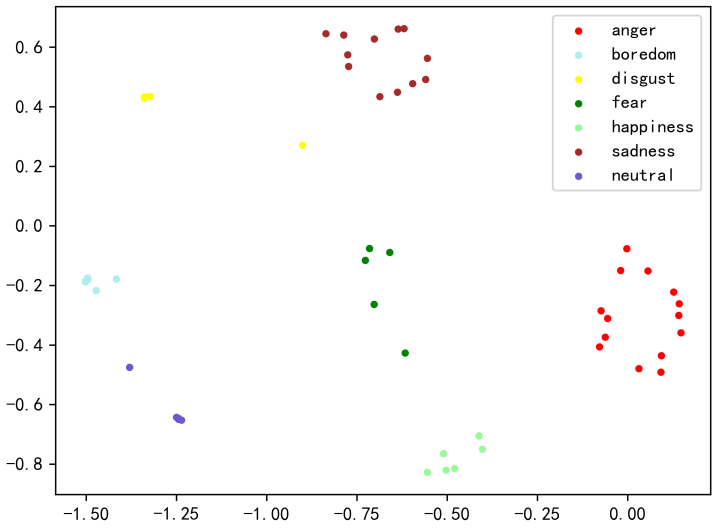

Figure 6. Visualization of t-SNE for feature distribution on the EMO-DB dataset.

4.4.2. Ablation Experiment

In the same way, we would like to show the role of the skip connection, the FFN and the SCFFN in emotion embedding extraction, as well as the role of the modules of contrastive learning, Macaron, FFN, Conv and BiLSTM in ConLearnNet, from the experiments on the EMO-DB database. Table 8 shows the results of the ablation study of the emotion embedding extractor on the EMO-DB dataset. From the table, we have the following observations.

From Table 8, several conclusions can be obtained:

- The SCFFN using the skip connection structure improves the WA and UAR by 1.72% and 1.43%, respectively, over the SCFFN without the skip connection structure.
- The use of FFN in the SCFFN model improved the WA and UAR by 1.23% and 0.29%, respectively, over the use of the FC.

- The use of SCFFN in the emotion extractor improves the emotion recognition results WA and UAR by 0.90% and 1.77%, respectively, over no SCFFN.

Table 8. Ablation experiment results of the emotion extractor on the EMO-DB dataset in terms of WA (%) and UAR (%).

Models	WA	UAR
ConLearnNet	97.20	96.41
w/o skip connection	95.48	94.98
w/o FFN	95.97	96.12
w/o SCFFN	96.30	94.64

Table 9 shows the results of the ablation study of the ConLearnNet on the EMO-DB dataset. From the table, we have the following observations.

Table 9. Ablation experiment results of the ConLearnNet on the EMO-DB dataset in terms of WA (%) and UAR (%).

Models	WA	UAR
ConLearnNet	97.20	96.41
w/o contrastive learning	96.18	95.96
w/o Macaron	96.19	96.02
w/o FFN	95.85	95.60
w/o Conv	96.11	96.03
w/o BiLSTM	93.27	93.60

From Table 9, several conclusions can be obtained:

- The use of supervised contrastive learning absolutely improves WA and UAR by 1.02% and 0.47%, respectively.
- The WA and UAR results after using the Macaron structure of the FFN improved by 1.01% and 0.39%.
- The addition of the FFN layer can improve the WA and UAR by 1.35% and 0.81%, respectively.
- The use of the Conv block can bring an improvement of 1.09% and 0.38% to the WA and UAR, respectively. The module does not improve as well on the EMO-DB as on the IEMOCAP, as shown in Table 4, suggesting that the module is more useful for the identification of datasets with larger amounts of data and that it has a greater improvement for models with low identification performance.
- The addition of the BiLSTM layer can improve WA and UAR by 3.97% and 2.81%. The improvement of the recognition effect of this module for the EMO-DB is also not as good as that for IEMOCAP as shown in Table 4, mainly because of the small sample size of the dataset, which itself is already able to achieve a good recognition rate, so the improvement is not significant, but it can further improve the SER performance of the model.

From the above analysis, it can be seen that all the modules in the emotion extractor and the ConLearnNet contribute positively to the overall performance of the system on the EMO-DB dataset.

4.4.3. Comparison with Commonly Used Features

To verify the effectiveness of the proposed W2V2 fine-tuning under the model of ConLearnNet, the extracted emotion embedding is compared with commonly used features in SER such as 3-D log-Mel and W2V2 on the EMO-DB dataset.

The experimental results are shown in Table 10. The results show that the WA and UAR obtained using 3-D log-Mels are only 88.81% and 89.16%, respectively. Meanwhile,

the WA and UAR obtained by W2V2 are 95.13% and 95.43%, respectively, while the WA and UAR of emotion embedding can reach 97.20% and 96.41%, respectively. This means that there is more emotion information in emotion embedding than that in 3-D log-Mels (W2V2).

Table 10. Comparison with commonly used features on the EMO-DB dataset in terms of WA (%) and UAR (%).

Features	WA	UAR
3-D log-Mels	88.81	89.16
W2V2	95.13	95.43
Emotion embedding	97.20	96.41

4.4.4. Confusion Matrix Analysis

To further analyze the experimental results, a confusion matrix is used on the EMO-DB dataset to more accurately analyze the actual recognition results of each type of emotion. By observing the recognition performance of the system for each type of emotion, we can make targeted improvements to the system for the emotion with poor recognition effect. Figure 7 shows the confusion matrix of the system when SER is performed on the EMO-DB database and the WA and UAR are 97.20% and 96.41%, respectively.

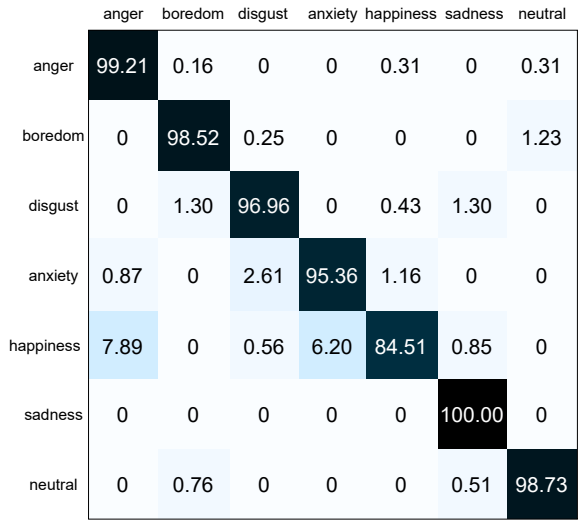

Figure 7. Confusion matrix of ConLearnNet on the EMO-DB dataset with WA of 97.20% and UAR of 96.41%.

The confusion matrix on the EMO-DB shows that the system achieves good discrimination for each category. Observing the confusion matrix in Figure 7, we can see that the system recognizes each type of emotion very well on the EMO-DB database, and can perform over 90% on each emotion category except happiness.

4.4.5. Comparison with the State-of-the-Art Systems

Table 11 shows the experimental results of our system and other state-of-the-art systems on the EMO-DB dataset, and the results show that our method outperforms these methods. Compared with the best system in Table 11 (MFCC-TIM-Net [11]) , WA can be absolutely improved by 1.50% and UAR by 1.24%.

Table 11. Comparison with the state-of-the-art systems on the EMO-DB dataset in terms of WA (%) and UAR (%).

Systems	Features	Models	WA	UAR
1	Raw Speech	1BTPDN [31]	89.16	88.46
2	MFCC	GM-TCN [32]	91.39	90.48
3	MFCC	CPAC [33]	94.95	94.22
4	MFCC	TIM-Net [11]	95.70	95.17
Proposed	Emotion embedding	ConLearnNet	97.20	96.41

5. Discussion

In our studies, ablation experiments were conducted on the IEMOCAP and the EMO-DB databases, respectively. The experimental results show that the proposed model, ConLearnNet is the optimal model among the sets examined, and the SER results on both datasets can be greatly improved after adding contrastive learning, which fully reflects the importance of contrastive learning for improving the feature representation. However, by observing the t-SNE visualization results in Figures 4 and 6, we can find that even with the introduction of the supervised contrastive loss function, the learned feature representation on the IEMOCAP is still inadequate, and we would like to further optimize the loss function afterward so that the model can generate a more easily classifiable feature representation.

In addition, under the model of ConLearnNet, 3-D log-Mel and W2V2 are used to compare the proposed emotion embedding. The experimental results on both datasets demonstrate the effectiveness of the proposed emotion feature. Compared with 3-D log-Mel, our emotion embedding retains more emotion-related information and is better adapted to the classification task. Compared to W2V2, the fine-tuned model is more adaptable to extract emotion-related information and the extracted feature is more beneficial to the emotion classification task.

To explore the system's recognition effect for each emotion, we also used a confusion matrix for observation. We observed that our system has average recognition for the angry and neutral emotions on the IEMOCAP dataset, but is strong at recognizing sad and happy emotions, whereas on the EMO-DB dataset, each emotion has a high recognition ability. The accuracy of the IEMOCAP dataset is much lower than that of the EMO-DB dataset, and the reasons for this include the following three. The first one is that the emotion recognition of the EMO-DB contains seven categories of emotions, while the IEMOCAP only recognizes four categories of emotions. The second is that the EMO-DB has fewer test utterances and only 53 utterances need to be categorized. The third is that the EMO-DB dataset has no spontaneous samples, so the emotion information in speech is more standardized.

Our studies also demonstrate that our system can obtain good recognition results on both English and German corpora. Further work will be carried out on a wider variety of corpora in order to further promote the system for practical applications.

6. Conclusions

In this paper, in order to extract emotional representation well, SCFFN is proposed to train an emotion embedding extractor at the base of the pre-trained W2V2 model to extract emotion embedding that can well characterize the emotion signals. In addition, to classify the emotion signals better, different from the traditional model that usually consists of the function of feature learning and classification, a new model that has a new function of contrastive learning to supervise features in the model training stage is proposed. Contrastive learning is used to make samples belonging to the same category exhibit similar feature representations while those from different categories exhibit discriminative representations.

The experimental results on the IEMOCAP and the EMO-DB datasets show that the proposed emotion embedding can perform better than commonly used features such as

3-D log-Mel and W2V2 extracted features, and that contrastive learning plays an important role in the system. In addition, we also find that the proposed system can achieve better performance than the state-of-the-art systems on each of the two datasets.

Author Contributions: Conceptualization, C.S. and J.Y.; methodology, C.S., J.Y. and X.H. (Xin Huang); software, C.S.; writing—original draft preparation, C.S.; writing—review and editing, J.Y. and X.H. (Xin Huang); supervision, Y.Z. and X.H. (Xianhua Hou). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NSFC (62001173, 62171188).

Data Availability Statement: Data available in a publicly accessible repository that does not issue DOIs. Publicly available datasets were analyzed in this study. These data can be found here: https://sail.usc.edu/iemocap/index.html; http://emodb.bilderbar.info/docu/ (accessed on 1 September 2023).

Acknowledgments: The author gratefully acknowledges the support of 2022 Guangdong Hong Kong-Macao Greater Bay Area Exchange Programs of South China Normal University (SCNU).

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

References

1. Chen, L.; Su, W.; Feng, Y.; Wu, M.; She, J.; Hirota, K. Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. *Inf. Sci.* **2020**, *509*, 150–163. [CrossRef]
2. Iulamanova, A.; Bogdanova, D.; Kotelnikov, V. Decision Support in the Automated Compilation of Individual Training Module Based on the Emotional State of Students. *IFAC-PapersOnLine* **2021**, *54*, 85–90. [CrossRef]
3. Cen, L.; Wu, F.; Yu, Z.L.; Hu, F. Chapter 2—A Real-Time Speech Emotion Recognition System and its Application in Online Learning. In *Emotions, Technology, Design, and Learning*; Academic Press: San Diego, CA, USA, 2016; pp. 27–46.
4. Bänziger, T.; Scherer, K.R. The role of intonation in emotional expressions. *Speech Commun.* **2005**, *46*, 252–267. [CrossRef]
5. Zhu, L.; Chen, L.; Zhao, D.; Zhou, J.; Zhang, W. Emotion Recognition from Chinese Speech for Smart Affective Services Using a Combination of SVM and DBN. *Sensors* **2017**, *17*, 1694. [CrossRef] [PubMed]
6. 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]
7. Andy Jason, C.; Kumar, S. An Appraisal on Speech and Emotion Recognition Technologies based on Machine Learning. *Int. J. Recent Technol. Eng.* **2020**, *8*, 2266–2276.
8. Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Epps, J. Direct modelling of speech emotion from raw speech. *arXiv* **2020**, arXiv:1904.03833.
9. Aftab, A.; Morsali, A.; Ghaemmaghami, S.; Champagne, B. LIGHT-SERNET: A Lightweight Fully Convolutional Neural Network for Speech Emotion Recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 6912–6916. [CrossRef]
10. Zhong, Y.; Hu, Y.; Huang, H.; Silamu, W. A Lightweight Model Based on Separable Convolution for Speech Emotion Recognition. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 3331–3335. [CrossRef]
11. Ye, J.; Wen, X.C.; Wei, Y.; Xu, Y.; Liu, K.; Shan, H. Temporal Modeling Matters: A Novel Temporal Emotional Modeling Approach for Speech Emotion Recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5. [CrossRef]
12. Yue, P.; Qu, L.; Zheng, S.; Li, T. Multi-task Learning for Speech Emotion and Emotion Intensity Recognition. In Proceedings of the 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, 7–10 November 2022; pp. 1232–1237. [CrossRef]
13. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; Volume 33.
14. Fan, Z.; Li, M.; Zhou, S.; Xu, B. Exploring wav2vec 2.0 on Speaker Verification and Language Identification. In Proceedings of the Interspeech 2021, Brno, Czech Republic, 30 August–3 September 2021; pp. 1509–1513. [CrossRef]
15. Pepino, L.; Riera, P.; Ferrer, L. Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. In Proceedings of the Interspeech 2021, Brno, Czech Republic, 30 August–3 September 2021; pp. 3400–3404. [CrossRef]
16. Morais, E.; Hoory, R.; Zhu, W.; Gat, I.; Damasceno, M.; Aronowitz, H. Speech Emotion Recognition Using Self-Supervised Features. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 6922–6926. [CrossRef]
17. Wang, Y.; Boumadane, A.; Heba, A. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *arXiv* **2022**, arXiv:2111.02735.
18. Wang, X.; Zhao, S.; Qin, Y. Supervised Contrastive Learning with Nearest Neighbor Search for Speech Emotion Recognition. In Proceedings of the Interspeech 2023, Dublin, Ireland, 20–24 August 2023. [CrossRef]

19. 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]
20. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the Interspeech 2005, Lisbon, Portugal, 4–8 September 2005; Volume 5, pp. 1517–1520.
21. Chen, L.W.; Rudnicky, A. Exploring Wav2vec 2.0 Fine Tuning for Improved Speech Emotion Recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5. [CrossRef]
22. Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Qadir, J.; Schuller, B.W. Survey of Deep Representation Learning for Speech Emotion Recognition. *IEEE Trans. Affect. Comput.* **2021**, *14*, 1634–1654. [CrossRef]
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
24. 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. [CrossRef]
25. Dauphin, Y.N.; Fan, A.; Auli, M.; Grangier, D. Language modeling with gated convolutional networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 933–941.
26. Lu, Y.; Li, Z.; He, D.; Sun, Z.; Dong, B.; Qin, T.; Wang, L.; Liu, T.Y. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv* **2019**, arXiv:1906.02762.
27. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv* **2020**, arXiv:2005.08100.
28. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised Contrastive Learning. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: New York, NY, USA, 2020; Volume 33, pp. 18661–18673.
29. Singh, P.; Waldekar, S.; Sahidullah, M.; Saha, G. Analysis of constant-Q filterbank based representations for speech emotion recognition. *Digit. Signal Process.* **2022**, *130*, 103712. [CrossRef]
30. Peng, Z.; Lu, Y.; Pan, S.; Liu, Y. Efficient Speech Emotion Recognition Using Multi-Scale CNN and Attention. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 3020–3024. [CrossRef]
31. Sönmez, Y.A.; Varol, A. A Speech Emotion Recognition Model Based on Multi-Level Local Binary and Local Ternary Patterns. *IEEE Access* **2020**, *8*, 190784–190796. [CrossRef]
32. Ye, J.X.; Wen, X.C.; Wang, X.Z.; Xu, Y.; Luo, Y.; Wu, C.L.; Chen, L.Y.; Liu, K.H. GM-TCNet: Gated Multi-scale Temporal Convolutional Network using Emotion Causality for Speech Emotion Recognition. *Speech Commun.* **2022**, *145*, 21–35. [CrossRef]
33. Wen, X.C.; Ye, J.X.; Luo, Y.; Xu, Y.; Wang, X.Z.; Wu, C.L.; Liu, K.H. CTL-MTNet: A Novel CapsNet and Transfer Learning-Based Mixed Task Net for the Single-Corpus and Cross-Corpus Speech Emotion Recognition. *arXiv* **2022**, arXiv:2207.10644.