Universiteit Leiden

Master Computer Science

Feature Extraction from Spectrums for Speech Emotion Recognition

Name:	Jincheng Li
Student ID:	S2534355
Date:	20/08/2021
Specialisation:	Computer Science and Advanced Data Analytics
1st supervisor:	Dr. E.M. Bakker
2nd supervisor:	Prof.dr. M.S.K. Lew

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

A speech emotion recognition (SER) system is a collection of methods that process and classify speech signals to detect the embedded emotions. In this work, we will focus on the feature processing and the classifier of an SER system. Previous works on SER show that the feature extraction process is an important task. Some emotion classes are difficult to distinguish, and the recognition rate is low. In this work, we study polymorphic spectrograms and employ deep end-to-end learning to improve the effectiveness of speech emotion recognition. To address the problems of high false recognition rate and low discrimination of traditional speech emotion recognition, we propose a multi-level speech emotion recognition framework, which firstly calculates the high similarity emotion class by a hash algorithm and adopts the speech emotion model based on long and short term memory network to generate the temporal speech emotion feature vector. Furthermore, it adds a multi-sampling rate data enhancement method to enrich the original samples, leading to a multi-level SER framework with improved accuracy. The main contribution of this paper is the design of an automatic SER system, using polymorphic spectrograms, with the components TF-CNN and Multi-SER.

Contents

1 Introduction **4**

2 Background and Related Work **5**
2.1 Representation of Emotions 6
2.2 Datasets 8
2.3 Data Augmentation and Self-Supervised Learning 9
2.4 Features and Feature Sets 10
2.5 Machine Learning Algorithms for SER 11
2.6 Transfer Learning 13
2.7 Deep Network for SER 13

3 Fundamentals **14**
3.1 Mel-frequency Cepstral Coefficients(MFCCs) 14
3.2 Convolutional Neural Networks 16
3.3 Long and Short-term Memory Networks 17
3.4 Emotion Similarity 18

4 Baselines **19**
4.1 The Baseline Method 19
4.2 The state-of-the-art SER model on IEMOCAP 19

5 Method **20**
5.1 Feature Extraction Using CNN 20
5.2 Multi-level Recognition Framework 23

6 Experiments **24**
6.1 File Processing 24
6.2 Train Test Split 26
6.3 Hyperparameter Setting 26
6.4 Experimental Set-up 26
6.5 Initial Experimental Results 26
6.6 Experimental Results: Baseline with Data Augmentation 27
6.7 Experimental Results: TF-CNN 28
6.8 Experimental Results: Multi-level SER 29
6.9 Results and Discussion 30

7 Conclusion **31**
7.1 Future work 32

1 Introduction

In the early years, it was considered that emotions are controlled by an individual and were seen as a kind of disruption of organized rational thought [1]. However, recently, emotional intelligence is recognized as an essential assessment of intelligence. As a consequence, it is important that conversational AIs have the capability to understand and express different emotions. In this context, the emotion recognition function could improve the effectiveness of human-machine interaction, for example, when integrated in smart devices. Furthermore, it can be applied to in-home or military healthcare scenarios for the early detection of psychiatric diseases [2]. With the recent development and widespread use of deep learning and smartphones, many people interact with virtual personal assistants, such as Siri and Cortana [3]. However, most systems do not comprehend the user's emotion, therefore, the SER technology is essential to further improve the performance and user experience of these kinds of conversational AIs.

There is a wide range of applications of speech emotion recognition, although many are still in the development stage. In remote voice services, the attendant can sense the emotional state of the user through the computer and can adjust the service at any time, thus improving the quality of the service. During criminal interrogation, computers can use voice recognition supplemented by heartbeat, brain waves, etc. to determine whether the person is lying or not. In toys, emotional toys can be added so that children's emotional processing skills can be improved at an early age. In remote voice instruction for teachers, both students and teachers can use voice recognition to perceive the emotions of both parties, thus improving the quality of teaching. Overall, when computers are given the ability to recognise and process human emotions, they can be made to work better for and with humans.

SER is a very promising and application. Although deep learning and other machine learning technologies have been successful in various fields, some characteristics of emotion and speech make SER a hard task. Because emotion is subjective, researchers have defined several emotion models. Therefore, there is no consensus about the output of SER systems Also it is hard to create a speech corpus in which people talk with each other in a real environment with real emotions. Instead, researchers invite actors into the laboratories or studios to record utterances with played emotions, there are examples with 2000 to 10000 utterances per corpus. However, thousands of data points clearly do not cover a continuous coordinate space, therefore, most of the time SER is considered as a classification problem, where subtle emotions are not captured because of its discrete characters.

Industrial output is difficult, and the development cycle is long: the research of speech emotion recognition algorithms and the construction of the corpus is still not perfect, the accuracy of the models on the corpus needs to be further improved, and the real-time testing is poor, which is not suitable for rapid productization and market launch, in the long term, this area still needs in-depth research and exploration.

Now, one of the challenges of SER is that the emotive features of speech are not as intuitive as those in the image domain. The traditional method requires manual feature extraction, selecting some emotive features from many speech features, which can be highly subjective and varies from corpus to corpus. Therefore, the focus of this thesis is the design, implementation, and evaluation of a multi-input feature extraction method, which uses the spectrogram

and the Mel spectrogram as input features, and a CNN to extract time-domain features and frequency domain features. Whereas in our selected baseline, for every only 13 MFCC features, in our method the spectrogram and Mel-spectrogram take the place of the MFCC matrix. In addition, an SER framework is built which calculates the emotion similarity, and then uses an LSTM model for the final emotion recognition. The IEMOCAP corpus is used for our experiments. The unbalance in the number of samples per category is handled by using the data augmentation method. The overall architecture of the SER system is shown in Fig.1. The rest of the paper is organized as follows, Section 2 will discuss the background of Speech Emotion Recognition, which contains some important components of SER. Section 3 will illustrate the baseline. Then Section 4 will propose the architecture of my SER, TF-CNN, and Multi-level SER, which are the main contribution of this paper. Finally in Section 5 contains the experiment and evaluation of the proposed model.

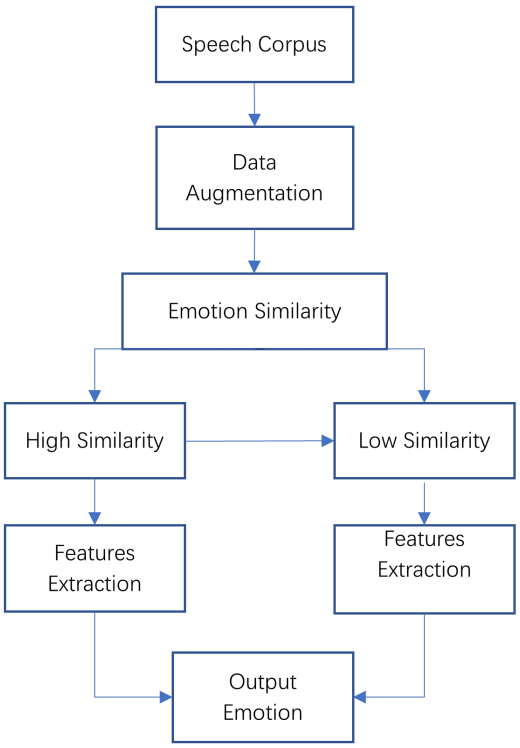

Figure 1: The overall architecture SER

2 Background and Related Work

Research on speech emotion recognition algorithms has been a major focus in the field of audio research. Especially in recent years, with the development of computer vision and artificial intelligence, more and more techniques as well as migrate to the field of audio algorithms. The general framework is shown in Fig.2. Current emotional speech recognition algorithms can be divided into machine learning methods based on statistics and model construction methods based on deep learning.

Figure 2: The typical framework of SER

Most researchers study speech emotion recognition as a classification problem. Each utterance is assigned a label. The utterance is a small piece of speech. In [4], an ensemble framework is used with a multitask DNN so that classified the gender of the speaker, therefore, this method is also suitable for language recognition.

[5] uses a spectrogram as input, two different convolution kernels were subsequently to extract time-domain features and frequency domain features separately, and then concatenate. The last layer used attention pooling. On the IEMOCAP database, it obtained 71.8% Weighted Accuracy (classification accuracy of all utterances), and 68% unweighted accuracy (averages the accuracy of each emotion class). They used 10-fold cross-validation with only one testing set. However, one of the baselines in [6], used 5-fold cross-validation, and the testing set was split for validation and testing. Therefore, they could not compare directly because they should do a 5-fold-cross-validation, too. Another problem is that before the concatenation, it did a softmax to the feature map to the bottom-up attention, there are three spectrograms in python_speech_map: amplitude, energy, log, the log might be the best because the range of amplitude and energy is too large.

[6] is the baseline of [5], they used the same database. The difference is that the pre-processing. In this work, the speech was segmented up to 3 seconds, and for the previous one, it is 2 seconds. The interesting is that this paper introduces a two-step method, initially, it passes through a four-classifier (4 emotions). If it is a natural class, then another three binary-classifiers are used to determine the final emotion, this could improve UA. The explanation is that most of the nature of a non-neutral emotion is neutral, however, the nature emotion class only occupies a small part, therefore further judgments are needed for the neutral category.

Most models just simply use the information from the spectrogram, which could not capture enough emotional features. In [7], it combined the CRNN with handcrafted high-level statistic features, HSF-CRNN.

2.1 Representation of Emotions

Current research scholars have divided speech emotion into two categories, discrete and continuous models. For each model, there are different classifications. The categories in the discrete model consist of basic and derived emotions, with basic emotions including happy, neutral, and sad, and derived emotions such as cold anger and hot anger. Usually, it represents as a one-hot vector. Usually, the label is assigned by human raters, however, sometimes for an utterance it could not obtain a majority decision, therefore, the discrete label is fuzzy sometimes.

In addition, some researchers consider emotions to be the result of a combination of factors

and to be a continuously changing value, a representation known as the continuous model. In contrast to discrete representations, the dimensional emotion model is more informative and is represented as a point on a coordinate axis, with the emotion space consisting of multiple emotion attributes (two or three). An emotion can be represented precisely as a point in space, with one direction of the spatial axis representing the range of positive emotions and the other the range of negative emotions, with the corresponding value being the strength or weakness of the emotion. For example, the neutral state should be at the origin of the coordinates. The point in space determines the extent to which a certain emotion is mapped onto the positive emotion and the extent to which it is mapped onto the negative emotion, thus enabling a quantitative representation of the emotion.

Fig.3 below represents a two-dimensional model that is expressed as arousal-valence(or activation-valence) space theory, which is one of the most recognized measures in the research field. The horizontal coordinate indicates the degree of validity, and the vertical coordinate indicates the degree of arousal. On the valence dimension, relatively positive emotions are valued positively (happy, relaxed, etc.), while negative emotions are valued negatively (Angry, Sad). The degree of activation indicates how strong the emotion is, e.g., Excited and Angry are both highly activated, while Neutral and Sad are less activated, which is consistent with human emotional expression.

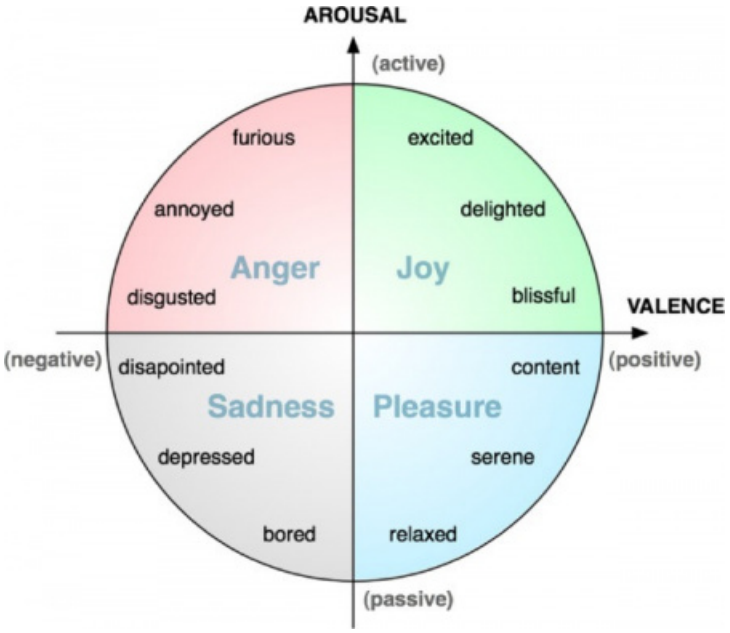

Figure 3: The 2-dimensional VA model [8]

Fig.13 below shows the three-dimensional model of emotion evaluation in the dimensions of arousal, valence, and intensity. Plutchik considers emotions to be multidimensional, with three important characteristics being the degree of emotional strength, emotional similarity, and emotional polarization. The three important characteristics of emotion are strength and weakness, emotional similarity, and emotional polarisation. Strength and weakness are different levels of happiness or sadness, etc. Similarity means that different emotions are expressed with a certain degree of similarity. Polarization refers to two extremes of emotion, such as pathos and ecstasy. The model can represent eight different levels of emotion. The model is

presented as a cone-like structure, with the strongest emotions located at the top and those at the bottom tending to be calm. The diagram shows that the closer the emotions are the more similar they are, the more polarized the two emotions are, and the two emotions at opposite corners are the most different.

The above descriptions of the two models show that each has certain advantages. The most obvious advantage of the discrete model is that it is simple to describe, using one label to describe a category of emotions, easy to understand and can describe emotions qualitatively. The continuous model, on the other hand, is more fine-grained in its ability to describe differences in emotions, can provide quantitative descriptions in multiple dimensions, is better suited to expressing complex emotions, and is more descriptive. The Tab.1 shows the main difference between the two emotional models. Therefore, many works combine these two kinds of emotion models. As is shown in Fig.3, the data points in quadrant 1 are regarded as Joy, and then Anger, Sad and Pleasure in sequence. If the attributes of the utterance are provided, the average value could be used to map to a discrete label. This explanation is less fuzzy. In this work, this emotion model will be applied, as is shown in Tab.4.

	The Discrete Model	The Continuous Model
Emotional description style	Adjective labels	Coordinate points in Cartesian space
Ability to describe emotions	A limited number of categories	Arbitrary number of categories
Time of Proposed	1980s	2000s
Advantages	Simplicity, Easy to understand, Easy to get started	Infinite ability to describe emotions
Disadvantages	Limited expressive ability	2consume large amounts of computational resources. High complexity

Table 1: The difference between Discrete and Continuous Emotional Model

2.2 Datasets

Speech databases are crucial for SER research. There are currently no unique regulations or general standards for the recording and construction of speech databases. SER databases in different languages such as English, Chinese, German, Urdu and other Western Languages [9]. There are discrete emotion databases and continuous emotion databases. The discrete type is mainly labelled with a single adjective-style label or a one-hot vector, while the continuous type is represented as a value in a spatial coordinate system. A continuous emotion database [10], exists as a continuous space and maps emotions to a point on some 3D or 2D coordinate. In the three-dimensional emotion space, each dimension is defined separately: Valence/Evaluation: the main function is to classify feelings as positive or negative. Activation/Arousal: the primary role is to reflect the fierceness of emotion by showing the level of physiological activation of the nerves associated with the emotion. Control/Power: it is used to reflect the degree of subjectivity that emotion has and to distinguish whether the emotion is due to the surrounding environment or is generated by the individual's subjective initiative. Furthermore, there are natural, performative, and guided SER databases.

In a performance-based speech database, the speaker does not necessarily express his or her real emotions but rather performs according to a developed scenario and script. The advantages are ease of access, flexibility in arranging the type of emotion and the character of the performer, etc. The disadvantage is that it is less realistic and more difficult to reproduce real scenarios. In natural speech databases, where speakers express their real emotions in real scenarios, access to the database is more difficult and costly, and the complexity of text and emotion changes is

difficult to control. In guided speech databases, speakers express their real emotions under the guidance of an external session, which can be used as a compromise between the first two types.

Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database [11] was published from SAIL lab at USC, which is the most widely used database for training an SER system, or other kinds of emotion recognition and analysis system. It is an English corpus, 5 males and 5 females were invited to record 12 hours of audio-visual data. After the recording, each session was manually segmented into many utterances, then, they were annotated by at least 3 human annotators. We will use IEMOCAP for our studies.

2.3 Data Augmentation and Self-Supervised Learning

When training a deep neural network, a large amount of data should be feed into the network, however, usually, a speech corpus is not enough because of the data sparsity. What is more, the data distribution is unbalanced. For example, in the Cortana database, over 60% sentences were labelled as neutral, because this dataset is collected by talking with the chat board. Data Augmentation is a kind of method that transform input data in a way but does not change the label [12], as a result, many new data points could be added into the dataset, to improve model robustness and avoid over-fitting. Traditionally, the techniques for speech data enhancement perform various operations on the time domain, however, in this case, depends on the methods, the waveform audio has to be converted to spectrum to feed into the neural network as the input. Nevertheless, Park et al. [13] proposed a novel approach to speech data augmentation by performing operations directly on a spectrogram. Like the data augmentation methods used in computer vision, such as Flip, Rotate, Scale, Random Crop or Pad, Colour jittering or Noising. Overall, the vector of pictures or a frame of video is transformed.

Different speakers have different vocal tract lengths, which is the source of the speech variations [14]. For adult females, the vocal tract is about 13cm and for adult males, it might be over 18cm. As a result, the centre frequencies of the vocal cords can differ by up to 25% between speakers. Generally speaking, in classical speech recognition, to remove speaker variations of training data, a method called Vocal Tract Normalization (VTLN) technique is widely used [15]. For the whole database, using a warp factor α to linearly warp the frequency axis of speech signals. Inspired by this, to augment the corpus, for each utterance, a random warp factor α is generated to warp the frequency axis. This method is called Vocal Tract Length Perturbation [12].

Many experts believe that deep learning is essentially about two things: Representation Learning and Inductive Bias Learning. In the case of representation learning, direct supervised learning of semantics performs well, but it requires many samples and is often designed for a specific task, making it difficult to be transferable. In addition, many speech corpora were not labelled, such as the RECOLA dataset, therefore, the model should have the ability to extract knowledge from unlabelled data. Therefore, in the field of SER, self-supervised is the future. There are two main types of methods for self-supervised learning, generative method, and contrastive method. For contrastive methods, the model does not need to know the details of the features, if the learned features are sufficient to distinguish it from other samples. In [16], the team used self-supervised contrast learning to learn emotional features, inspired by

the SimCLR [17] approach. During the training process of SimCLR, because the augmented data has the same label, then the contrastive loss could be calculated, and their best result could match some supervised learning architectures. In the abstract, the team claimed that their learning algorithm does not need a memory bank. However, there should be thousands of training batch size, using a TPU with 32-128 cores.

The Speech SimCLR [16] is inspired by the success of SimCLR because the visual signal and speech signal are all continuous. Inspired by the SimCLR, it also tried to maximum the contrastive loss between different augmented samples. In addition, it also learned reconstruction loss of input representation. In the beginning, for the data augmentation, they used the WavAugment tool [18].

2.4 Features and Feature Sets

Different extracted speech features directly affect the accuracy of the recognition results. Speech contains a variety of features, and some feature parameters can reflect the difference between emotions. Commonly used features for SER include phonetic features, prosody features and spectral-based correlation features. Statistics on these features are suitable for SER analysis.

Rhythmic features are weakly correlated with the specific textual content but focus on factors such as sound length, pitch, and speed. The characteristics vary from person to person and in different emotions, for example, an extrovert person is characterised by a high volume and speed of speech. Whereas more introverted person often speaks in a low and slow voice. There can also be clear variations for different emotions. For example, when a person is sad, he or she speaks in a breathless, soft, and slow voice, while when he or she is excited and happy, the voice is crisp and loud, which shows that there is a strong relationship between rhythmic characteristics and the expression of emotions. Rhythmic characteristics include important factors such as duration, fundamental frequency, and energy.

The speaker's voice is well textured, i.e., the voice is measured in terms of its quality, portrayed in terms of purity, clarity, and ease of recognition. For example, children's voices are generally pure and clear, while older people's voices are blurrier and more muffled. Researchers believe that voice quality characteristics are strongly related to emotion. Some of the most used characteristics of sound quality today are format frequency, jitter and shimmer, bandwidth, glottal parameters, etc.

In the field of speech emotion, the emotions depend on the voice characteristics and linguistic content [19]. The latter one often works with Natural Language Process (NLP); therefore most SER projects focus on the voice features. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) feature set [20] contains 62 features, these 62 features are all high-level statistics functional (HSF) features, which are calculated from 18 Low-Level Description (LLD) features. Among those 18 LLD features, there are 6 frequency related features, 3 Energy/Amplitude related features and 9 Spectral features. Besides, eGeMAPS is the extended version of GeMAPS, adding some features based on 18 LLDs, including 5 spectral features and two frequency-related features. After that, calculate the arithmetic average and coefficient of variation, which means that by calculating the standard deviation and then normalizing with the arithmetic average,

14 statistical characteristics could be obtained. In addition, the arithmetic average of spectral flux is calculated only in the unvoiced region, and then the arithmetic average and coefficient of variation of 5 spectral features are calculated only in the voiced region, and 11 statistical features are obtained. Finally, there is an equivalent sound level. Therefore, there are 26 extended features, with 62 features in GeMAPS, there are 88 features in eGeMAPS.

In the baseline [21], 13 MFCC features per frame were used, and finally, the input is an MFCC matrix. In some similar works [8], they used 29 features per frame: 26 MFCC features along with F_0, energy and voice probability. The core idea is that using CNN to extract HSF from some LLDs such as MFCC or F_0. However, spectrogram and Mel spectrogram also contains many useful features, therefore in this paper, we will try to use CNN to extract features from spectrogram and Mel spectrogram.

2.5 Machine Learning Algorithms for SER

Statistically based machine learning related algorithms are primarily a process of generalisation, summarization and inference using statistical related mathematical modelling theory. One of the more difficult processes is the construction of speech emotion recognition algorithms, where the accuracy of recognition relies heavily on the extraction of features from the audio signal. The audio signal is rich in feature information, like the various features extracted from images, where the signal can be processed globally and locally, but also through frequency, amplitude, and energy.

After the feature extraction phase has been carried out, it needs to be processed by machine learning methods. Support vector machines (SVM), Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), K-nearest neighbours (KNN) and other related algorithms have been used and promoted in the field of speech emotion recognition. Depending on the specific requirements, each method has its advantages and corresponding limitations of use. With the continuous development of technology, integrated methods integrating these basic algorithmic frameworks have emerged, which can better integrate the advantages of each of these algorithms and have strong application value.

The HMM model has shown good results in the tasks of speech emotion recognition and multi-classification and can obtain more accurate emotion recognition in a specific database. In [22], they proposed an HMM-based model for six classifications of speech emotions. The acoustic features used are LFPC, LPCC, MFCC, etc. Four states are created between the six emotions and the speaker, and the topology of the HMM model is a fully connected model. The corpus dataset is based on Burmese and Mandarin, with a total of 720 dialogues, and the training and test sets are partitioned in a 6:4 ratio. The accuracy of the speaker recognition sentiment classification is 65.8% under normal conditions, with 78.5% for Burmese and 75.5% for Mandarin. This shows that the Hidden Markov Model has certain advantages in speech emotion recognition, but the data sets in the field of speech emotion recognition are different, and the results of the model are very different, and the data sets used in this method are easily distinguishable from those in the industry. The accuracy of speech emotion recognition is greatly reduced when the HMM model is trained on a less differentiated and more complex data set. [23]

Gaussian mixture model (GMM) is a linear combination of several different Gaussian distributions to simulate real speech emotion data by mixing the distributions of Gaussian random variables. The model can simulate arbitrarily shaped density distributions and can be equated to a Hidden Markov Model with continuity and only one state. Since the Gaussian mixture model is less demanding than the Hidden Markov model for training and testing, the model has a wide range of applications in acoustic modelling. The advantage of GMM is that it is significantly better than HMM when only global features are available for speech emotion utterances, but the disadvantage is that all training and testing samples are modelled based on independent vectors, so GMM cannot use the temporal information of speech emotion samples. EM algorithms can usually solve this problem better by first estimating the hidden variables of each Gaussian model by assuming that the parameters of each Gaussian distribution are known, and then determining the Gaussian distribution in turn by the estimated hidden variables until the model converges. In addition, the optimal Gaussian component can be determined using kurtosis-based goodness-of-fit measures (GOF), Akaike information criterion, minimum description length and model order criterion.

Support vector machines (SVM) also play a major role in speech emotion recognition tasks, and SVMs are a kind of relatively stable method for machine learning classification. The basic principle of SVM is to segment the samples by finding hyperplanes to maximize the intervals. When the training samples are linearly separable, the linearly separable support vectors are learned by hard and soft spacing; when the training samples are linearly non-separable, the non-linear mapping of low-dimensional features to high-latitude features is required, and the non-linearly separable support vectors are learned by using kernel functions. In the speech emotion recognition model, support vector machines are not often used, although there is good theoretical support for the classification problem, it is not recommended to perform nonlinear hyper-segmentation to prevent the final model overfitting and other problems[24].

In the early stage of speech emotion recognition research, SVM models were relatively widely used, and a typical architecture is presented in[25]. The first two approaches use SVM to model each emotion and train five speech emotions. The difference between the two is that the former uses distance-maximization for the difference is that the former uses distance maximization for classification, while the latter passes the distance judgment to a three-layer perceptron and uses a nonlinear approach for classification. A third approach is an integrated approach, where multiple models are fused. Each of the three methods was tested with the FERMUS III The recognition accuracies of the three classification methods were 76.1%, 75.5%, and 81.3%, respectively, when they were speaker independent. 75.5% and 81.3%, respectively.

In general, SVM is relatively widely used in practice, but in the traditional method of speech emotion recognition, HMM and GMM have been dominant, and the advantages and disadvantages of each model in different application scenarios are also different. The advantages and disadvantages of each model vary in different application scenarios. Tab.2 shows some results for different classifiers on IEMOCAP.

Based Model	Features	Accuracy
SVM[26]	GSV-cov	67.8%
GMM[27]	60-dimensional MFCC	70.8%
HMM [22]	MFCC and epoch features	64.2%
KNN[28]	IS10	56.7%

Table 2: Some example work using machine learning based models

2.6 Transfer Learning

Data scarcity is another problem, generally, it is not easy to find suitable resources. In addition, due to the language variation and diversity, a model that is trained from a single corpus is not robustly applicable to other corpus or languages. Transfer learning methods can leverage the knowledge from one domain to another, therefore it could be used to address the problem of cross-corpus emotion recognition. If a pre-trained model is available, we can apply the parameter tuning of the model [29], even if the pre-trained model is used for another domain such as image recognition.

In [9], five corpora in three languages were used: English, German and Italian. Although those corpora all used the discrete emotional model, the models are different. The author did not re-label those corpora manually, instead, they built a positive/negative valence mapping. For example, negative valance contains such as Anger, Sadness, Disgust and so on, whereas positive valance contains Neutral and Joy. They used Deep Belief Networks(DBNs) as the classifier, which contains 3 RBM layers.

2.7 Deep Network for SER

Convolutional neural networks are a widely used deep learning method in recent years, especially in the fields of vision and natural language processing. They discovered that the network structural units of neurons may form complex networks, thus enhancing the computational power. Subsequently, a forward-propagating neural network was developed based on machine learning, and related concepts such as local perception were developed to make neural networks more powerful. The neural network has a strong ability to extract local features. CNN's can obtain effective representations of the original image CNN's can obtain effective representations of original images and discover abstract patterns of images, and have a good performance on large image data sets, which directly accelerates the development of deep learning networks. The development of deep learning networks has been directly accelerated. In recent years, deep learning related frameworks have also performed very well in the field of speech recognition. In recent years, deep learning related frameworks have also performed very well in the field of speech recognition, especially in the field of speech emotion recognition. The basic network architecture of CNN consists of convolutional layers, pooling layers, and a network of networks. The basic network architecture of CNN includes a convolutional layer, pooling layer, and fully connected layer.[30]

Because the length of each utterance could not be the same, sometimes experts try to segment the utterance. However, the emotion of sub-utterance could not represent the emotion of the whole utterance, assuming that each utterance has only one emotion. Therefore, a fully

convolutional neural network (FCN) is used to handle the variable-length speech spectrogram, so that the speech does not need to be segmented [19]. In addition, the input of the FCN could be the speech spectrogram itself, therefore we do not need to do the feature extraction process. Typical CNN such as AlexNext usually has several fully connected layers after the convolutional layers, and the feature map generated by the convolutional layer is mapped into a fixed-length feature vector. Therefore, typical CNN's only accept fixed-size input. In [29], Jonathan claimed that FCN could accept an input image of arbitrary size and use a deconvolution layer to upsample the feature map of the last convolution layer to restore it to the same size as the input image, thus generating a prediction for each pixel while preserving the spatial information in the original input image. This allows a prediction for each pixel, while preserving the spatial information in the original input image, and finally performs pixel-by-pixel classification on the upsampled feature map. Finally, this is equivalent to one training sample per pixel.

The only difference between the fully connected and convolutional layers is that the neurons in the convolutional layer are connected to only one local region in the input data and the neurons in the convolutional column share parameters. However, in both types of layers, neurons compute dot products, therefore their functional forms are the same. Therefore, it is possible to interconvert the fully connected and convolutional layers. For any convolutional layer, there should exist a fully connected layer that implements the same forward propagation function as it does. The weight matrix is a huge matrix with all zeros except some specific blocks, and in most of these blocks, the elements are equal. Conversely, any fully connected layer can be transformed into a convolutional layer.

3 Fundamentals

3.1 Mel-frequency Cepstral Coefficients(MFCCs)

In the previous work, they used OpenSimle to extract MFCCs features. I decided to calculate the feature set from scratch, according to the tutorial in [31]. The Mel-frequency is proposed based on the human hearing characteristics, which has a non-linear relationship with the Hz frequency. MFCC uses this relationship to calculate the Hz spectrum characteristics, mainly used for feature extraction of speech data and dimension reduction. For example, usually for a frame, there would be 256 or 512 sampling points, after the MFCC process, the most important 40-dimensional data can be extracted and finally, the first 13 dimensions are preserved.
The general process of calculating MFCCs is shown in Fig.4. Generally, it should go through

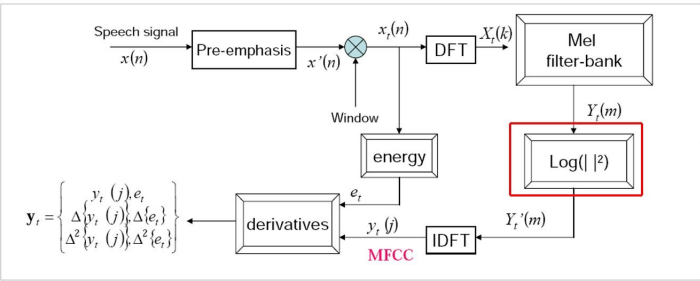

Figure 4: MFCC Process [31]

pre-emphasis, framing, windowing, Fast Fourier transform (FFT), Mel Filter Banks, and discrete cosine transform(DCT). The most important ones are FFT and filter banks because these two carry out the main dimensional reduction operation.
Here I used an utterance from CASIA corpus as an example. The raw signal is shown in Fig.5. The sampling rate is 16000Hz and there are 43876 sample points. The pre-emphasis filter is a high pass filter in the following equation:

$$H(t) = x(t) - \mu x(t-1)$$

The value of μ should be between 0.9 to 1, usually, 0.95 or 0.97 are taken, and I apply 0.97.

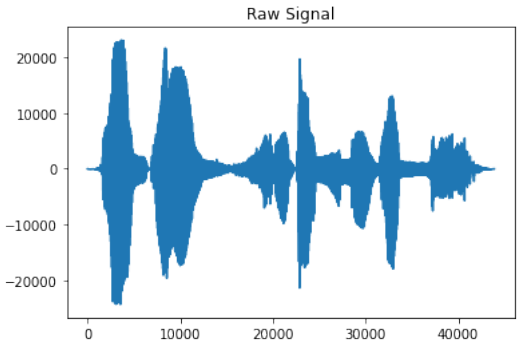

Figure 5: Raw signal

After the pre-emphasis, the signal is shown in Fig.6. The purpose of the pre-emphasis is to enhance the high frequency part, flatten the frequency spectrum of the signal, and guarantee that we can use the same SNR to calculate the frequency spectrum in the whole frequency band from low frequency to high frequency. Meanwhile, it is also to eliminate the effects of the vocal cords and lips during the process, to compensate for the high frequency part of the voice signal suppressed by the pronunciation system, and to highlight the high frequency format.

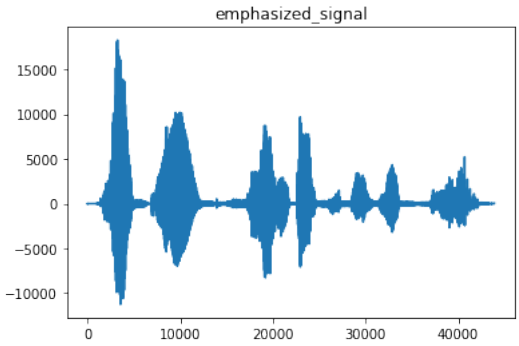

Figure 6: Emphasized Signal

To facilitate the speech analysis, the speech can be divided into small segments, which are called frames. First gather N sampling points into an observation unit, which is called a frame. Ordinarily, the value of N is 256 or 512, and the time covered is about 20 30ms. To avoid excessive changes between two adjacent frames, there will be an overlapping area between two adjacent frames. This overlapping area contains M sampling points, usually, the value of M is

about 1/2 or 1/3 of N [32].

The voice is constantly changing in a long-range and can not be processed without fixed characteristics. Therefore, each frame is substituted into the window function, and the value outside the window is set to 0. The purpose is to eliminate the signal disturbance that may be caused at both ends of each frame. In this work, Hamming window is applied. The Hamming window has the following equation:

$$w[n] = 0.54 - 0.46cos(\frac{2\pi n}{N-1})$$

Since the transformation of a signal in the time domain is not easy to see the characteristics of the signal, it is usually converted to energy distribution in the frequency domain to observe the different energy distributions, which can represent the characteristics of different speech sounds. Therefore, after multiplying the Hamming window, each frame must undergo a fast Fourier transform to obtain the energy distribution in the frequency spectrum. The spectrum of each frame is obtained by the fast Fourier transform of each frame after the windowing. The power spectrum of the speech signal is obtained by taking the mode square of the spectrum.

Since the human ear has different sensitivities to different frequencies and a non-linear relationship, we divide the spectrum into multiple Mel filter sets according to the sensitivity of the human ear. In the Mel scale range, the central frequencies of each filter are linearly distributed at equal intervals, but not at equal intervals in the frequency range, and this is formed by the formula for Hertz frequency to Mel frequency conversion. The discrete cosine transform is often used in signal processing and image processing for lossy data compression of signals and images, due to the strong "energy concentration" property of the discrete cosine transform: the energy of most natural signals (including sound and images) is concentrated in the low-frequency part of the discrete cosine transform, which means that each frame of data is compressed in a single pass of the dimension.

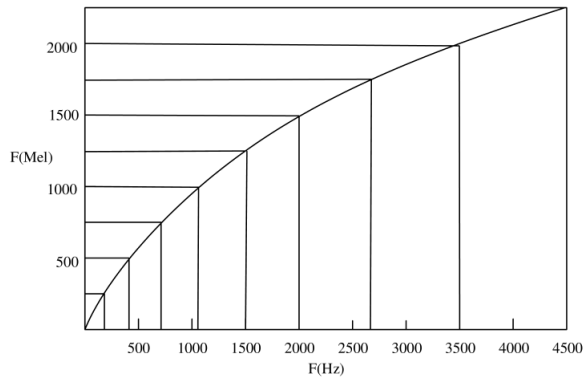

Figure 7: The relationship between normal scale and Mel frequency scale

3.2 Convolutional Neural Networks

A convolutional neural network is an end-to-end feature extraction structure, which can significantly reduce the workload of speech emotion feature extraction. With the backpropagation

mechanism, convolutional neural networks can automatically learn and optimize the network parameters. A convolutional neural network consists of several different layer structures, namely input layer, convolutional layer, pooling layer and fully connected layer. In recent years, deep learning methods have evolved rapidly, from the AlexNet to the current stage of better performance, such as VGGNet, GoogLeNet and ResNet, the model structure has become increasingly large and performance has gradually increased.

As the core component of the convolutional neural network structure, the main task of a convolutional layer is to extract features from the input data. With the deepening of layers, the extracted features gradually change from low-level features to high-level features, in which the central role is played by the convolutional kernel, which is designed as a set of parameters of a specific shape, and the designed convolutional kernel performs multiple operations on the input unit to complete a feature extraction. The convolutional neural network has the features of local connectivity, parameter sharing, and translation invariance, which can reduce the number of parameters while fully extracting features. Different convolutional kernels can extract different features from the same area of the image, which is equivalent to people looking at the same image from different angles and points of interest to obtain different aspects of information (e.g., edges, colours, contours, etc.), and then perform to reach a final output.

The pooling layer is also named as downsampling layer, which can reduce the computational effort by removing the redundant information from the feature map. The pooling layer can be divided into Max Pooling, Average Pooling and so on, according to the calculation method. The pooling layer usually defines a local window (e.g. 2*2, 3*3) on the feature map after the convolution calculation, which is slid over the whole image and pooled to calculate the statistics of all values in the window. The idea behind pooling is that a feature in a local area can be approximated by a statistic in the window. Pooling reduces the resolution of the feature map while removing some of the noise and preventing overfitting.

3.3 Long and Short-term Memory Networks

Long and short-term memory networks (LSTM) [33]have shown good results in speech recognition tasks. The LSTM was introduced in 1997 and has been used in a variety of applications, including image, natural language processing and speech. The LSTM is essentially a recurrent neural network, which was proposed to solve the problem of long term memory in RNNs due to gradient loss. The LSTM alleviates this problem by setting up multiple gate structures that constantly update the current cell state, allowing information to be remembered over a longer period. RNNs are designed to share parameters by having modules of neural units with identical parameters. The main difference between an LSTM and a normal RNN is the different designs of each cell. A normal RNN cell has a simple design, usually with a single activation function, while an LSTM achieves feature forgetting and updating by setting up multiple different gate structures.

Each cell of the LSTM contains three gate structures, namely the forgetting gate, the input gate and the output gate. The state of the previous moment is fed into the current moment and processed by non-linear functions (including Sigmoid and Tanh, etc.), discarding information that does not affect the current state and retaining useful information. By adding

and filtering each time, it can effectively deal with the problem of long term memory in RNNs.

The forgetting gate, which calculates the state h_{t-1} at the previous moment and the input x_t at this moment to obtain the information to be thrown away, selectively forgets the information at the previous moment. The input gate determines the information that needs to be stored in the current cell state, and selectively remembers the information, using the sigmoid gate, and creates a vector of candidate values through the tanh layer. Afterwards, the old cell state is updated in conjunction with the forgetting gate for the previous moment, C_t is the current cell state after the update. The output gate can get the current new cell state according to the input gate and the update gate, and the update of the cell state has been completed, and it is necessary to calculate the information of the output state, i.e., to calculate what information needs to be output from the current new cell state, which is also realized by the Sigmoid gate and Tanh gate. In this paper, we used double-layer LSTM[34]. The brief double-layer LSTM structure is shown in Fig.8 below.

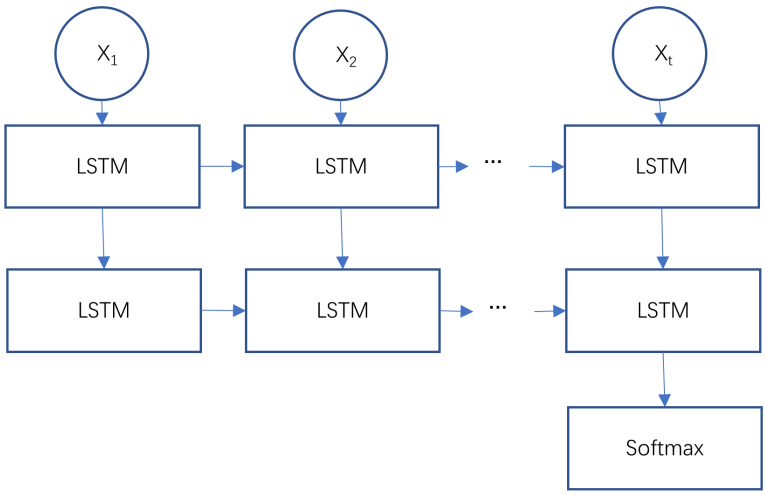

Figure 8: The double-layer LSTM

3.4 Emotion Similarity

Spectrograms contain a lot of information. We can calculate the hamming distance between two spectrograms to get the similarity. Here is a detailed procedure for calculating the similarity of emotion categories:

- Construct the emotion set. Define the emotion set $E = E_1, E_2, E_3, ..., E_c$, where c is the number of categories.
- The images in each category are scaled down to $n * n$ size, and the image structure should be kept.

- The images are transformed into $n * n$ grayscale images, denoted as G;
- Then calculate the average of all pixels in the greyscale graph G, which denoted as p_A;
- Iterate through all pixels p_i in G, if $p_i \geq p_A$, the hash value of that pixel is 1, else 0. Then, we could get a n_2 binary string, which is the Hash value of the graph, which denoted as H;
- Calculate the Hamming distance between the hash values H of the two images, the smaller the distance the more similar the two images are, the larger the difference a larger distance means a greater difference;
- Calculate the hash value of a category: if the number of samples in category E_1 is M, there are n^2 bits of hash value in every samples. E_1^i is the Hash value of the i_{th} sample in category E_1, $i = 1, 2, 3..., M$, E_1^{ij} is the hash value of the j_{th} bit. Therefore, the Hash value of E_1, HE_1 is calculated by:

$$HE_1^j = \begin{cases} 1, if count(E_1^{ij} = 1) > count(E_1^{ij} = 0) \\ \\ 0, if count(E_1^{ij} = 1) < count(E_1^{ij} = 0) \end{cases}$$

4 Baselines

4.1 The Baseline Method

In [21], the author's work proved that it is possible to train a Convolutional Neural Network by using a combination of different languages, and the model could recognize emotions independently from different languages. As to the dataset, they used an English corpus IEMOCAP [11] and a French dataset RECOLA. They used 13-dimensional MFCC features. The result shows that the accuracy of the emotion *joy* exceeds 90%, however, for the other three emotions, the accuracy is even hard to reach 20%, and theoretically, the accuracy of random classification is 25%. The author pointed out that the reason is that there are far more training data for joy than other categories. Therefore, we can improve this work by using more datasets or just using the data augmentation method. In the baseline, the CNN structure is quite a sample. It just has one convolutional layer and one dense layer. For the convolutional layer, they used 50 kernels of size 10×13, then, the max-pooling layer is applied with a 30-sized window and the stride of 3. Finally, the dense layer with 128 nodes with dropout and finally the 4-node output layer. The architecture of the baseline model is shown in Fig.9.

In this paper, the main problem is that the unbalanced distribution of data points leads to an unbalanced result. Therefore, the data augmentation method is used to get a balanced training dataset. In addition, only 13-dimensional MFCC features seem not enough, however, this paper is not focused on feature engineering, therefore, spectrogram and Mel spectrogram are used to extract features.

4.2 The state-of-the-art SER model on IEMOCAP

The method proposed in [35] holds the best result on the IEMOCAP so far. This work connected the phonetic theories and computational SER. Therefore, this work built two networks.

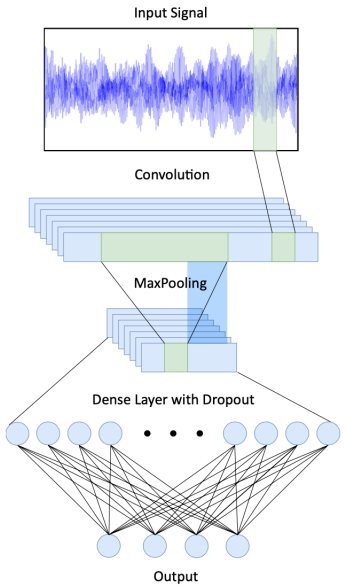

Figure 9: Architecture of CNN in baseline [21]

The first one is a Bi-LSTM model with an attention layer ($BLSTMATT$). The second model is a convolution-based self-attention(CSA)[35]. The $BLSTMATT$ focused on the sentence level so that it could be the varied length, however, CSA is one the frame level, therefore the length should be fixed. Both models used 23-dimensional log-Mel filter-bank features. As to the data, the first four sessions are used as a training set and session 5 is used as the test set. Finally, it reached 80.6% unweighted accuracy, which is the best one on the IEMOCAP so far.

5 Method

In the previous work, they just used grey-scale Mel-Spectrogram to extract features. In addition, in[8], the researchers combined MFCC and other phonetic features into a single feature matrix. In my work, I used coloured spectrogram and Mel-spectrogram as input to extract features, then fuse the two feature sets. Therefore, it is a multi-fusion feature extraction structure. What is more, it is difficult to distinguish some high similarity emotion categories, so that the overall recognition accuracy is affected. Therefore, we proposed a Multi-SER framework, for high similarity emotion classes, using a special Double-LSTM to do the fine-grained recognition.

5.1 Feature Extraction Using CNN

In the baseline and some other SER systems, the feature extraction is done by manual synthesis, however, there are many problems with this method. This paper proposes a multi-input feature extraction method, which is based on CNN. The model structure uses end-to-end training for learning, which can simplify the workflow and improve the overall recognition accuracy. Firstly, a speech segment is converted into spectrogram and Mel spectrogram, which is more suitable for the model. Then it is fed into the model for automatic feature extraction. The model is designed to take into account both the time and frequency domains of the spectrograms, and

finally, the high-level features of the images are fused to improve speech emotion recognition. The overall process of feature extraction is shown in Fig.10 below.

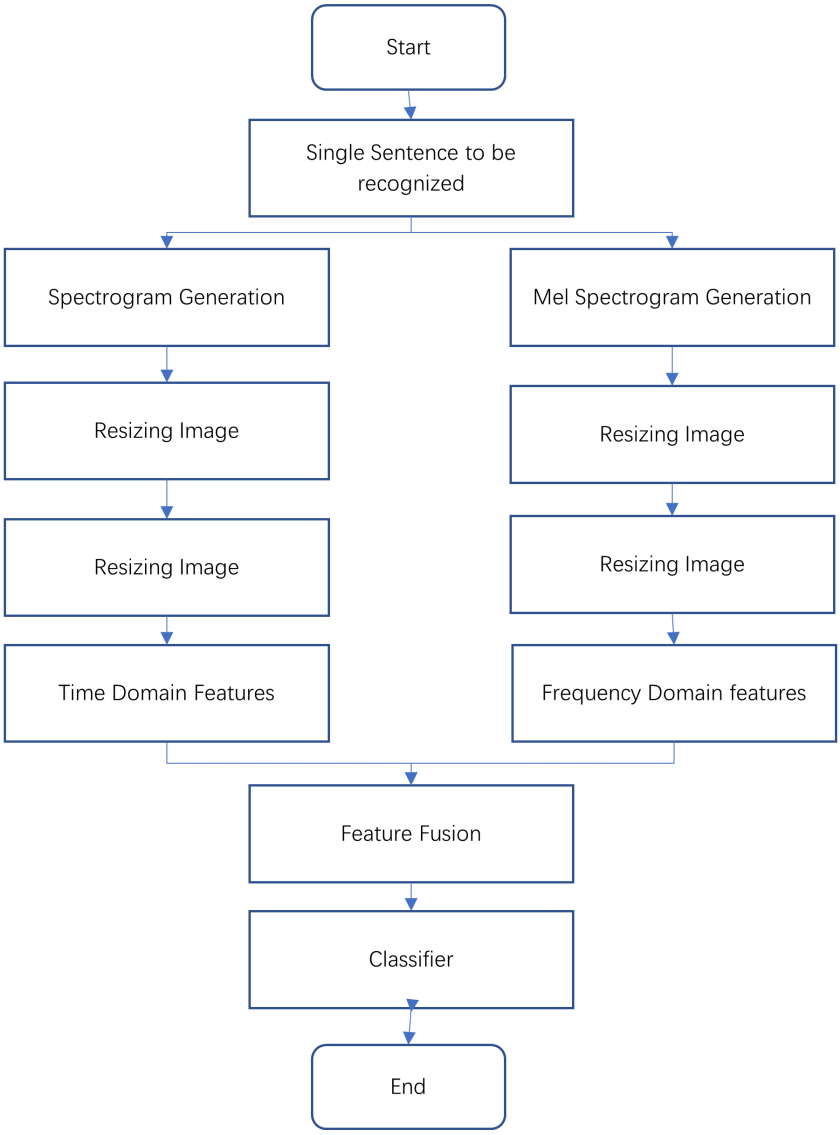

Figure 10: The Process of Time-Frequency Feature Fusion CNN

The spectrogram and Mel spectrogram by using the Python Library Librosa. The model could combine these two spectrograms and then extract the multi-spectral features. Meanwhile, because of the difference between the time domain and the frequency domain when doing the feature extraction, the multi-scale convolutional kernel is applied to solve this problem, so that the extracted features are more adequate and diverse. Then the model construction steps will be described.

Speech has the characteristic of short-time stability, so the dialogue segments in the database are segmented by time interval according to this characteristic. Assuming that the duration of

a single speech segment is x, the segmentation method with interval s and overlap m is used, then we could obtain $((x-s)/(s-m)+1)$ samples. Each sample is then normalized to the same size and used as the input of the model.

The horizontal and vertical coordinates represent different meanings in the speech spectrogram and Mel spectrogram, which contain both time and frequency domain representations. Conventional image recognition does not require the shape of the convolution kernel, but for special recognition tasks, it is necessary to design specific convolution kernels. The shape and size of the convolutional kernel in the convolutional layer directly affect the goodness of the extracted features. When the convolutional kernel receives a rectangular region at each position in the image, it means that each output contains a specific time-frequency information range. Therefore, this paper adopts a multi-scale convolutional kernel scheme and designs two sets of different shapes of convolutional kernels. If the window size of the convolutional kernel is $2 * k_1$ and $k_2 * 2$, in [31], they claimed that when $k_1 = 10$ and $k_2 = 8$, the best results are obtained.

As to the multi-fusion structure for the feature extraction. There are many kinds of representations of speech segment x, such as speech spectrograms and Mel spectrograms. As a time-frequency representation of the audio signal, speech spectrograms can describe the characteristics of speech in the time and frequency domains, and show different energies according to the colour, which is a description of speech itself. In contrast, the Mel spectrogram is transformed in the frequency domain into a frequency range to which the human ear is attuned, and the speech is described from the receiver side of the human ear. The model proposed in this paper, therefore, uses the speech spectrogram and the Mel spectrogram as input for feature extraction. If a speech spectrogram is S, the Mel spectrogram is M, and the kernel size is $8 * 2$. Then using the function below to calculate the feature maps S^* and M^*.

$$S^*_{i,j} = f(\sum_{m=0}^{7}\sum_{n=0}^{1} W_{m,n}X_{i+m,j+n} + W_b) \text{ [36]}$$

$$M^*_{i,j} = f(\sum_{m=0}^{7}\sum_{n=0}^{1} W_{m,n}X_{i+m,j+n} + W_b) \text{ [36]}$$

After we got the speech spectrogram and Mel spectrogram, we reshape them into $256*256*3$, which is a colour map with length 256, width 256 and 3 channels. Then feed them into the model. The speech spectrogram is fed into the first convolution layer, and the kernel size is set as $2 * 10$, the step size is 2 and the number is 64. The time-domain feature information is obtained from the original map through the transverse rectangular receiver field. After one layer of convolution, it is serially input to the maximum pooling layer with a window of $2 * 2$ and a step size of 1 for downsampling. Meanwhile, two spectrograms are fed into a parallel convolution layer with 64 convolution kernels of size $8 * 2$ and a step size of 2. The function of this layer is to obtain the frequency domain information from the spectrograms using a longitudinal rectangular receiver domain. After the second convolution layer is completed, it is fed serially into a maximum pooling layer with a window of $2 * 2$ and a step size of 1.Then the two types of feature maps are fed in parallel into the next convolutional layer with a 3*3 kernel size to extract advanced features. Finally, the last layer is a fully connected layer.

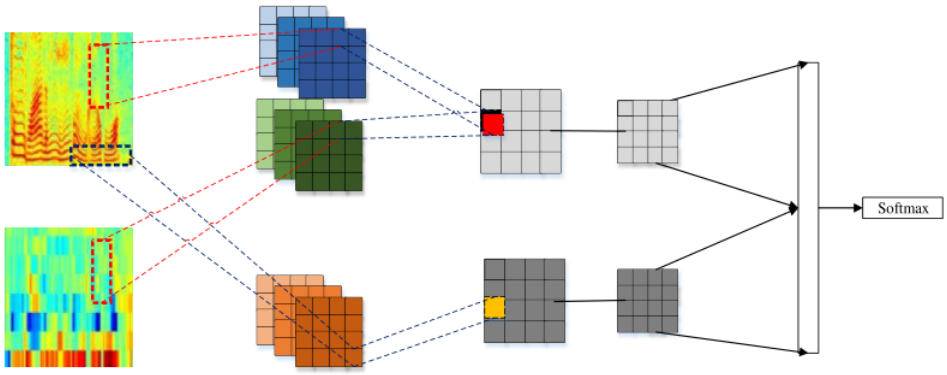

Figure 11: TF-CNN Model

5.2 Multi-level Recognition Framework

The model in the previous chapter effectively solves the problem of complex and inadequate feature extraction emotion. The analysis of speech emotions reveals that some emotion categories are difficult to distinguish due to their high similarity, which affects the overall recognition accuracy. This chapter proposes a multi-level speech emotion recognition framework, in which coarse-grained recognition is performed first, and then a hash algorithm is used to hash-code the speech spectrogram and calculate the category similarity to obtain a high similarity emotion set. In addition, the unevenly distributed data set is augmented.

In the first stage, coarse recognition was performed by calculating the sentiment similarity to obtain a high-similarity sentiment set; the second stage was designed as a two-stage recognition, where all the categories in the high-similarity category were first formed into a new class and added to the low-similarity sentiment category for training to generate Model 1. At the same time, only samples from the high-similarity sentiment categories are used for training, generating Model 2. With the multi-level recognition structure, then the results were obtained.

Due to the complexity of the human mind, there are many influences on the expression of emotions, some of which are difficult to distinguish and ambiguous, resulting in low recognition rates for some categories, which affects the overall accuracy of the results. Conversely, some categories are relatively easy to recognise, and the features are clearly distinguishable. To address this situation, the first step is to calculate the similarity of the emotion categories and obtain a set of highly similar sentiment categories. For each emotion class, select 10% spectrograms as the sample to calculate the similarity. The process was illustrated in Section 3.4 above.
The network consists of two main phases. First, remove the softmax layer, using a fully connected layer instead of as the input of the LSTM. Then the fully connected layer is connected to the LSTM network. In this paper, a two-layer 256-node LSTM network is used for the extraction of temporal features. Each memory unit learns the emotional features of the input speech at that time, and the emotional state is analyzed by the forgetting gate. The input vector is $1 * 1 * 512$, after the LSTM the size of the vector should be $1 * 1 * 4$. Then the vector will pass the softmax function and the final prediction of the categories will be obtained.

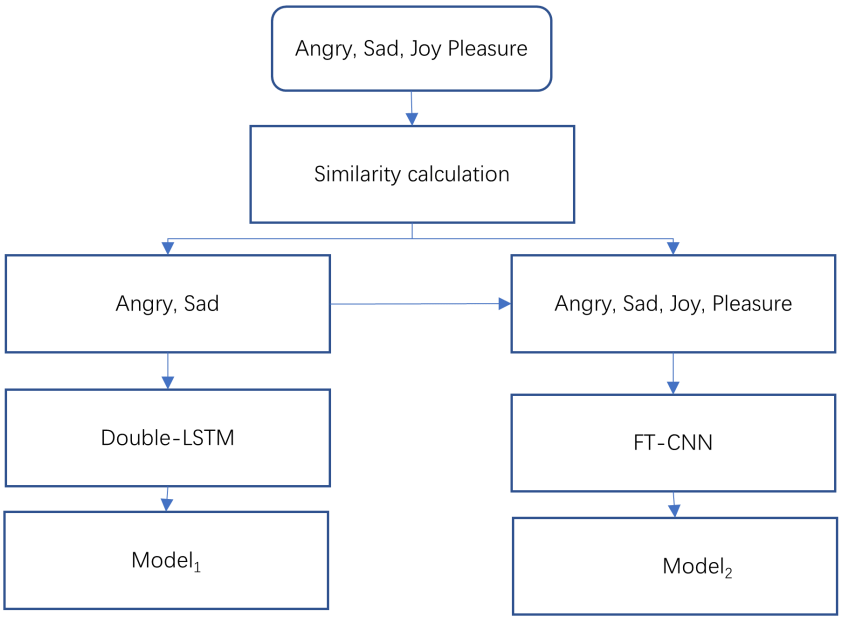

Figure 12: The flow chart of Multi-level Recognition Framework

6 Experiments

The ultimate goal of this research is to improve the recognition accuracy of the model predictions, and in this experiment, the model is evaluated using the confusion matrix and average accuracy on the test set. The confusion matrix is a frequently used evaluation method in classification tasks. From the confusion matrix, it is possible to visualise the rate at which each category is correctly classified and the rate at which it is misclassified into other categories.

6.1 File Processing

For the IEMOCAP dataset, there are 5 sessions. Each session contains about 30 dialogues, after the recording, each dialogue was split into single sentences. Each session has approximately 1800-2000 sessions. The Tab.3 shows the statistic of dialogue and sentences in different sessions. In total, there are 151 dialogues and 10039 sentences. In the experiment, each sentence will be a data point.

The publisher of the IEMOCAP also provided the annotated label. They used a tool called

	Session 1	Session 2	Session 3	Session 4	Session 5
Number of Dialog	28	30	32	30	31
Number of Sentences	1819	1811	2136	2103	2170

Table 3: The number of dialog and sentences in different sessions in IEMOCAP

ANVIL [37] ("annotation of video and spoken language"). Each piece of sentence will be labelled by at least two different human evaluators. Each sentence was annotated under the categorical emotional model and continuous emotional model. For the categorical emotional

model, there are 10 different emotions: neutral state, happiness, sadness, anger, surprise, fear, disgust, Frustration, Excited and others. As to the continuous emotional model, it used the Arousal-Valence-Dominance model, which is a three-dimensional emotional model. In this work, only the Arousal-Valence dimensions will be used. The arousal defines the strength of the emotion, and the valence describes the emotion as positive or negative. In the IEMOCAP, each dimensional attribute is an integer from 1 to 5, and each sentence was annotated by 2 or 3 human evaluators. Therefore, for each sentence the mean value of the arousal and valence. Due to the range of the value is 1 to 5, therefore 0 to 2.5 means low or negative, and 2.5 to 5 means high or positive. According to this, each sentence will be mapped into new labels, which is shown in Tab.4. And after the mapping, the number of data in different categories is shown in Tab.5

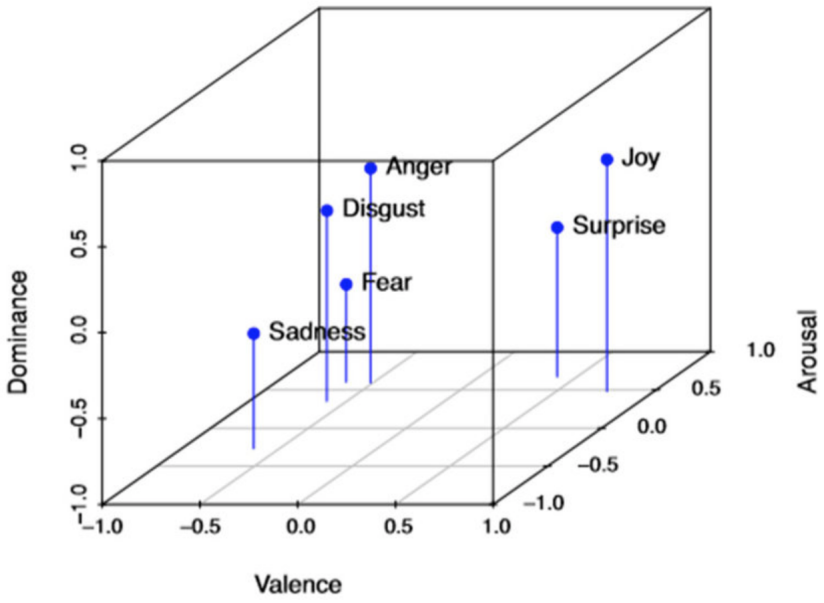

Figure 13: The VAD model [38]

Arousal/Valence	Negative[1, 2.5]	Positive(2.5, 5]
Low[1, 2.5]	Sad	Pleasure
High(2.5, 5]	Angry	Joy

Table 4: Mapping from VA value to new emotion labels [39]

	Session 1	Session 2	Session 3	Session 4	Session 5	Total
Sad	187	254	215	304	305	1265
Pleasure	395	446	560	586	593	2580
Angry	323	306	295	289	341	1554
Joy	914	805	1066	924	931	4640

Table 5: The number of data in different categories after mapping

6.2 Train Test Split

After the preprocessing of the IEMOCAP corpus, 10039 sentences was obtained. 80% of the speech from each of the four categories were selected as the training set, the remaining 10% as the validation set and the last 10% as the test set for training the model. The distribution of each emotion in each set in experiment is shown in the Tab.6.

	Training Set	Validation Set	Testing Set	Total
Sad	1012	127	126	1265
Pleasure	2064	258	258	2580
Angry	1243	155	156	1554
Joy	3712	464	464	4640
Total	8031	1004	1004	10039

Table 6: The number of data point in different sets

6.3 Hyperparameter Setting

For the training of all networks, including the baseline and others, the hyperparameters setting is shown in the Tab.7

Batch Size	50
Base Learning Rate	0.001
Solver(Optimizer)	Adam
Activation Function	Softmax
Dropout	0.5
Number of Data Points in Each Class	2500

Table 7: The Hyperparameters Setting

6.4 Experimental Set-up

For the testing process of the whole model, first, the test data is input into the $Model_1$ for coarse recognition, the result is recorded as Y_a. Then, determine whether Y_a is in the high similarity set S. If Y_a is Sad or Angry, then input the data into $Model_2$ to get the final result. If not, just output the Y_a.

6.5 Initial Experimental Results

In the previous work [21], the author only obtained 46% overall accuracy. In my implementation, the result is 51.3%. Although we all used the IEMOCAP dataset, they obtained 8029 data points, however in my work I got 10039 data points, and the data points in each categories are more evenly distributed. The confusion matrix is shown in the Fig.15 below. The result was much better than before, because in the previous, the accuracy of anger or sad is just about 2%. However in this work, the result is more even. The accuracy of each class is shown in the Tab.9.

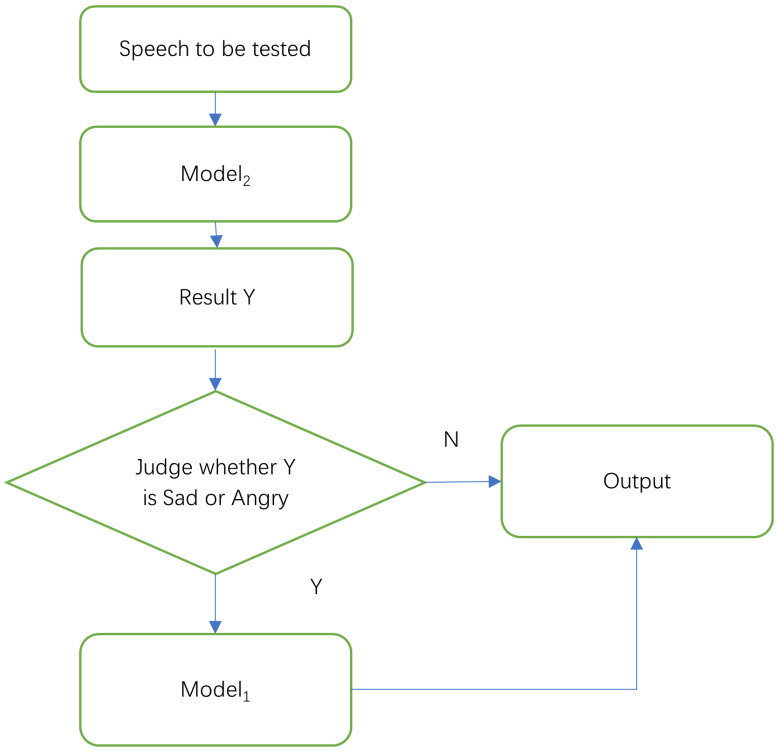

Figure 14: The testing flow chart

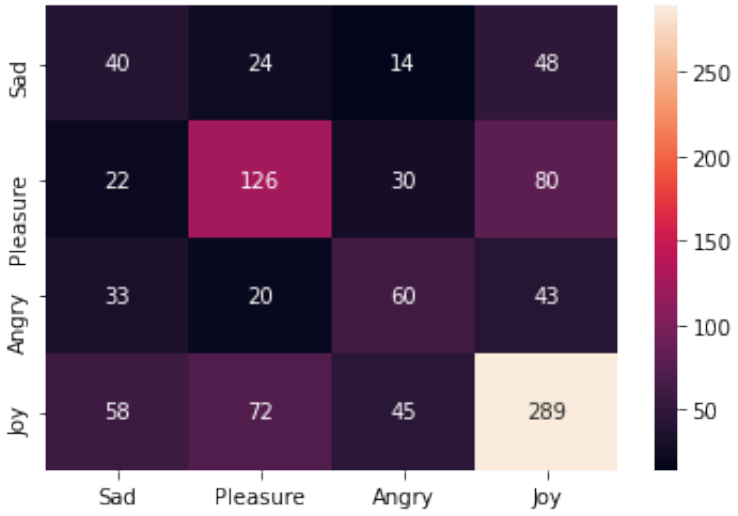

Figure 15: The confusion matrix of baseline

6.6 Experimental Results: Baseline with Data Augmentation

The result shows that the number of data points directly affects the accuracy rate. Therefore, the data augmentation method could be used to make the number of data points the same for each category. Here for each class, I used 2500 data points. Therefore for Pleasure and

Classes	Accuracy
Sad	31.7%
Pleasure	48.8%
Angry	38.7%
Joy	62.3%
Overall	51.3%

Table 8: The recognition accuracy for each categories

Joy, just randomly select 2500 data points, and for Sad and Angry, using VTLP method to makes it 2500 data points. There for, the new dataset has 10000 data points, and 8000 in the training set, 1000 in the validation set and 1000 in the test set. Then feed the new dataset into then the result on test set is shown in Fig.16. The overall accuracy is 53.4%, a little better than the result before. But most importantly, the accuracy of Sad and Angry have improved significantly. The accuracy is shown in Tab

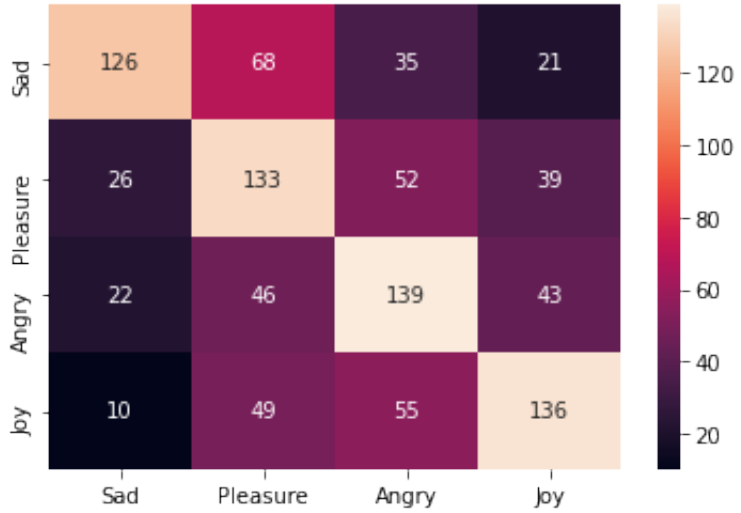

Figure 16: The confusion matrix of baseline on augmented dataset

Classes	Accuracy
Sad	50.4%
Pleasure	53.2%
Angry	55.6%
Joy	54.4%
Overall	53.4%

Table 9: The recognition accuracy for each categories

6.7 Experimental Results: TF-CNN

Then, I used the TF-CNN to extract the features instead of MFCC. To compare with the result before, the same feature vector size should be set. Therefore, the output of the TF-CNN should

be $1 * 13$. Then, feeding the new features to the same model in the baseline. Here I used the augmented dataset. The confusion matrix is shown in Fig.17. And the accuracy table is in Tab.10. The final average accuracy is 55.5%, therefore the feature extraction method based on the CNN is a little better than traditional MFCC.

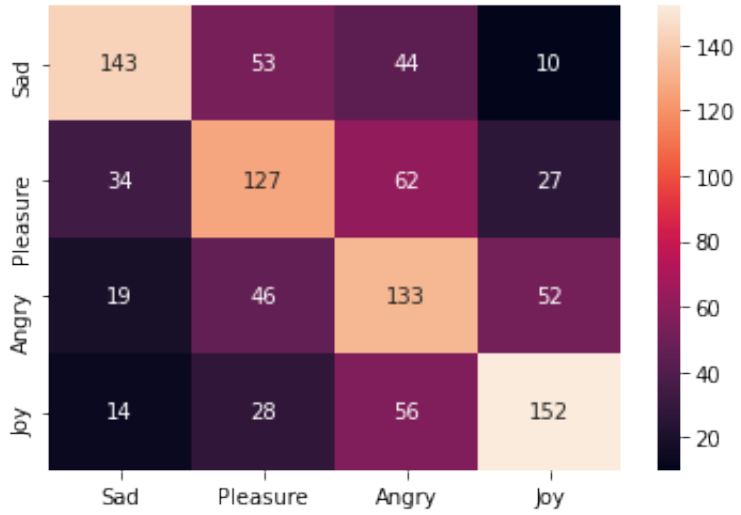

Figure 17: The confusion matrix of baseline by using features from TF-CNN

Classes	Accuracy
Sad	57.2%
Pleasure	50.8%
Angry	53.2%
Joy	60.8%
Overall	55.5%

Table 10: The recognition accuracy for each categories

6.8 Experimental Results: Multi-level SER

First, using the algorithm in Section 4.3 to calculate the similarity between each emotion. The result is shown in the Tab.11.

The result shows that Sad and Angry are the most similar. The value in the table means the

	Sad	Pleasure	Angry	Joy
Sad				
Pleasure	23			
Angry	12	25		
Joy	28	17	20	

Table 11: The similarity between different emotions

hamming distance, therefore smaller value indicates greater similarity. Therefore, the sad and angry will be added to the high similarity set. The flow chart of training is shown in Fig.**??**

and testing flow char is shown in Fig.14.

For the training process, because the high similarity set is $S = Sad, Angry$, first, Sad and Angry are grouped into a new class and labelled as HN class. Then,feed the data points in S to the double-LSTM model to get the $Model_1$. Meanwhile, feed all data points into the TF-CNN model to get the $Model_2$. Here the trained model previously could be used directly.

For the testing process, first, the test data is input into the $Model_1$ for coarse recognition, the result is recorded as Y_a. Then, determine whether Y_a is in the high similarity set S. If Y_a is Sad or Angry, then input the data into $Model_2$ to get the final result. If not, just output the Y_a. The result is shown in the Fig.18 and Tab.12. And the overall accuracy is 60.3%, which is the best one in this paper. The accuracy of Sad and Angry increased by 4.4% and 6.4%, and the overall accuracy increased by 4.8%. The result proved that this framework could be a solution to the problem of high similarity of emotions.

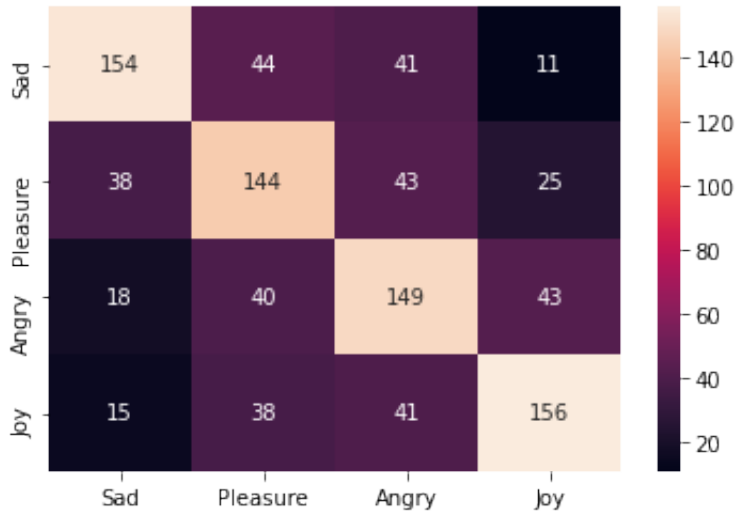

Figure 18: The confusion matrix

Classes	Accuracy
Sad	61.6%
Pleasure	57.6%
Angry	59.6%
Joy	62.4%
Overall	60.3%

Table 12: The recognition accuracy for each categories

6.9 Results and Discussion

The best result in this paper outperforms the baseline[21] by about 10%, especially in some emotion classes. However, the model in the baseline was designed for language-independent SER, but in his work, the "language-independent" just illustrated as using two corpora in

different languages (English and French), which means that his work did not use any special characteristic from English and French. In addition, his model is just a normal CNN, and this model could even be used for image recognition. And in my work, I used the same emotion model, and solve the uneven accuracy between different classes. If I got another corpus in a different language, it is also possible to train a language-independent SER model by using the framework in this paper. Finally, I just compare the result from the mono-lingual process in the baseline, which means that using just one language to train the model and test on the same language. Therefore, the result from my work could be compared with the result in the baseline.

As to the state-of-the-art SER model[35], although my accuracy has a large gap from his

Classes	Exp.1	Exp.2	Tf-CNN	Multi-level SER
Sad	31.7%	50.4%	57.2%	61.6%
Pleasure	48.8%	53.2%	50.8%	57.6%
Angry	38.7%	55.6%	53.2%	59.6%
Joy	62.3%	54.4%	60.8%	62.4%
Overall	51.3%	53.4%	55.5%	60.3%

Table 13: The recognition accuracy for each categories

work, we both trained an LSTM model and a CNN model, to work on the sentence level and frame level, respectively. Because each sentence is assigned a label. However, emotions are varying momentarily in a sentence, therefore we should also work on the frame level. Therefore, the CNN-LSTM based models are very popular in recent years. However, in[35], these two models are independent with each other, they can work well alone. What is more, he used attention mechanisms to pay more attention weight to the consonant-vowel boundaries, and in my work, these two models should work together.

7 Conclusion

This paper is research of the feature extraction and recognition phases of speech emotion recognition. The speech emotion recognition system is mainly used for the automatic recognition of random speech segments. The first one is the speech input module, which selects the speech to be recognized in the file for loading and displays the speech amplitude graph to show the general trend of the speech; the second one is the speech pre-processing module, which pre-processes the speech to be recognized, including pre-emphasis, frame addition and endpoint detection; after that, the speech is converted to image form This module converts the speech into a speech spectrogram and a Meier spectrogram using a short-time Fourier variation and a nonlinear Meier scale filter set, which are further processed as input to the recognition network. The most important module is the recognition model module.

Traditional speech emotion recognition requires manual extraction of features, and the parameters used vary from database to database, making the process complex and inefficient. In this paper, I implemented the FT-CNN model, which considers the expression of speech in both time and frequency domains and uses the horizontal and vertical convolution kernels for multi-scale feature extraction. The experimental validation shows that the effectiveness of TF-CNN network.

Due to the complexity of speech expression, some emotions are easily confused and highly similar, this thesis implemented a multi-level speech emotion recognition framework based on double LSTM, which uses a hash algorithm to calculate the hash value of images and further obtains the category hash value and similarity. The Double-LSTM network structure is proposed for emotions with high similarity, which not only uses a convolutional neural network for feature extraction of speech spectral map in the first stage but also uses LSTM network for modelling inter-speech temporality. In addition, the training set is augmented with a multi-sample sampling rate to allow the network to learn more fully and to further improve the recognition accuracy.

7.1 Future work

Emotion recognition is not only able to obtain information through speech but also exists in various other human expressions such as facial expressions, text, brain waves, etc. We can consider the integration of different modal information The overall accuracy can be improved by processing different modal information and then fusing them with certain strategies. In addition, in this paper, experimental studies have been carried out on the English database only, because I only obtained the IEMOCAP. But the ultimate problem is to achieve final recognition in multiple languages, for example, Chinese, German, and Japanese. If it is possible, make a model for cross-lingual SER would be more useful.

References

[1] Young, P. T., 1943. *Emotion in man and animal; its nature and relation to attitude and motive.* J.Wiley & Sons, Incorporated.

[2] Tokuno, S., Tsumatori, G., Shono, S., Takei, E., Yamamoto, T., Suzuki, G., Mituyoshi, S., and Shimura, M., 2011. "Usage of emotion recognition in military health care". In 2011 Defense Science Research Conference and Expo (DSR), IEEE, pp. 1–5.

[3] Hoy, M. B., 2018. "Alexa, siri, cortana, and more: an introduction to voice assistants". *Medical reference services quarterly,* **37**(1), pp. 81–88.

[4] Tao, F., Liu, G., and Zhao, Q., 2018. "An ensemble framework of voice-based emotion recognition system for films and tv programs". In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 6209–6213.

[5] Li, P., Song, Y., McLoughlin, I. V., Guo, W., and Dai, L.-R., 2018. "An attention pooling based representation learning method for speech emotion recognition". International Speech Communication Association.

[6] Satt, A., Rozenberg, S., and Hoory, R., 2017. "Efficient emotion recognition from speech using deep learning on spectrograms.". In Interspeech, pp. 1089–1093.

[7] Luo, D., Zou, Y., and Huang, D., 2018. "Investigation on joint representation learning for robust feature extraction in speech emotion recognition.". In Interspeech, pp. 152–156.

[8] Parthasarathy, S., and Tashev, I., 2018. "Convolutional neural network techniques for speech emotion recognition". In 2018 16th international workshop on acoustic signal enhancement (IWAENC), IEEE, pp. 121–125.

[9] Latif, S., Rana, R., Younis, S., Qadir, J., and Epps, J., 2018. "Cross corpus speech emotion classification-an effective transfer learning technique". *arXiv preprint arXiv:1801.06353*.

[10] Schröder, M., 2004. "Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions". In Tutorial and research workshop on affective dialogue systems, Springer, pp. 209–220.

[11] Busso, C., Bulut, M., Lee, C.-c., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S., 2008. "Iemocap: interactive emotional dyadic motion capture database," language resources and evaluation". Citeseer.

[12] Jaitly, N., and Hinton, G. E., 2013. "Vocal tract length perturbation (vtlp) improves speech recognition". In Proc. ICML Workshop on Deep Learning for Audio, Speech and Language, Vol. 117.

[13] Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V., 2019. "Specaugment: A simple data augmentation method for automatic speech recognition". *arXiv arXiv:1904.08779*.

[14] Ko, T., Peddinti, V., Povey, D., and Khudanpur, S., 2015. "Audio augmentation for speech recognition". In Sixteenth Annual Conference of the International Speech Communication Association.

[15] Lee, L., and Rose, R., 1998. "A frequency warping approach to speaker normalization". *IEEE Transactions on speech and audio processing,* **6**(1), pp. 49–60.

[16] Jiang, D., Li, W., Cao, M., Zhang, R., Zou, W., Han, K., and Li, X., 2020. "Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning". *arXiv preprint arXiv:2010.13991*.

[17] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G., 2020. "A simple framework for contrastive learning of visual representations". In International conference on machine learning, PMLR, pp. 1597–1607.

[18] Kharitonov, E., Rivière, M., Synnaeve, G., Wolf, L., Mazaré, P.-E., Douze, M., and Dupoux, E., 2021. "Data augmenting contrastive learning of speech representations in the time domain". In 2021 IEEE Spoken Language Technology Workshop (SLT), IEEE, pp. 215–222.

[19] Zhang, Y., Du, J., Wang, Z., Zhang, J., and Tu, Y., 2018. "Attention based fully convolutional network for speech emotion recognition". In 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, pp. 1771–1775.

[20] Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al., 2015. "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing". *IEEE transactions on affective computing,* **7**(2), pp. 190–202.

[21] Strohm, F., 2018. "Language-independent emotion recognition from speech: Performance of activation function".

[22] El-Yazeed, M. A., El Gamal, M., and El Ayadi, M., 2004. "On the determination of optimal model order for gmm-based text-independent speaker identification". *EURASIP Journal on Advances in Signal Processing,* **2004**(8), pp. 1–10.

[23] Pierre-Yves, O., 2003. "The production and recognition of emotions in speech: features and algorithms". *International Journal of Human-Computer Studies,* **59**(1-2), pp. 157–183.

[24] Huang, C., Gong, W., Fu, W., and Feng, D., 2014. "A research of speech emotion recognition based on deep belief network and svm". *Mathematical Problems in Engineering,* **2014**.

[25] Schuller, B., Rigoll, G., and Lang, M., 2004. "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture". In 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, IEEE, pp. I–577.

[26] Chen, S., Jin, Q., Li, X., Yang, G., and Xu, J., 2014. "Speech emotion classification using acoustic features". In The 9th International Symposium on Chinese Spoken Language Processing, IEEE, pp. 579–583.

[27] Zheng, S., Du, J., Zhou, H., Bai, X., Lee, C.-H., and Li, S., 2021. "Speech emotion recognition based on acoustic segment model". In 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), IEEE, pp. 1–5.

[28] Paraskevopoulos, G., Tzinis, E., Ellinas, N., Giannakopoulos, T., and Potamianos, A., 2019. "Unsupervised low-rank representations for speech emotion recognition". pp. 939–943.

[29] Long, J., Shelhamer, E., and Darrell, T., 2015. "Fully convolutional networks for semantic segmentation". In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

[30] Chen, B., Yin, Q., and Guo, P., 2014. "A study of deep belief network based chinese speech emotion recognition". In 2014 Tenth International Conference on Computational Intelligence and Security, IEEE, pp. 180–184.

[31] Fayek, H. M., 2016. Speech processing for machine learning: Filter banks, mel-frequency cepstral coefficients (mfccs) and what's in-between.

[32] Räsänen, O. J., 2007. "Speech segmentation and clustering methods for a new speech recognition architecture". *helsinki university of technology*.

[33] Hochreiter, S., and Schmidhuber, J., 1997. "Long short-term memory". *Neural computation,* $\mathbf{9}(8)$, pp. 1735–1780.

[34] Zhou, J., and Xu, W., 2015. "End-to-end learning of semantic role labeling using recurrent neural networks". In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1127–1137.

[35] Jalal, M. A., Milner, R., and Hain, T., 2020. "Empirical interpretation of speech emotion perception with attention based model for speech emotion recognition.". In INTERSPEECH, pp. 4113–4117.

[36] Jiang, F., Guan, Z., Wang, X., Li, Z., Tan, R., and Qiu, C., 2021. "Study on prediction of compression performance of composite laminates after impact based on convolutional neural networks". *Applied Composite Materials*, pp. 1–21.

[37] Kipp, M., 2001. "Anvil-a generic annotation tool for multimodal dialogue". In Seventh European Conference on Speech Communication and Technology.

[38] Bălan, O., Moise, G., Petrescu, L., Moldoveanu, A., Leordeanu, M., and Moldoveanu, F., 2020. "Emotion classification based on biophysical signals and machine learning techniques". *Symmetry,* $\mathbf{12}(1)$, p. 21.

[39] Neumann, M., et al., 2018. "Cross-lingual and multilingual speech emotion recognition on english and french". In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 5769–5773.