Brunel
University
London

DOCTORAL THESIS

Music Emotion Recognition based on Feature Combination, Deep Learning and Chord Detection

A Thesis submitted to Brunel University London
in accordance with the requirements
for award of the degree of Doctor of Philosophy

in

the department of "Department of Electronic and Computer Engineering"

Fan Zhang

May 15, 2019

Declaration of Authorship

I, Fan Zhang, declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: .... DATE: ....

(Signature of student)

Abstract

As one of the most classic human inventions, music appeared in many artworks, such as songs, movies and theatres. It can be seen as another language, used to express the authors thoughts and emotion. In many cases, music can express the meaning and emotion emerged which is the authors hope and the audience feeling. However, the emotions which appear during human enjoying the music is complex and difficult to precisely explain. Therefore, Music Emotion Recognition (MER) is an interesting research topic in artificial intelligence field for recognising the emotions from the music. The recognition methods and tools for the music signals are growing fast recently. With recent development of the signal processing, machine learning and algorithm optimization, the recognition accuracy is approaching perfection. In this thesis, the research is focused on three different significant parts of MER, that are features, learning methods and music emotion theory, to explain and illustrate how to effectively build MER systems.

Firstly, an automatic MER system for classing 4 emotions was proposed where OpenSMILE is used for feature extraction and IS09 feature was selected. After the combination with STAT statistic features, Random Forest classifier produced the best performance than previous systems. It shows that this approach of feature selection and machine learning can indeed improve the accuracy of MER by at least 3.5% from other combinations under suitable parameter setting and the performance of system was improved by new features combination by IS09 and STAT reaching 83.8% accuracy.

Secondly, another MER system for 4 emotions was proposed based on the dynamic property of music signals where the features are extracted from segments of music signals instead of the whole recording in APM database. Then Long Shot-Term Memory (LSTM) deep learning model was used for classification. The model can use the dynamic continuous information between the different time frame segments for more effective emotion recognition. However, the final performance just achieved 65.7% which was not as good as expected. The reason might be that the database is not suitable to the LSTM as the initial thoughts. The information between the segments might be not good enough to improve the performance of recognition in comparison with the traditional methods. The complex deep learning method do not suitable for every database was proved by the conclusion,which shown that the LSTM dynamic deep learning method did not work well in this continuous database.

Finally, it was targeted to recognise the emotion by the identification of chord inside as these chords have particular emotion information inside stated in previous theoretical work. The research starts by building a new chord database that uses the Adobe audition to extract the chord clip from the piano chord teaching audio. Then the FFT features based on the 1000 points sampling pre-process data and STAT features were extracted for the selected samples from the database. After the calculation and comparison using Euclidean distance and correlation, the results shown the STAT features work well in most of chords except the Augmented chord. The new approach of recognise 6 emotions from the music was first time used in this research and approached 75% accuracy of chord identification.

In summary, the research proposed new MER methods through the three different approaches. Some of them achieved good recognition performance and some of them will have more broad application prospects

Acknowledgements

Firstly, Id like to thank my principal supervisor, Dr. Hongying Meng. Dr Meng was my supervisor for my final sassy when I working for my MSc degree. He is very kind and patient to teach and guide me on my research. After I finish my MSc course, I found I was very fascinated by the area I researched on. So I started my PhD research follow the guild of Dr. Meng. Although I got the weak basic knowledge and understanding in the emotion recognition field. Dr. Meng patiently guild me from the basis of the audio emotion recognition in artificial intelligence area at the beginning to the specific suitable research direction at last. At the same time, he gave me a lot of help and encourage on my work for improvement. I definitely cannot finish this without his contribution.

Then I will thank my second supervisor Dr Nikolaos Boulgouris. On my way to the finish of the PhD career, Dr Boulgouris suggested me a lot of advise. He also help my with his rich experience when I feel confused or get in wrong way.

And I also have to thank my research development advisor Professor Maozhen Li. Prof Li is very kind and happy to help whenever I need help. He always concerned about my research progress when we meet in every corner of campus. And he help me a lot on my paper writing and research direction chosen.I am really respect for there three supervisors who spend their much vigour on me.

And following I will thanks for my two friends: Jingxin Liu and Rui Qin. As the person who start the PhD latest, I received a lot of help of these two “brothers”. They gave me a large mount of suggestion during four years. And they also help me on my research and learning so many times. Nobody can work alone without friends.

At last, I need to thank my family. They are who gave me the chance and encouraged me to start my PhD career.And they are always the fast one who gives a helping hand when I really need one. I love you.

Contents

1 Introduction **4**

1.1 Background 4

1.2 Motivation 7

1.3 Research Challenges 7

1.4 Aim and Objectives 9

1.4.1 Aim 9

1.4.2 Objectives 9

1.5 Thesis Contribution 10

1.6 Thesis Overview 11

2 Literature Review **13**

2.1 Sound and Perception 13

2.1.1 Sound 13

2.1.2 Speech 14

2.1.3 Music 15

2.1.4 Audio 16

2.1.5 Perception 16

2.2 Emotion 19

2.2.1 Discrete Emotion 19

2.2.2 Dimensional Emotion 20

2.2.3 Emotion Recognition 21

2.2.4 Music Emotion Recognition (MER) 23

2.3 Features Extracted for Emotion Recognition 24

2.3.1 Fundamental Frequency 25

2.3.2 Spectrum 25

2.3.3 Zero-crossing Rate (ZCR) 26

2.3.4 Mel-frequency Cepstral Coefficients (MFCCs) 27

2.4 Machine Learning Methods 29

2.4.1 k-Nearest Neighbors (kNNs) 30

2.4.2 Random Forest (RF) 31

2.4.3 Naïve Bayes (NB) 33

2.5 Deep Learning Methods 35

2.5.1 Convolutional Neural Network (CNN) 35

2.5.2 Recurrent Neural Network (RNN) 37

2.5.3 Long Short-Term Memory (LSTM) 39

2.6 Music Emotion Recognition (MER) Application 39

2.7 Summary 40

3 MER Based on Feature Combination and Random Forest Classification 42

3.1 Introduction 42

3.2 Related Work 43

3.3 Emotion Recognition System 46

3.3.1 System Overview 46

3.3.2 IS09 Audio Feature Extraction 46

3.3.3 The Global Statistics Feature Inspired by EEG Signal Analysis . . 48

3.3.4 Classification 50

3.4 Experiment Results 51

3.4.1 APM Database 52

3.4.2 Experimental Setting 52

3.4.3 Performance Measurement 53

3.5 Summary 57

4 Emotion Detection Based on Dynamic Deep Learning Models 59

4.1 Introduction 59

4.2 Related Works 60

4.3 LSTM Based Emotion Detection System 62

4.3.1 System Overview 63

4.3.2 Feature Extraction 63

4.3.3 Long Shot-Term Memory (LSTM) Modelling 65

4.3.4 Parameter Setting 72

4.4 Experiment 74

4.4.1 Data Preparation 74

4.4.2 Experimental Results 75

4.4.3 Comparison and Discussion 78

4.5 Summary 79

5 Music Emotion Identification Through Chords 81

5.1 Introduction 81

5.2 Emotion of the Chords 83

5.2.1 C Minor Chord . . . 84
5.2.2 Minor Sixth Chord . . . 85
5.2.3 Natural Minor Chord . . . 86
5.2.4 Neapolitan Sixth Chord . . . 86
5.2.5 Subdominant Chord . . . 87
5.2.6 Augmented Chord . . . 89
5.3 Chord Identification . . . 90
5.3.1 Data Preparation . . . 90
5.3.2 Chord Recognition Method . . . 94
5.4 Experiments . . . 96
5.5 Summary . . . 106

6 Conclusion and Future Works 109
6.1 Conclusion . . . 109
6.1.1 Feature Combination and RF Classification . . . 110
6.1.2 Dynamic Deep Learning Models . . . 110
6.1.3 Music Emotion Identification Through Chords . . . 111
6.2 Future Works . . . 112

List of Figures

2.1 The Sound [1]. 14

2.2 The speech [2]. 15

2.3 Anatomy of the ear [3]. 17

2.4 Wheel of emotions [4]. 20

2.5 Multi-dimensional emotion analysis [5]. 21

2.6 Emotion representation by A-V [6]. 22

2.7 The signal and ZCR [7]. (a. the original audio signal; b. ZCR of the signal) 26

2.8 The example of kNNs working [8]. 31

2.9 RF classifier in which several decision trees are selected and combined together for the classification automatically. 32

2.10 The CNN structure [9]. 35

2.11 The RNN structure [10]. 37

3.1 The overview of the MER system. (a). original signal, (b). left channel, (c). right channel, (d). statistics features, (e). typical audio feature, (f). two channel and all features combination, (g). random forest classifier) (h).emotion prediction. 47

3.2 Five common audio features selected from IS09. 48

3.3 Statistic Feature. 49

3.4 Four audio signals for 4 different emotions.(take continuous 45,000 sampling points from the music recording for approximately one second). . . 52

3.5 The performance of RF with different tree numbers setting. 55

4.1 Time and frequency domain[11]. 60

4.2 LSTM based emotion detection system from music signals. (a) Music signal, (b) Feature selection, (c) LSTM model, (d) Emotion classification. 63

4.3 IS09 features extracted from APM database. 64

4.4 Statistic Feature for all the 400 music sample in APM database. 65

4.5 LSTM model. 66

4.6 LSTM hidden unit. 67

4.7 Normal RNN structure (The A block means the same structure unit in different time sequence)[12]. 68

4.8 LSTM structure[12]. 68

4.9 Icon explanation[12]. 69

4.10 The cell statement in the LSTM[12]. 69

4.11 The sigmoid layer[12]. 70

4.12 The forget gate[12]. 70

4.13 The input gate. 71

4.14 The update of cell state [12]. 72

4.15 The output gate [12]. 73

4.16 Four audio signals for 4 different emotions. (take continuous 44, 100 sampling points from the music recording for approximately one second). 75

5.1 Music and emotions [13]. 82

5.2 C minor chord. 85

5.3 Minor sixth chord. 86

5.4 Natural minor chord. 87

5.5 Neapolitan sixth chord. 88

5.6 Subdominant chord. 88

5.7 Augmented chord. 89

5.8 Chord identification system (a.The chord database; b.The FFT features; c.The STAT features; d.The identified chords). 91

5.9 Chord database build by piano music signals(a. Original music signals; b. Chord split; c. The selected chord class). 91

5.10 Six kinds of chords and their corresponding emotions. 92

5.11 Adobe Adudition work space. 93

5.12 Chord extraction. 94

5.13 Six kinds of selected chords. 95

5.14 The comparison figure between True FFT and Sample FFT of Augmented Chord.(a. benchmark Augmented chord; b. sample Augmented chord) . . 97

5.15 The comparison figure between True FFT and Sample FFT of Minor sixth Chord.(a. benchmark Minor sixth chord; b. sample Minor sixth chord) . . 97

5.16 The comparison figure between True FFT and Sample FFT of C Minor Chord.(a. benchmark C Minor chord; b. sample C Minor chord) 98

5.17 The comparison figure between True FFT and Sample FFT of Natural Minor Chord.(a. benchmark Natural Minor chord; b. sample Natural Minor chord) 98

5.18 The comparison figure between True FFT and Sample FFT of Neapolitan Chord.(a. benchmark Neapolitan chord; b. sample Neapolitan chord) . . . 99

5.19 The comparison figure between True FFT and Sample FFT of Subdominant Chord.(a. benchmark Subdominant chord; b. sample Subdominant chord) 99
5.20 Emotion classification based on the chord. (a. Original music signals; b. The identified chords; c. The Theory of Musical Equilibration; d. The emotions [14]) 107

List of Tables

3.1 Accuracy of individual features and Combined features using three different machine learning methods. 54

3.2 Comparison between IS09 Combination features and standard openSMILE features in APM database and DEAM database with RF method. . 55

3.3 The results of using RF in different tree numbers setting. 56

3.4 Comparison with existing results on APM database. 56

4.1 The accuracy(%) comparison between different length segments running in different learning methods. 77

4.2 The accuracy(%) between different features and length segments on default setting running in LSTM. 77

4.3 The accuracy(%) test on epoch parameter. 78

4.4 The accuracy(%) test on output parameter. 78

4.5 The accuracy(%) test on mini batch size parameter. 78

4.6 The accuracy(%) test on shuffle parameter. 79

4.7 The final accuracy(%) with parameters. 79

4.8 The accuracy(%) comparison with other works in same database. 79

5.1 The emotions in the chords. 90

5.2 Chords database detail. 92

5.3 The Euclidean distance comparison test with FFT features in Chord database (The meaning of abbreviation : Aug - Augmented chord, m - Minor chords, m6 - Minor 6th chord, Sub-d - Major subdominant chord). 99

5.4 FFT Euclidean distance confusion matrix. 101

5.5 The Euclidean distance comparison test with STAT features in Chord database (The meaning of abbreviation : Aug - Augmented chord, m - Minor chords, m6 - Minor 6th chord, Sub-d - Major subdominant chord). 102

5.6 STAT Euclidean distance confusion matrix. 104

5.7 The Correlation by Pearson correlation coefficient comparison test with STAT features in Chord database (The meaning of abbreviation : Aug - Augmented chord, m - Minor chords, m6 - Minor 6th chord, Sub-d - Major subdominant chord). 104
5.8 STAT correlation confusion matrix. 106
5.9 The identification accuracy of both STAT and FFT chord feature. 106

List of Abbreviation

A-V	Arousal-Valence
ACF	Aggregate Channel Feature
AI	Artificial Intelligence
ANCW	Affective Norms for Chinese Words
ANEW	Affective Norms for English Words
ANN	Artificial Neural Network
ASR	Average Silence Ratio
BoW	Bag-of-Audio-Words
BPNN	Back Propagation Neural Network
BPTT	Backpropagation Through Time
CCC	Concordance Correlation Coefficient
CNN	Convolutional Neural Network
CQT	Constant Q transform
DBLSTM	Deep Bidirectional LSTM
DCT	Discrete Cosine Transform
DFT	Discrete Fourier transform
DWCH	Daubechies wavelets coefficient histogram
EEG	Electroencephalogram
FD	Fractal Dimension
FFT	Fast Fourier transform
FkNNs	Fuzzy kNNs classifier
FV	Fisher Vector Encoding
GMM	Gaussian Mixture Model
GPR	Gaussian Process Regression
GSV	GMM Super Vector
HCI	Human-computer Interaction
HMMs	Hidden Markov Models
HOC	Higher Order Crossing
IEMOCAP	Interactive Emotional Dyadic Motion Capture
kNNs	k-Nearest Neighbours
LLD	Low-Level Descriptors
LSTM	Long Shot-Term Memory
MCA	Multi-scale Contextbased Attention
MER	Music Emotion Recognition
MEVD	Music Emotion Variation Detection
MFCCs	Mel-Frequency Cepstral Coefficients

MLR	Multiple Linear Regression
MSFs	Modulation Spectral Features
NB	Naïve Bayes
NN	Neural Network
PAD	Pleasure (Valence), Arousal and Dominance
PCA	Principal Component Analysis
PCP	Pitch Class Profile
PSD	Power Spectral Density
RBMs	Restricted Boltzmann Machines
RBF	Radial Basis Function
RF	Random Forest
RMS	Root Mean Square
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SCF	Spectral Crest Factor
SER	Speech Emotion Recognition
SFM	Spectral Flatness Measure
SVM	Support Vector Machine
SVR	Support Vector Regression
VAD	Voice Activity Detection
VP	Voiceprob
ZCR	Zero-Crossing Rate

List of Publications

1. Liu, J., Meng, H., Li, M., **Zhang, F.**, Qin, R. and Nandi, A.K., 2018. Emotion detection from EEG recordings based on supervised and unsupervised dimension reduction. Concurrency and Computation: Practice and Experience, p.e4446.

2. Zhang, W., **Zhang, F.**, Chen, W., Jiang, Y. and Song, D., 2018. Fault State Recognition of Rolling Bearing Based Fully Convolutional Network. Computing in Science Engineering, 1: pp. 1-1.

3. Jan, A., Meng, H., Gaus, Y.F.B.A. and **Zhang, F.**, 2018, Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. IEEE Transactions on Cognitive and Developmental Systems, 10(3), PP.668-680.

4. Jan, A., Gaus, Y.F.B.A., Meng, H. and **Zhang, F.**, 2016. BUL in MediaEval 2016 Emotional Impact of Movies Task. MediaEval 2016 Workshop. Available in "http://slim-sig.irisa.fr/me16proc/MediaEval_2016_paper_43.pdf".

5. **Zhang, F.**, Meng, H. and Li, M., 2016, August. Emotion extraction and recognition from music. In Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016 12th International Conference on . IEEE, pp. 1728-1733.

6. Gaus, Y.F.B.A, Olugbade, T., Jan, A., Qin, R. Liu, J. **Zhang, F.**, Meng, H., Bianchi-Berthouze ,N., 2015, Social Touch Gesture Recognition using Random Forest and Boosting on Distinct Feature Sets, Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, New York, NY, USA., pp 399-406.

7. Gaus, Y.F.B.A., Meng, H., Jan, A., **Zhang, F.** and Turabzadeh, S., 2015, May. Automatic affective dimension recognition from naturalistic facial expressions based on wavelet filtering and PLS regression. In Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops, (Vol. 5)., pp. 1-6, IEEE.

8. Jan A., Meng H, Gaus Y.F.B.A, **Zhang F.**, Turabzadeh S., 07 Nov 2014, Automatic Depression Scale Prediction using Facial Expression Dynamics and Regression, Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (AVEC14). ACM, ACM New York, NY, USA. pp. 73-80.

Chapter 1

Introduction

1.1 Background

Music is an art that has a long history and cannot explore its origins. Its origin may have been created by humans mimicking the sounds of nature [15]. As a member of the sounds of nature, music naturally has the same characteristics of all sounds. It is generated by the vibration of the object to generate a regular wave. These waves through the medium is received by the human eardrum and transmitted to the brain. So that people can hear these wonderful sounds [16]. Although the principle of sounding is the same, but different people speak different sounds, different vibration frequencies, different amplitudes, different wavelength combinations, etc. make each sound completely different [17]. When these completely different sounds are combined according to certain rules according to the creators' ideas and ideas, a wonderful piece of music is born [18].

Music appears in any civilisation currently known [19]. It varies with the humanities, history, climate and other factors of the civilization that exists. The history of music development is like the epitome of human development history. From the oldest ancient music to the Baroque music [20], then the Renaissance music [21], the development of music theory from the parallel three [22] to the natural scale, and then the emergence of mature staves and chord techniques, all show the human with the times progress is more abundant and effective, and expresses your own thoughts and emotion through music [23] [24].

Emotion, as a state of reaction to human psychology and physiology [25]. It can produce different reactions with different sensations; thoughts and behaviors, when human beings are subjected to internal or external stimuli. Under normal circumstances, the six most common human emotions (happy, anger, sad, fear, disgust, and surprise) can basically reflect a person's feelings [26]. But when there are some complicated situations, other complex emotions will be also occurred. As one of the most common external stimuli, music can easily affect human emotions [27]. How different concerts affect human emotions has always been a hot research in the field of psychology and music. But the methods for quantify the definition and describe emotions has always been one of the difficulties in this type of research [28].

With the development of technology, the emergence of Artificial Intelligence (AI) [29] has shown outstanding performance in various fields [30]. As the cutting-edge technology, AI has shown initial success in simulating human communication, perception and action [31]. It integrates high-definition disciplines such as bionics, psychology [32], statistics, and advanced mathematics as the core of technology [33], and has begun to emerge in medicine [34] [35], education [36] [37], sociology [38], and construction. Moreover, in some areas where AI can be used maturely, it is surprising that the work efficiency of AI is excellent. The establishment of AI is the principle that simulates the process of human thinking and learning, making it possible for AI to distinguish and understand human emotions in different situations. The process of defining human emotions based on different external manifestations of human beings, such as facial expressions [39] [40], body movements [41], sounds [42] [43], brain waves [44] [45] and skin electrical signals [46], is called emotion recognition [47].

When music, emotion recognition and AI collide, an interesting study was born: Music Emotion Recognition (MER) [48]. This research aims to build a system that can accurately and efficiently identify emotions that are desired to be expressed in music through research in the field of AI. Because the high frequency of new music created, higher quality requirements for music researching by emotion in music market is achieved [28]. Moreover, MER as useful part in AI research and human emotions simulation. So MER has a very high research value [18].

As the emotions in music are discovered by more and more people, more people are beginning to invest in it [49]. However, to accurately define and describe the emotions in music is not easy to judge. Different music may have different emotional expressions at different times, or the same melody may appear differently in different places. Even a person with sensitive thinking can express feelings that are felt in a piece of music. But he may cannot be expressed that accurate [50]. Then, how to make the machines and computers, which do not have the same thoughts as human beings, to express the emotions contained in music accurately has become an important topic of AI in the audio field [51].

Similar to emotion recognition in other fields, the core goal of MER is to use the current data to perform a series of processing and operations, so as to achieve the same purpose of the final predicted emotion and real emotion. To accomplish such a goal, it is necessary to simulate the process of human feelings. When an audience hears a piece of emotional music, the melody of the music, which impresses the audience and touches him, such as memories or imagination [52]. This situation may be the mother's care, let him feel comfortable and relaxed; or the fierce battlefield, let him feel excited; or the death of a loved one, makes him feel sad. The parts of these music that cause emotional changes need to be separated and remembered by machines without thinking. These parts are called features [53]. For training the machine to remember the characteristics which features specifically represent, a model similar to human logic is needed. Just like humans learn to talk. When a parent tells a child something like an apple is called an 'apple', the child knows what the red fruit is called. In the same way, researchers need to tell the machine first, which emotions these features represent. Then through the training of the logical model, let the machine form a mapping of "this kind of music should be such a feeling" [28]. Finally, when the machine hears a new piece of music, it can judge what emotions the music contains through the "knowledge" he learned before.

In this training machine learning how to recognize the emotions in music, the first problem that needs to be dealt with is: how to accurately separate the parts of the music that impress the audience? In the field of AI, the problem is equivalent to how to extract a suitable feature in the music data to describe the sentiment classification. This is also the first important component of the sentiment recognition system: feature extraction and selection [53]. In MER, for effectively and accurately teaching the computer how to distinguish different emotions, it need using appropriate features [54].If the wrong features are selected during training, it will probably compromise the efficiency and accuracy of the recognition.

After completing the feature selection, a model is needed to construct for letting the computer learn the relationship between features and emotions more quickly and effectively. An AI is like a child at the beginning of creation, knowing nothing about what is in front. The task of this model is like a human parent telling a child that the fruit of this red prototype is called apple. When the machine finishes training the model, it will learn that: which feature it has means which kind of emotion is expressed. This process, which is training and educating machines with models, is called machine learning [55]. In the field of AI, a large number of models are invented every day in order to find the best model [56] [57] [58]. As a method of training computers, machine learning has evolved, from the simplest k-Nearest Neighbours (kNNs) to the complex various in-depth learning methods. In so many learning method models [50], which one is the most suitable for musical emotion recognition, no one can assert.

1.2 Motivation

With the rapid development of science and technology, AI technology is becoming more and more mature [59] [60] [61]. Theories in various fields of AI are also gradually improving. More and more researchers have joined the exploration of this new world [62]. As an important part of the human-computer interaction (HCI) of AI, the field of emotion recognition has been the focus of research by researchers. As the research progressed step by step, the researchers found that more unknown parts of the field appeared. Musical emotion recognition also plays its role in the part of this field [63]. As a carrier for human emotions and thoughts, music has great research value in the field of emotion recognition.

Like the popular speech recognition research of previous years, music can be seen as another language of human beings [64]. Studying MER has greatly promoted the field of emotion recognition in HCI. Moreover, music, as an art form that has been integrated into human society [65], has a great influence on the industry where it appears [66]. For example, studying the emotions of the background music of a film and television work will help the audience understand the feelings that the work itself wants to express; or the music containing certain specific emotions may be of great help in treating mental illness [67] [**?**]. However, MER is not perfect in all aspects. There is still more room for improvement from signal processing to recognition model.

In summary, the main motivation of this research is to research the process and method of MER. Then understand and try to improve the current MER method. Make the recognition system perform better and increase the accuracy of recognition.

1.3 Research Challenges

In recent years, MER has been continuously studied as an useful branch of HCI in the field of AI [54] [68]. As an important part of the expression of human emotions, music has always played a very important role in the field of psychology [69] [70]. HCI research is devoted to the study of the interaction between humans and computers [71]. Music has become an bridge to study how to make computers understand and learn emotions. The computer can learn the emotions, which humans will produce, by understanding the emotions that music displays. On the other hand, there are also some demands for MER applications in the real world. For example, some music companies need to quickly identify and classify large quantities of music according to the emotional types of music [28]; some film and television post-production companies need to use post-synthesis to enhance the expressiveness of a certain piece of music with key emotions [72];Even in some medical fields, MER is already trying to participate in the treatment of Alzheimer's disease [73]. With the deepening of MER research, the key points and difficulties in the research are gradually exposed to the researchers' vision.

Feature extraction, as the first part of the MER system, has always been the top priority in research. A great feature can effectively describe and embody the emotions contained in music. An excellent feature needs to accurately reflect the difference between emotions and music. At the same time, it also needs to be easily extracted. This feature can't be too big and verbose. How to extract this feature from so many different styles and different forms of music is also the key to the study of musical emotion recognition. In recent years, a large number of different features have been excavated as the research progressed. Some of their features can be very effective in the identification. Other features are not good enough to affect the performance of the entire system. Therefore, extracting a feature, which can be very effective in expressing emotions from music, is the most important and crucial step in the MER research [28].

The significant impact of machine learning on musical emotion recognition systems follows feature extraction. The machine learning has been developed from the concept of AI has been rapidly developing. From the earliest Hebbian theory learning rules [74] to the simplest pattern recognition algorithm: kNNs, and now the deep learning, the evolution of machine learning algorithm models has never stopped. However, not the most advanced machine learning algorithms are necessarily the best algorithms. For a specific recognition task, only the most appropriate algorithm can be called the best algorithm model. For the MER system, it is also fascinated by the evolution of the model. After the experimental and research of a large number of traditional machine learning methods, can the deep learning model, such as Long Shot-Term Memory (LSTM), further improve the accuracy of the MER system? This is a question that needs to be studied.

Finally, music is a unique expression that has been inherited for thousands of years. Whether it contains some rules of emotional expression is also worthy of research and discussion. Music, like any other language discipline in the world, has its own independent theoretical system. Among these systems, is there a theoretical law that can assist computers in the identification of emotions, such as chords? As an important part of music, chords [75] play a role in giving music a new expressive power in limited temperament changes. Just like the mathematical system that originally had only “natural numbers”, the concept of “decimal” was suddenly added, which changed the whole field of mathematics. Can chords play a role like vocabulary doing in human language, changing the meaning of the whole sentence when the vocabulary difference [76]?

1.4 Aim and Objectives

1.4.1 Aim

The aim of this research is to build a system for MER. The researching for the theory, methods and applications of emotion detection and analysis based on music signals need be done for this aim.

1.4.2 Objectives

For the objective of whole research, it is split to 3 parts of objectives from the basic knowledge to the state-of-the-art methods. These objectives focus to solve the research questions which mentioned previous section.

a. Investigate the Best Feature for MER

At the starting, the basic signals knowledge of the audio has been researched. It started from the principle of the sound making with reviewed the knowledge of the human vocal structure and the musical instruments sound principle. Then the basic signal property has been researched, such as loudness and frequency. Some signal progress skill has been learned. After these learning, for the feature extraction (which is always the key point of this research), the traditional, popular and useful features has been learned and analyzedand also some features combination has been done from the research. At laset, the most suitable features always been found during the research and experiment. Some new features with new method even in different research field has been extracted for the music signals.

b. Explore Advanced Learning Method for MER

For the core part of the emotion recognition, machine learning has been studied. The research began from the most basic part of machine learning methods, to the most popular method in the world. Many of the machine learning methods has been learned, practiced and applied. The most traditional machine learning methods had been reviewed. As the forefront and the most promising machine learning method deep learning will be the working focus. The cutting edge and the most practical deep learning method, such as LSTM and Convolutional Neural Network (CNN) will be leaned and researched for the music part and audio part experiment. After all this deep learning methods study, an effective deep learning methods for MER system should be found and worked in .

c. Investigate MER Through Chord Identification

As the last part of the objectives, the basic music theory of music should be learned and research. Look for the relationship between the difference between chords and musical emotions. Chords are an important part of music that can express musical changes. Many times the difference in chords directly affects the entire musical style and even the difference in feeling. Like words, chords are also “adjectives“ that describe music. In the same way, chords should also have the ability to influence the thoughts that music wants to express; study the relationship between chords and emotions; find out if it can be researched and tried to establish a recognition system between chords and emotions based on the relationship between the two.

1.5 Thesis Contribution

The main contribution of this paper is that it has carried out in-depth and extensive research on MER from three different aspectsand three different systems have been established for musical emotion recognition.

1. Based on data pre-processing, a variety of selected effective feature sets and a musical emotion recognition system for the machine learning method of the music database are proposed. The new feature combination which is propose first time for MER was used in this system. Two kinds of 13 different features selected for MER are extracted, and two types of features are effectively combined. At the same time, a Random Forest (RF) algorithm is used as a machine learning method. Combine the above two elements to complete the construction of the MER system. Later, a large number of parameter tuning experiments were performed on the parameters in the RF machine learning algorithm. Finally, the system makes accurate recognition of music emotions to around 84%.

2. Based on the dynamic nature of the music signal, a large number of different segmented pre-processed music features are extracted. For combining the dynamic deep learning methods,LSTM is chosen as a learning model with dynamic temporal continuity to solve the problem of information before and after communication in continuous data. Because LSTM method can connect the information between previous timing and current timing by training the continuous signal data for predicting the result which need to consider the previous situation. Five parameters with great influence were extensively analyzed and tested. Finally, the combination of features and parameters for the LSTM were obtained and the performance only achieved 65.8% accuracy. The conclusion prove that deep learning method not always the best not approach for MER based on different database.

3. The new approach for recognise the emotions from music is found. According to the chord sentiment theory mentioned in Daniela and Bernd Willimek [77], six different chords are selected as the basis for unique musical emotion recognition. Based on the different characteristics of the six chords, a new chord database was created. Using the chords of the database, Fast Fourier transform (FFT) and STAT statistical features are extracted from the chords for six different chords. Subsequently, the Euclidean distance method and correlation detection were used as the basis for chord identification. Therefore, MER can be performed according to the chord type and chord emotion theory in music.

1.6 Thesis Overview

This thesis has 6 chapters. The first chapter is introduction, which is the chapter been read now. This chapter gives the information about the background of the research area, research question, motivation, aim and objective, contribution and the thesis overview.

The second chapter is the literature review. In this chapter, large amount of the literature about MER field has been read and study. From the most basic knowledge of voice and sound to the top and popular deep learning methods, selected papers have been learned with demand.

In Chapter 3, an automatically musical emotion recognition system based on a new feature combination with two features, IS09 and STAT, was proposed. The two channel music signals pre-process before extract in IS09 and STAT effective features. After the combination features include the traditional and useful low level freatures set IS09 and new statistics features STAT strong joining, the RF machine learning methods used as the core model. Because it very suitable for this kind of classification question, it can cooperate wonderfully with those selected features to predict the classification of emotion.

In Chapter 4, a research focus on the MER system based on LSTM deep learning method is provided. The segmentation music signal is used to extract the featrues in this system. Then the LSTM deep learning method was provide to class them. After many experiment for looking for the best features and parameter combination. The system look forward to use the dynamic continuous information between the different time frame segments for more effective MER.

In Chapter 5, a system of chord identification based on a new built chord database for chord and emotion relationship research is approached. The research starts as building a new chord database which use the Adobe audition to extract the chord clip from the piano chord teaching audio. Then the FFT features based on the 1000 points sampling pre-process data and STAT features were extracted for the selected features from the database. After calculating and comparison the Euclidean distance and correlation, the results shown the STAT features work well in most of chords except the Augmented chord. The STAT features resulting in almost wrong results in Augmented chord. On the other hands, the FFT features provide the best results in Augmented chord, Minor chord and Minor 6th chord. However, the other three chord basically was not work on it. But it can be believe that if combining these two features together, the result will improve.

In the last chapter, the conclusion and the future work have been illustrated. All the works have been summarized there. The results and affect to the MER field have been analyzed and shown and the thesis will talk about the future works for this research, include the think about improvement of the researches and the application of the research using in.

Chapter 2

Literature Review

2.1 Sound and Perception

In the boundless universe, is never lack of sound: the human voice, the voice of the car engine, the voice of the cat and the sound waves. This sound is through the object vibration, and was received by the ears. Now that to understand and analysis of audio signals, understanding of the students in principle is necessary. Among them, the most representative is the principle of human speech and the principle of musical instruments

2.1.1 Sound

The vibration of the vibration source must depend on external power, the power has two sources [78]: 1) Mechanical force is the energy source of physical instrument, depending on the mechanical force vibration source have strings, the head and the soundboard, they emit frequency completely determined by the material properties and the shape of vibration source itself. The role of mechanical force on the source of vibration to be continuous. For example, power of the piano strings only depends on the piano hammers the strings of the moment. 2) Airflow is the energy source of breath instrument, depending on the air vibration source some of the vocal cords, reed (reed woodwind instruments), gas spring (depending on the air around the cavity vibration, no reed woodwind instruments), lip spring (the vibration of the lips, brass wind instrument) and the metal spring (harmonica and accordion) they emit frequency is generally not too rigid.

Source itself has a fixed frequency, known as the natural frequency, but the frequency of

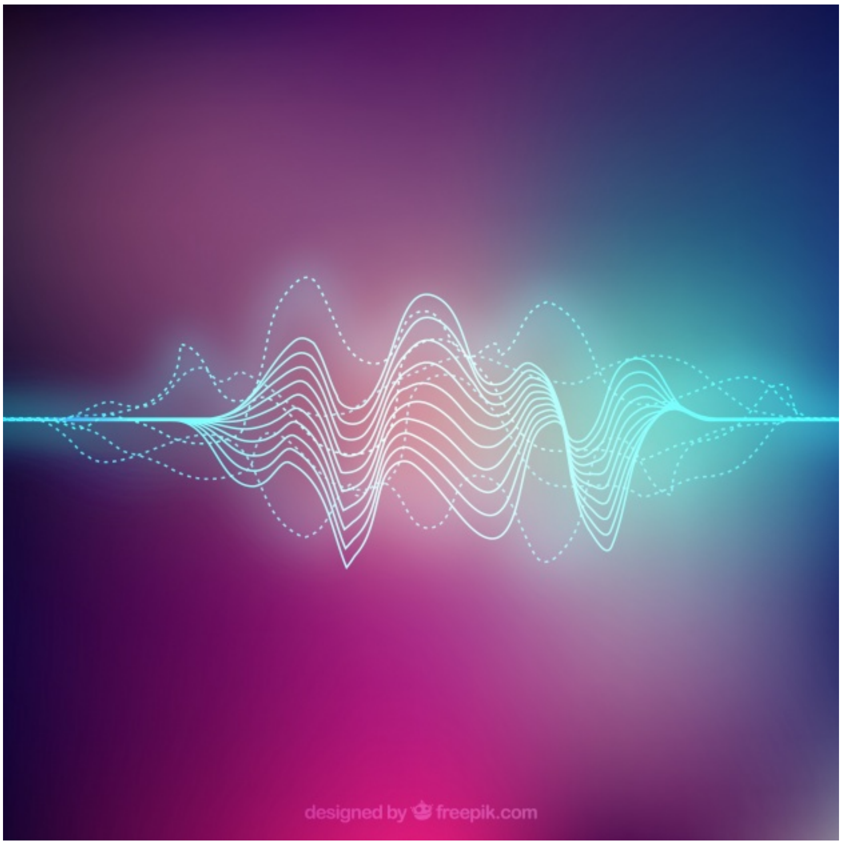


Figure 2.1: The Sound [1].

the outside world, especially the power and the influence of the resonance cavity, such as when the strings hit, frequency will be slightly higher than its natural frequency, the frequency of the reed is completely can the air around.

2.1.2 Speech

Human beings are born to pronunciation, but the vocalism principle is very complex. Human mouth, tongue and lips have great flexibility, allowing us to form accurate sound in a wide range. Chimps cannot do this. Some people than others to develop a more sophisticated sound system [79]. At the moment, a singer can become plump vibrato from his normal voice, let the sounds of nature and beautiful clear voice filled the hall.

It works like this: When people speak or sing, lungs of people exhaled air through the throat or larynx control. Approximately the size of a walnut throat. In men's neck, it is mainly composed of cartilage and muscle. At its top is pulled past the vocal cords, composed by two layers of mucous membranes. Exhaled air from lungs, pushed forward by the throat, so that the vocal cords vibrate, sound.

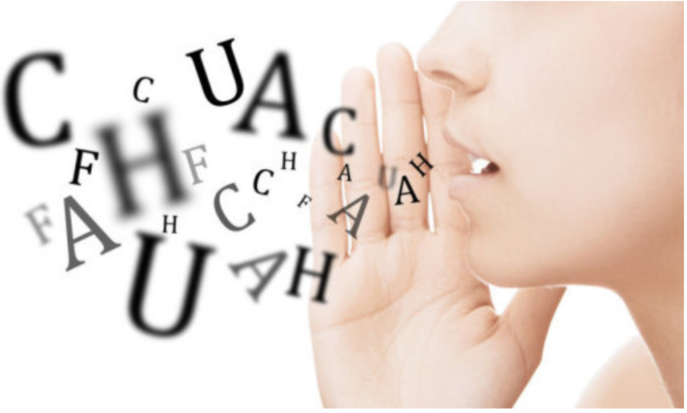


Figure 2.2: The speech [2].

Vibration of vocal cord area quivers frequency determines the pitch. It let people tighten the vocal tone sharpened, so that the voice sounded high-pitched shrill; relax the vocal cords, then the sound muffled. The shape the throat, mouth, tongue and lips of people put these sounds together to form words and sentences, is simply a complex dance. It requires a lot of fine motor control.

2.1.3 Music

In modern times, music is widely used in human society. People can hear music from all over the world, from the common restaurants and bars to the magnificent Opera House and Cathedra [28]. Everyone has a favorite piece of music, because this piece can bring him the most wonderful feeling. A well-behaved piece of music can also bring the audience into the wonderful world that the author hopes to create. Music is endowed with a variety of emotional expressions along with the creator's thoughts, as their work spreads through the ages [80]. These works are often used in a variety of film and television works or an atmosphere [81]. But how to accurately and efficiently recognize the emotions contained in music has always been an important lesson in various music schools [82].

For the musical instruments [83], music as one of common sound in human's life, it almost can cover all the frequency which can be heard by human. So it is another significant source and data collection. So if the instruments be wanted to make some voice, two of the element can be ignorethe source of vibration and the resonance cavity [84].

2.1.4 Audio

The audio signal is a kind of analog signal [85]. But it is needed to put his processing into discrete time signal, so in this step, the voice signal needs to be sampled and quantified [86]. Sampling is the use of Nyquist [87] theory and the highest frequency of sound to determine a sampling frequency and then according to the frequency do sampling. Nyquist theory pointed out that the sampling frequency can't higher than twice the highest frequency sound signal. So it can be put the performance of noise reduction with digital signal to the original sound. This is called a nondestructive digital. If it is not followed the theorem can cause error (aliasing effect). Quantitative is divided into paragraphs, the signal amplitude and discrete continuous range. Time domain and frequency domain is the most common and most effective approachs to illustrate the property of an audio signal. The signals can be analysis by different method with time domain and frequency domain.

2.1.5 Perception

Auditory system by auditory organs at all levels of auditory center and connect to the Internet. Known as ear auditory organs, with special differentiated cells in its structure, can feel the sound wave mechanical vibration and sound energy is converted to a nerve impulse, called acoustic sensors. he ear of the higher animals can be divided into the outer ear, middle ear and inner ear. Outer ear includes the pinna and concha, mainly set sound effect; Ears of some animals can free rotation, easy to “catch” the sound.

Middle ear including tympanic membrane [88], auditory ossicle chain, middle ear, the middle ear muscle, eustachian tube, such as structure, main function sound transmission. Tympanic membrane structure is enclosed in the external auditory canal side a layer of film. Sound waves come from the external auditory canal, acting on the eardrum, which generates a corresponding vibration. Mammalian auditory ossicle chain is made up of three small bones (malleus, incus, stapes.) consisting of lever system, one end for the hammer handle, attached to the tympanic membrane inside, on the other side of the stapes backplane, capping the eggs on the round diaphragm in the inner ear, the eardrum vibration through the lever system can effectively to the inner ear, the eardrum for tympanum, auditory ossicle chain and the middle ear muscle are among them. The middle ear muscle or muscle ears, there are two blocks: the contraction of the eardrum tensor by pull malleus and make the tympanic membrane tension, stapes stapes fixation muscle contraction, its role is limited sound to the transmission of the inner ear. Eustachian tube (eustachian tubes) by middle ear to the pharynx, normally closed, open when swallowed, and some mouth movements, often can make drum indoor air pressure and air pressure balance.

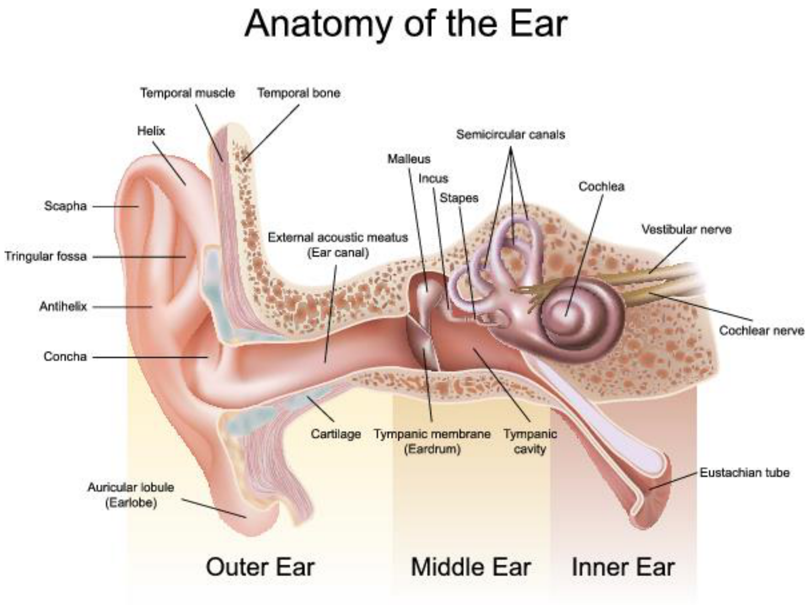


Figure 2.3: Anatomy of the ear [3].

Part of the inner ear [89], balance, called vestibular organ, can feel the other part of the stimuli is called the cochlea, bone shell is wrapped in tubular structures, curl number ring (two and a half circle) in a snail, so named. A tubular structure near one end of the stapes bottom coarser, called the base, the other end is fine, called the worm. Cochlear bone shell with membranous separation of three parallel pipeline structure, stretching from the base to the worm, respectively called court order, scala tympani and cochlear duct (or order). Court order and drum in the base of each have a window, respectively called oval window (vestibular window) and round window, both window film. Round out the window as the tympanic cavity, oval window is stapes floor block. Court order and drum at the top of the worm holes (worm) through, full of lymph in this two order, called lymph. Cochlear duct between vestibular order with scala tympani, also full of lymphatic fluid, called lymph. Separate the cochlear duct and the scala tympani membrane structure called basement membrane. By the receptor cell (receptor), the voice of the nerve endings and other structural components feel device is arranged on the basement membrane, called spiral device or of corti. If the curl of cochlear straightening, it will look from the transverse section, basement membrane, corti and adjacent structures. Sound x receptor cell is neat rows of 3 hair cells and 1 inline hair cells, supported by support cells, placed on the basement membrane. There are many fine hair top cilia, its buttercup connected with spiral device at the top of the cover film. Dominate hair nerve by the longitudinal axis (worm shaft) in the cochlea spiral ganglion. Another spiral ganglion neurons axon constitute the auditory nerve, along the worm shaft out, through the skull into the brain stem.

Auditory pathway between the center at all levels is rather complicated. Mammals of the level 1 listening to central is cochlear nucleus of medulla oblongata, it accepts ipsilateral auditory nerve fibers [90]. Most nerve fibers from cochlear nuclear cross to the contralateral, small in the same side, in the olive nuclear change neurons directly upward, or of the lateral lemniscus, arrived in the midbrain corpora quadrigemina inferior colliculus, upward from inferior colliculus fiber and a small number of directly from olives on nuclear fiber termination of medial geniculate body in the thalamus. Fiber bundle upward from medial geniculate body disperse into radial, tin steel, ended in auditory cortex, the brain is the centre of the hearing at the highest.

Ear sensitivity is based on frequency of the sound [91]. Resonance characteristics of external auditory canal and middle ear acoustic impedance frequency characteristic, cochlear adept the mechanical properties of the wave, the spiral structure of filter characteristics and physiological characteristics of the receptor cell, to the ears of the sensitivity of different frequencies of sound feeling is not the same. All kinds of animals have their hearing is more sensitive frequency range, the person is roughly 1000 - 8000 Hz, beyond the limits of the passive sensitivity.

When the intensity is big enough, is limited to not cause hearing other than feeling) can be heard in person about 20 - 20000 Hz frequency range, therefore, used to call this the audio on, more than 20000 Hz frequency ultrasonic, under 20 Hz infra-sound [90]. Animals listen to more difficult to accurately measure the frequency range, overall varies between species.

The basic function of auditory system is sound and speech recognition. Feel the voice called the ability of listening, usually expressed in the discretion of the auditory threshold, sensitive hearing or hearing threshold low said good, ability to identify voice expressed by all kinds of threshold.

Auditory Threshold

Enough to cause the minimum intensity of the sound of auditory, usually expressed in decibels [92]. Human auditory threshold can be subjective sensation as a measurement index, animal auditory threshold is conditioned reflex, behavior observation or electrophysiological method determination. Normal ear auditory threshold level owing to the different frequency, different kinds of animals is not the same, all kinds of mammals auditory sensitive frequency range is not the same, but they are quite close to the best threshold and threshold pressure in roughly 0.00002 mpa, the sound pressure make the tympanic membrane vibration displacement amplitude of about 0.1 nm. If this is of high sensitivity, improve again, may be due to often hear the sound of the air molecules Brownian motion and constantly all day and night.

2.2 Emotion

Emotion, as a feeling that accompanies human life, has been very difficult to define. Because it is very complicated and difficult to describe accurately [93]. There are many descriptions and definitions of emotions in history. From the beginning of James's theory [27], this theory holds that human emotions are generated by human beings receiving external stimuli that cause the nervous system to act and generate emotions. To Schachter's Two-factor theory [94], this theory describes that emotions are triggered by the receipt of physiological stimuli and emotional experiences and Goldie [95] believes that emotions are caused by the perception of the body caused by the reaction of the body.

In general, emotion is a group of physiological responses produced by the human body under the multiple stimuli of external, physical, physiological and spiritual [96]. Commonly speaking, a complete emotion requires the human body to produce some physiological changes after recognizing the stimulation generated outside the body. The psychological state has changed and the body interacts and expresses to the outside world. At the same time, it has an impact on the next behavior. This series of changes requires coordinated synchronization to complete. Therefore, the accurate description of emotions is also very complicated [97].

2.2.1 Discrete Emotion

In order to accurately describe and define emotions, humans have conducted many studies on it. As early as 1935, Hevner [26] had a preliminary study of emotions. It summarizes some of the relevant types of emotions by using music to stimulate the subject. Now, researchers use a variety of different methods to describe the abstract concept of emotion in a three-dimensional approach.

In the study of emotional definition, psychologists and researchers classify, summarize, and summarize emotions using two different methods. In the first method, based on Ekman's [98] research, emotion is considered to be a collection of discrete sensations. Some people think that. A large number of complex emotions can be summarized into six different basic types (anger, disgust, fear, happiness, sadness and surprise). The rest of the emotions that are subtly different from the basic categories can be classified and summa-

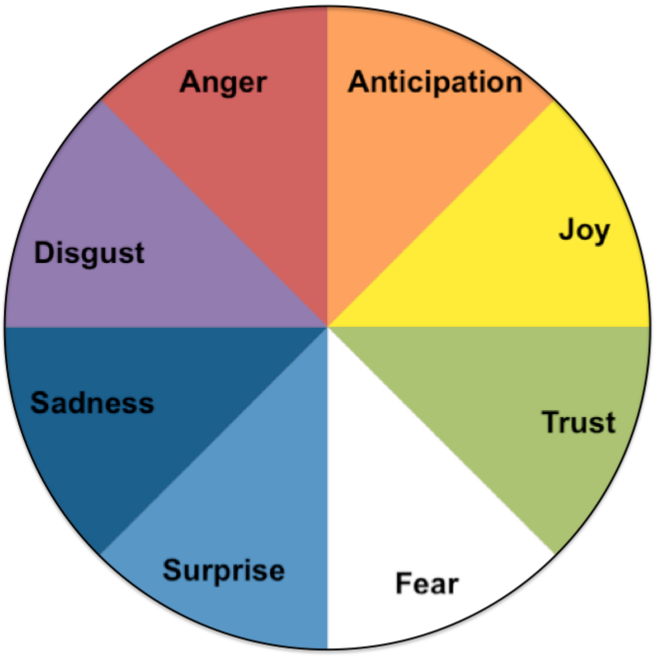


Figure 2.4: Wheel of emotions [4].

rized according to these main categories. These six basic types of emotions are not related to each other and are completely discrete. Sloboda and Juslin [99] believes that according to Hevner's research, there should be eight basic types of emotions. Still others think that there are nine [100].

2.2.2 Dimensional Emotion

But the human emotions is very complex . Sometimes it is impossible to meet the requirements by simply relying on a few basic emotions. Psychologists began to think about how to more effectively and accurately describe the emotions they want to express.

For accurately distinguish between different emotions, psychologists quantify these emotions according to different emotional descriptions. Then they used factor analysis techniques to measure the correlation between these emotions. Finally, using the dimension description method, the coordinates are used to locate various emotions. This method of classifying emotions is called dimensional emotion. After a series of research on positioning and positioning, the researchers summarized three dimensions for expressing emotions: valence, arousal and potency. All three dimensions can express emotion to a certain extent [101] [102] [103].

Until Russell [104] in 1980, a two-dimensional ring model based on arousal and valence

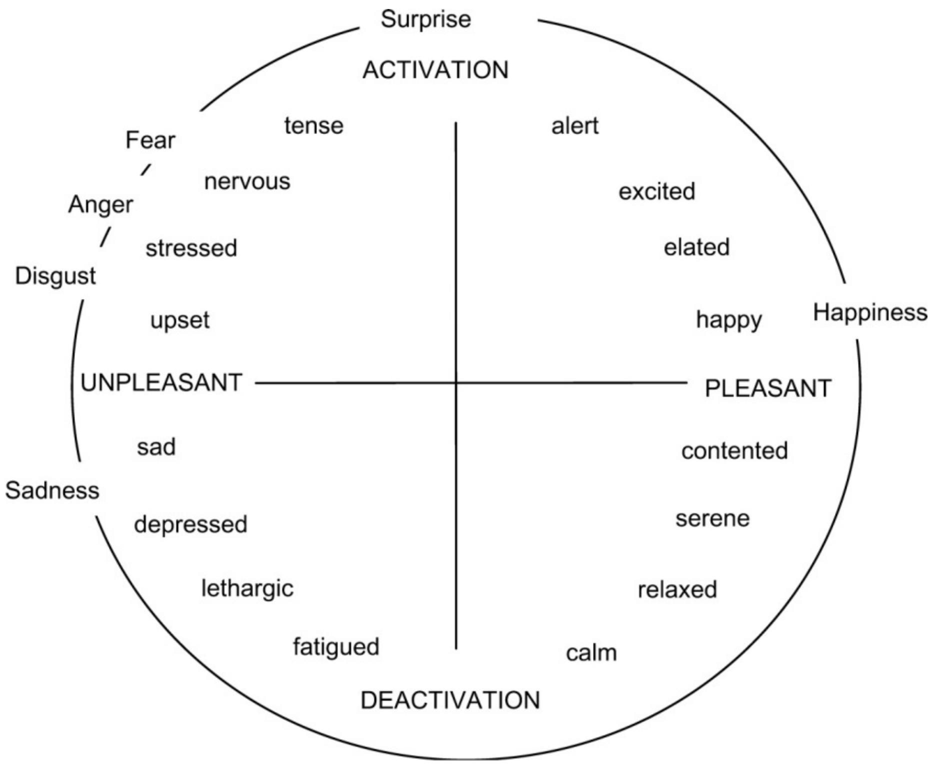


Figure 2.5: Multi-dimensional emotion analysis [5].

was invented. Russell found in the study that the two-dimensional circular model can very accurately describe the emotional position that the subject wants to express, and it is very simple to use. The Arousal-Valence (A-V) model shown in Figure 2.6. It only needs to be based on the feeling of quantifying emotions, and it can be drawn on the circle according to the degree of A-V and he thinks that A-V are more suitable for expressing emotions than other dimensions. The position of different emotions can indeed be displayed very accurately on this two-dimensional plane [105]. Thus, dimensional emotions and A-V models are beginning to be effectively used in the emotional realm.

2.2.3 Emotion Recognition

Emotion recognition of speech signals can be seen as a pattern recognition problem, the establishment of Speech Emotion Recognition (SER) system mainly has the following key points: classification of emotion, setting up the emotional speech database, choosing speech emotion feature parameters and construction of emotion recognition model [106]. In the past few decades, what characteristics can effectively reflect the emotion in the speech signal, scholars have made a lot of research. Due to people's perception of voice is very diverse, comprehensive consider acoustic feature of emotion is a very difficult work, considering the processing capacity of the computer, only by some parameters

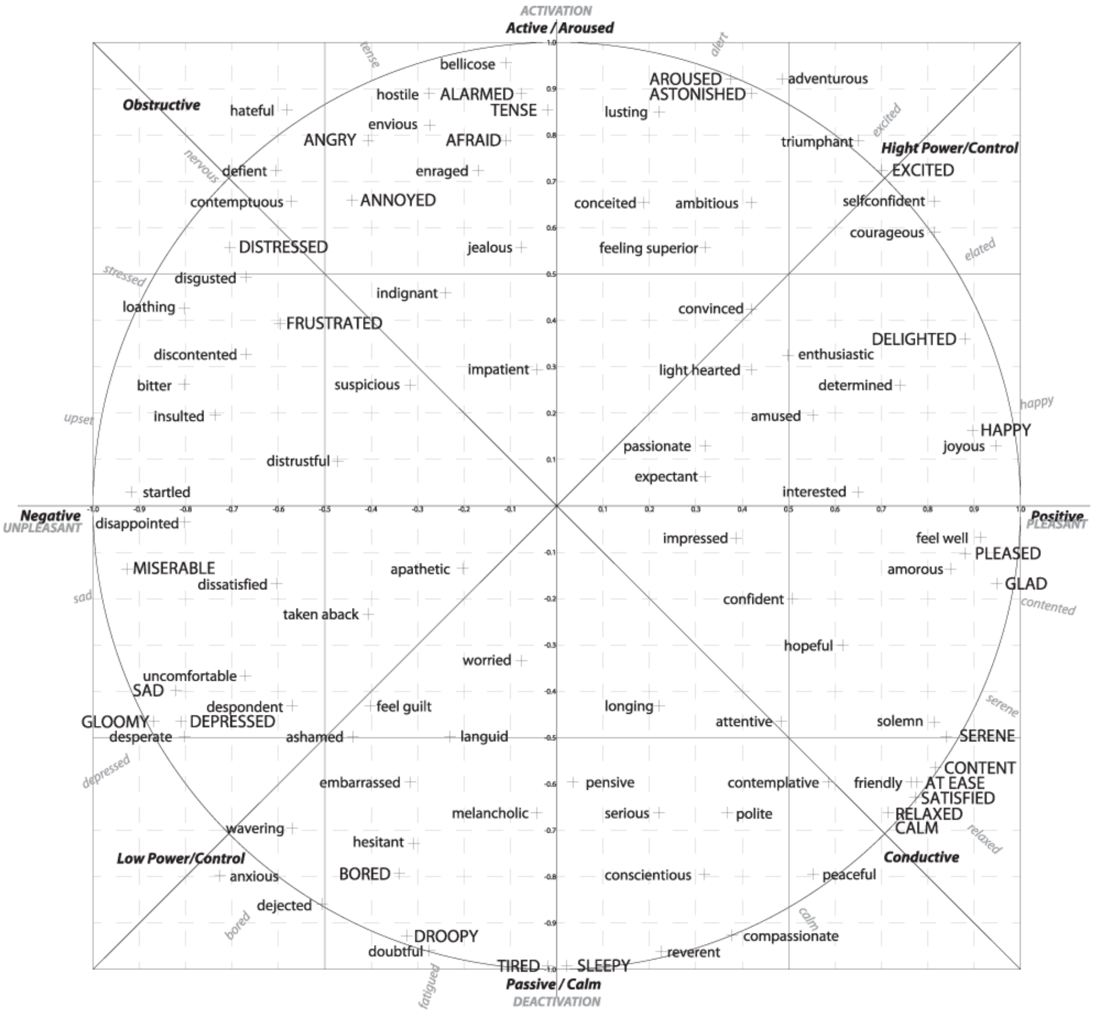


Figure 2.6: Emotion representation by A-V [6].

from a certain extent, to summarized the acoustic characteristics of the emotional speech. Researchers in psychology and language psychology provides a large number of research results about phonetics and metrics, can be used to analyze the emotional speech feature. Throughout all kinds of literature and staff from all over the world in recent decades of research, according to the characteristics of the used by emotion recognition, almost are mostly based on prosodic features, such as pitch, intensity and duration of these types and these characteristics derived on the basis of a large number of parameters, such as the basic characteristics of mean, median, range, variance, etc. In some literature also considered the phonetic features, such as resonance peak information, etc.

Now a lot of methods for audio emotion recognition, such as the main element analysis Principal Component Analysis (PCA) and maximum likelihood the Naïve Bayes (NB) classifier and kNNs classifier, artificial Neural Network (ANN) and Support Vector Machine (SVM) and Hidden Markov Model (HMMs), etc.

Although audio emotion information processing has made a series of progress in many approaches, but in the face of the real HCI, there are many problems remain to be further research [107]. Audio emotion feature analysis and emotion modeling statistical analysis requires a lot of experiments, which requires large-scale, high sense of reality of emotional speech database. SER of the existing research results are based on a smaller range of material. Used to study the speech samples mostly belong to artificial materials, and in the majority with simulated corpus, so in the next few years of study, establish a standardized large emotional speech database researchers will be the direction of the efforts, and be sure to solve the problem, only solved the problem, studies of emotional speech Can be long-term development. Characteristic is the key of SER, acoustic characteristics of voice and emotion state of the relationship between the conclusion of research also lack of certainty. Most of the existing research by analyzing specific speech acoustic features and the relationship between the specific emotional state. This kind of research idea has a lot of limitations, some phonetic characteristics and relationship between emotional state is not explicit, direct, therefore, a deeper research of the relationship between phonetics characteristic and the emotion, should be combined with emotional cognitive research results and the language habits of the people, more essentially mining the approach in which people express emotion, emotional acoustic expression model is set up. From the perspective of signal processing, at the same time should be more attention to the combination of features and derivative of the contribution of emotional expression, seek reasonable emotional acoustic representation.

2.2.4 Music Emotion Recognition (MER)

MER is one of the special and interesting part in emotion recognition. With the more requirement for convenience in music searching area, the classification for the music follow the emotions has been proposed as the effective approach to searching and classing the music [28]. From the day this question been proposed, the researcher focused the key which is the conceptualization and correlation of emotions from music. MER has been researched surround this aim and many methods and results have been proposed about it.

The traditional methods for the MER is extract the featrues from the music signals first. Then the features is sent to the machine learning model for train based on the class labels. At last predict the prediction labels according to the test data with the trained machine learning network. There are some research example for these methods. Bischoff [108] used the SVM [109] classifiers with Radial Basis Function (RBF) on the audio information and social annotations from “Last.fm” database for recognizing the emotions. Liu et al. [49] used the Root Mean Square(RMS), spectral features after K-L transform through the Gaussian Mixture Model (GMM) for recognizing the emotions. Wieczorkowska [110] used the various features and the features for the speech detected on kNNs model to extract the emotions from the music and Yang [111] did the similar approach. But he use the developed model Fuzzy kNNs classifier (FkNNs) on the 15 features chosen features for MER and Feng [112] used the Average Silence Ratio (ASR) features working with the neural networks recognize the emotions from the music.

In 2008, Yang et al. [113] used the Daubechies wavelets coefficient histogram (DWCH) algorithm [114], spectral contrast algorithm with PsySoud [115] and Marsyas [116] extracting the features in previous work. Then he trained the data with features by AdaBoost.RT [117], Support Vector Regression (SVR) [118], and the Multiple Linear Regression (MLR) [119]. Later, he used the Music Emotion Variation Detection (MEVD) [111] [120] [121] to do the same ground true data and features for fear comparison. At last, he got his result that the regression approach has more promising prediction results than normal A-V [111] [120] [121] [122] computation algorithms in doing MER and also the regression approach can be applied in MEVD.

In 2009, Han et al. [55] used Juslin's theory [18] based on the emotion model proposed by Thayer [102] to analyse the 11 kinds of dispersed emotions (angry, excited, happy, relaxed, sad and etc.). He found these emotions due to 7 music characteristics which are pitch, loudness, key, harmonics and etc. Then he used the low-level features to train the regression functions by SVR. After he predicted the AV values and got the results, he compared them with the results obtained by GMM [123] and SVM [109]. The final results illustrated that, the SVR could raise the accuracy from 63. 03% to 94. 55%, but the GMM can only grow 1. 2% (91. 52% to 92. 73%).

Through these few parts of the MER researches, the 2 significant part of the MER system has been found as features extraction and machine learning. In the follow sections, both of them will be reviewed individually.

2.3 Features Extracted for Emotion Recognition

The most effective and the most significant part of emotion recognition research are features which are extracted and selected from an original signal. Generally, a feature in point can greatly improve the correct rate of recognition. There are so many feature found by the researches, such as F0, RMS energy, Rhythm strength etc. [53]. Most of them are effective if be used in reasonable situation. So some common and famous features should be mentioned.

2.3.1 Fundamental Frequency

The fundamental frequency, as a basic property of all complex waves signal, is widely used as a feature in emotion recognition and call as F0. The fundamental frequency, as its name suggests, is the lowest frequency of any freely oscillating composite wave. Because of this, it usually also shows the highest amplitude in the entire composite wave. Due to its particularity, the fundamental frequency is usually the most basic representation of the most obvious features of a set of wave signals. So it is one of the most basic and traditional feature [124].

Busso [125], in his work, he focused on the property and functions of fundamental frequency and extracted the pitch features of fundamental frequency. Then he built a system for SER with the fundamental frequency pitch features. Lin [126] build another SER with the features, which include the fundamental frequency, and HMMs and Support SVM as the classifier. It was shown in the survey done by Ayadi [53] that the F0 is a such popular feature in this area.

2.3.2 Spectrum

The spectrum is well known for its excellent performance in the frequency domain of the signal [53]. The spectrum is a feature that visually shows the properties of a signal that changes over time in the frequency domain. The spectrum is the result of a Fourier transform [127] of a signal that changes over time and is projected onto the frequency domain. It visually exhibits the properties of a signal in the frequency domain based on frequency and amplitude or frequency and phase. Due to its nature, the spectrum can well exhibit the frequency domain characteristics of the signal. This allows researchers to study the frequency domain of the signal more clearly and conveniently. In the field of emotion recognition, spectrum is also a classic and basic feature.

In Wu's [128] work, he used the Modulation Spectral Features (MSFs) working on the SVM classifier for the 7 kinds of discrete emotion recognition from speech signals and his system shown the excellent performance on 91. 6% accuracy and Lin [129] used the advanced spectral feature Power Spectral Density (PSD) in his work for recognition the emotions from the Electroencephalogram (EEG) signals and in Yang's [52] research for the MER, he used the feature contained spectrum into the MLR [119], SVR [118], and AdaBoost. RT [117] for training for the regression of the A-V emotion recognition.

2.3.3 Zero-crossing Rate (ZCR)

ZCR is an another useful features in emotion recognition [130]. A normal signal due to the characteristics of the oscillation of the wave, there will be positive vibration and reverse vibration. When the signal is processed into an audio signal, it is found that when the vibration changes from positive to negative due to the vibration of the wave, a phenomenon occurs in which the amplitude changes from a positive maximum value to a zero value to a negative maximum value. Each time the amplitude changes and passes through the zero frequency it will affect the properties of this signal wave. Therefore, the ZCR is a characteristic of the signal property by expressing the frequency of a signal passing through the zero point. The example figure and equation of ZCR in illustrate in Figure 2.7 and Equation 2.1. The *IIA* in the equation means that if *A* is true, the function equal to 1, otherwise equal to 0.

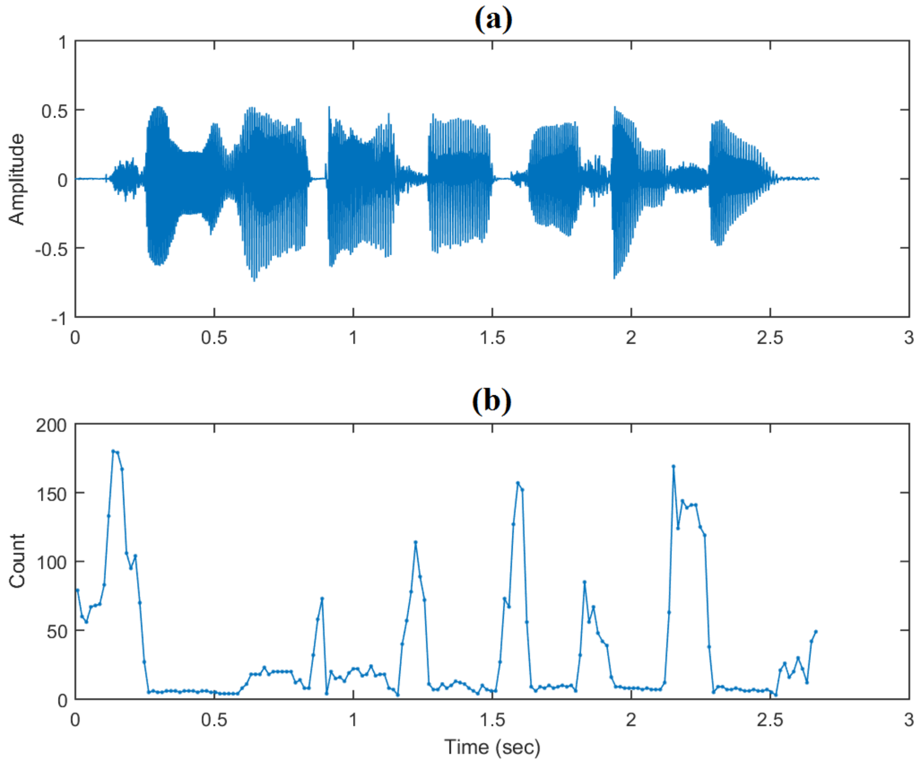


Figure 2.7: The signal and ZCR [7]. (a. the original audio signal; b. ZCR of the signal)

$$ZCR = \frac{1}{T-1}\sum_{t=1}^{T-1} II\left\{s_t s_{t-1} < 0\right\} \tag{2.1}$$

ZCR also have so many use cases in emotion recognition area [53]. Wöllmer [131] used the Low-Level Descriptors (LLD) features which include the ZCR on the LSTM model for the classification of A-V emotional dimensions. Lee [132] built a system for automated emotion state tracking by emotion recognition thinking with the features containing ZCR and used the hierarchical decision tree machine learning model for achieving. Ooi [133] used a effective Voice Activity Detection (VAD) [134] technique with the ZCR and other energy feature through the features-level fusion to predict the emotions from the audio signals.

2.3.4 Mel-frequency Cepstral Coefficients (MFCCs)

Mel-frequency cepstral coefficients (MFCCs) [135] is the one of the most useful and effective feature for MER [136]. As mentioned in the previous chapters, the principle of sounding such as human speech or instrumental music is due to the air flowing in the resonant cavity of the object and being affected by the vibration to make a sound. At the same time, human ears also have a certain specificity when receiving sound signals. If it can reasonably use science and technology to capture the changed conditions of the air according to different vibration modes, and the characteristics of human auditory perception sound, it can undoubtedly effectively improve the effect of the features. The MFCCs is a special feature that meets this goal.

The MFCCs feature consists of several different parts, Mel-frequency, cepstrum and coefficients. This is also the formation process of MFCCs. At the very beginning, the signal is converted to a spectrum by Fourier transform. The inverse Fourier transform, which is the homomorphic signal processing, is then performed on these spectra to obtain the low-frequency portion of these spectra, that is, the cepstrum. The Mel frequency is a unique frequency designed to simulate the human auditory system. It shows that humans perceive the received sound as if it were more efficient to receive certain frequencies. The equation of transformation from the normal frequency to Mel frequency is shown in Equation 2.2. When the Mel frequency is used as a filter to filter the spectrum of the first signal, the Mel spectrum of this signal is obtained. The inverse Fourier transform of the absolute values of these spectra is then referred to as homomorphic signal processing. Then the low-frequency part of these spectra is obtained, which is the cepstrum. Finally, the MFCCs can be obtained by performing Discrete Cosine Transform (DCT) [137] on the obtained cepstrum for coefficients.

$$Mel\left(f\right) = 2595 \times ln\left(1 + \frac{f}{700}\right) \tag{2.2}$$

In the AVEC2013 challenge, James Williamson [138] got the best result in DSC part with a significant audio features: MFCCs. They focused on the audio features analysis and found 2 very effective features: MFCCs and the formant frequencies. Then Then they use the close relationship between the formant frequency and the δ-Mel cepstral coefficient for reveals motor symptoms of depression. At last, they got the best result with the GMM-Based regression analysis machine learning.

Bitouck [139] and his partner improve the MFCCs feature which based on three vocabulary vocalization methods used in human speech, and combined with prosody function to extract effective features and them classed the emotions with these features by the HMMs and SVM for nice result.

In MediaEval 2013, Konstantin Markov [140] processed the music signals by extracting some features, like Spectral Crest Factor (SCF), MFCCs, Spectral Flatness Measure (SFM) etc. Then he used the Gaussian Processes regression to get the results as the Gaussian Process Regression (GPR) got the better results for static emotion estimation when it compared with the SVR. Anna Aljanaki [141], has done the data filtering before which delete some useless information, such as containing speech, noise and environmental sounds. Then he used 3 toolboxes (PSYsound [115], VAMP [142] and MIRToolbox [143]) to extract the 44 features as loudness, mode, RMS and MFCCs 1-13 etc. Next step, he finished the feature selection in both arousal and valence with RReliefF feature selection algorithm in Weka in order to top 10 significant features [113]. At last, he also used most regression machine learning method in Weka, such as SVR, M5Rules and multiple regression. Then the conclusion showed that the results of multiple regression were not weak than the outcome of M5Rules. Because the valence is a considerable correlation with arousal, the prediction accuracy of arousal was better than that for valence.

2.4 Machine Learning Methods

As part of the audio signal about the emotional feature was extracted the previous methods, a new problem come to researchers. How to detect a signal without feelings of computer perception and passed in the emotion? In order to solve this problem, machine learning the theory and method was invented. This approach aims to let the computer receives a large number of audio and mood of the relationship between the data. Then, the computer according to these relationships creates a rule to determine which characteristics corresponding to which emotions .

For successful using the machine learning, the system must has a correct progress according to the question which need to be solved. Then a model based on the known information from training data will be built for training the computer to complete the question solving mission.

The questions which need to use machine leaning normally have 4 types: classification, regression, clustering and rule extraction. For solve the classification question, a model will be built based on class labeled data ,which has the discrete value label, such as the fruit classification. The system will training by the model for judge the classes of the new sample data. For regression question, the labels of the training data are defined by the real numbers. The model will be built according to the number for more meticulous prediction from the new sample, such as the exchange rate trend and for the clustering, there is no label for the training data. However, a similarity measure is given for generalizing the data according to this measure, such as split the cars by appearance. For rule extraction, it will focus on the relationship of the statistics between the data. [144]

After understanding the type of the questions, the styles of the machine learning for different types of data are reviewed.

Supervised learning: Input data has a category tag or result tag, called training data. The model is obtained from the training process. Using the model, new samples can be speculated, and the accuracy of these predictions can be calculated. The training process often requires a certain degree of precision on the training set, without under-fitting or over-fitting. The general problem solved by supervised learning is classification and regression, such as the Back Propagation Neural Network (BPNN) [145].

Unsupervised Learning: There is no markup on the input data, and the model is constructed by inferring the existing structure in the data. The general problem solved is rule learning and clustering, such as the kNNs algorithm.

Semi-Supervised Learning: Input data is a mixture of annotated data and non-annotated data. It is also used to solve the prediction problem, but the model must take into account the existing structure and predictions in the learning data, that is, the above supervised learning. Integration with unsupervised learning. The problem to be solved by this method is still the regression of classification, which is generally extended on the algorithm of supervised learning, so that it can model unlabeled data. [146]

Reinforcement Learning: In this learning style, the model is built first, and then the data stimulus model is input. The input data is often from the environment. The result obtained by the model is called feedback, and the model is adjusted by using feedback. It differs from supervised learning in that feedback data is more from environmental feedback than by humans. The problem solved by this method is system and robot control, such as Temporal difference learning [147].

Different styles of machine learning methods focus on different questions. Suitable using the correct methods is the key to solve the questions.There are many kinds of methods for machine learning. Here some of popular and useful machine methods are introduced.

2.4.1 k-Nearest Neighbors (kNNs)

Nearby algorithm, or kNNs [148] classification algorithm is one of data mining classification techniques easiest approach. The so-called K-nearest neighbor is the k-nearest neighbor mean to say is that it can be used for each sample k nearest neighbors to represent.

The core idea of kNNs algorithm is that if a sample of k in the feature space most adjacent to the sample belongs to a category most, the sample also fall into this category, and the category having the sample characteristics [149]. The method in determining the classification decision based solely on the nearest one or several categories of samples to determine the category to be sub-sample belongs. kNNs method when the category decision, with only a very small amount of adjacent samples related [150]. Since kNNs method is mainly limited by the surrounding adjacent samples, rather than discrimination class field method to determine the category for, so for overlap or class field of more sample set is to be divided, kNNs method than other methods more suitable.

Figure 2.8 shows an example of kNN algorithm. If K=3, the red star belongs to class B, because the 2/3 of the sample near the star is class B. As the same reason, if k=6, that star

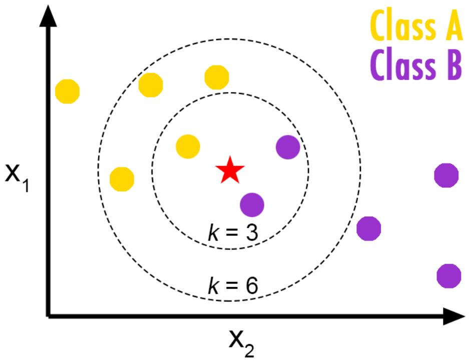


Figure 2.8: The example of kNNs working [8].

belongs to class A.

The challenge's name is: The First International Audio/Visual Emotion Challenge [151]. The Audio/Visual Emotion Challenge and Workshop (AVEC 2011 [151]) will be the first competition event aimed at comparison of multimedia processing and machine learning methods for automatic audio, visual and audiovisual emotion analysis, with all participants competing under strictly the same conditions. In the audio part of this challenge, Hongying Meng [152] and Nadia Bianchi-Berthouze as participators got the best result. They use kNNs and the HMMs classification in 3stage based on the features in 4 labels which provide by challenge and using PCA with them.

2.4.2 Random Forest (RF)

RF is one of the ensemble learning methods that is very popular for classification. The main idea is to form a forest by training and combining different kinds of decision trees, and the final classification result decided by voting of these trees [153]. The method combines the 'bagging' idea and the random selection of features.

The training process of RF includes:

1. Sampling the original training set several times with replacement, and generate several new training sets, then use the new training set train the base classifier;

2. Randomly selected some attributes of the new training set, then train a classification tree, and do not prune this classification tree.

3. Repeat last two steps times to obtain the classifiers.

After the training, it will use these classifiers to do the classification, the result decided by the voting of these classifiers.

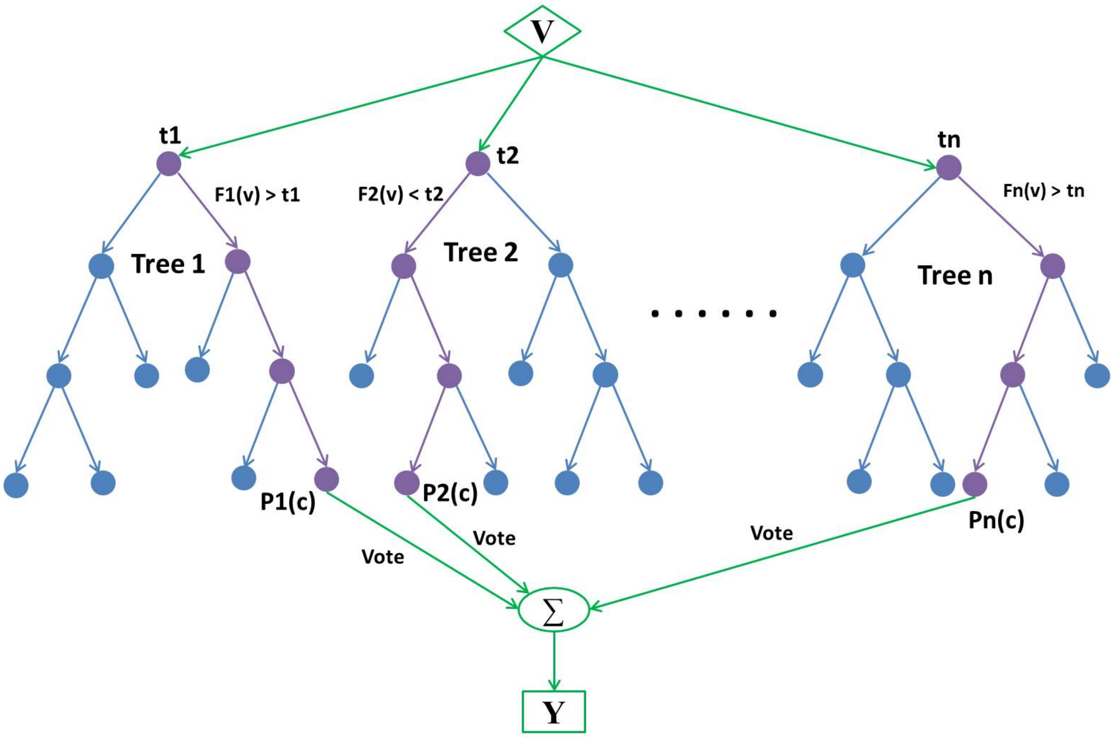


Figure 2.9: RF classifier in which several decision trees are selected and combined together for the classification automatically.

The Figure 2.9 shows that the basic concept of RF. V are the parts of features, Fn (V) are the class rule, t means the branch point in decision tree and P means the final decision result of each tree. The next step of the branch depend on if the feature is satisfy the requirement of the rule. After many of the trees go to final branch by the classifier, they will voting for which one is the most, which is the final result of the classification.

In comparison with bagging, RF [154] has the advantages of:

1. It can handle high-dimensional data. The training is based on the attributes which means it could improve the training speed.

2. Evaluate the importance of each attributes. According to the accuracy of the sub-classification tree, it can evaluate the importance of each attributes.

3. Deal with the missing attributes. Mainly because the sub-classification tree is built on different attributes, it can only select a part of available sub-classification trees to do the determination.

2.4.3 Naïve Bayes (NB)

Bayesian classification algorithm is a classification of statistics, it is a kind of use of probability and statistics knowledge classification algorithms. In many cases, NB [155] classification algorithm with decision trees and NN classification algorithm is comparable, the algorithm can be applied to large databases, and the method is simple, accurate classification rate of speed [156].

Before get in to NB classification method, one significant thing should be known: Bayes theorem.

When given a conditional probability, it is for get the probability of two events after the exchange, which is known P(AB) under the circumstances how to obtain P(BA). The explanation for what is the conditional probability is: P(AB). B represents the premise of the event has occurred, the probability of an event A occurs, the conditional probability of event A is called the next event B occurs. The basic formula is solved:

$$P(A \mid B) = P(AB)/P(B) \tag{2.3}$$

Bayes' theorem reason useful because there is often encounter such a situation in life: it can easily be directly obtained $P(A \mid B), P(B \mid A)$ it is difficult to directly draw, but it is more concerned with $P(B \mid A)$, so used Bayes' Theorem:

$$P(B \mid A) = P(A \mid B)P(B)/P(A) \tag{2.4}$$

It can be known that $P(A \mid B)$ of the cases, obtain $P(B \mid A)$.

NB classifier is a very simple classification algorithm, called it NB classifier is because of this approach is really thinking very simple, NB ideological foundation is this: For a given to be classified items appear to solve the probability of each category under the conditions of this appears, the largest of which, it believes this to be classified items belong to which category. The largest category of conditional probability will be chosen, which is the ideological basis NB.

The formal definition of Bayesian classifier as follows:

1. Let $x = a_1, a_2, \cdots, a_m$ be classified as an entry, and each a is a characteristic property of x.

2. There is a category collection $C = y_1, y_2, \cdots, y_n$.

3. Calculate $P(y_1 \mid x), P(y_2 \mid x), \cdots, P(y_n \mid x)$.

4. If $P(y_k \mid x) = maxP(y_1 \mid x), P(y_2 \mid x), \cdots, P(y_n \mid x)$, Then $x \in y_k$.

So the key now is how to calculate the conditional probability of each step 3. It can be did this:

1. Find a known collection of items to be classified classification, this set is called the training sample set.

2. Statistical obtain the conditional probability of each feature property is estimated in each category. That is:

$$\begin{aligned} P(a_1 \mid y_1), P(a_2 \mid y_1), \cdots, P(a_m \mid y_1); \\ P(a_1 \mid y_2), P(a_2 \mid y_2), \cdots, P(a_m \mid y_2); \cdots \\ P(a_1 \mid y_n), P(a_2 \mid y_n), \cdots, P(a_m \mid y_n) \end{aligned} \tag{2.5}$$

If the individual characteristic attributes are independent conditions, the following is derived based on Bayes' theorem:

$$P(y_i \mid x) = (P(x \mid y_i)P(y_i))/(P(x)) \tag{2.6}$$

Because the denominator is constant for all categories, because it simply maximize molecule can be and because each feature attribute is conditionally independent, so there is:

$$P(x \mid y_i)P(y_i) = P(a_1 \mid y_i)P(a_2 \mid y_i) \cdots P(a_m \mid y_i)P(y_i) = \prod_{j=1}^{m} P(a_j \mid y_i) \tag{2.7}$$

2.5 Deep Learning Methods

In current research field, lots of deep learning methods appear like the bamboo shoots grown after the rain every month. There are wide variety models which can be considered. So the deep learning methods selection needs rigorous and patient.

2.5.1 Convolutional Neural Network (CNN)

For the first models of the selection, the most famous and fashionable model CNN [157] [9] need to be considered. CNN is one of the classic and traditional deep learning model.

In 1962, biologists Hubel and Wiesel [158] obtained a hierarchical model of Hubel-Wiesel through a study of the cat's brain visual cortex. With the advancement of technology and new theories, Fukushima [159] and LeCun [9] have gradually improved and innovated according to the hierarchical model of Hubel-Wiesel. Finally, LeCun designed and trained the CNN (Lenet) using the error gradient backhaul method. As a deep learning architecture, CNN has the advantage of minimizing data pre-processing requirements. The biggest feature of CNN is sparse connection (local feeling) and weight sharing. Because this computational training process is the same as the convolution process, this network is called a CNN.

The basic structure of the CNN consists of an input layer, a convolution layer, a sampling layer, a fully connected layer, and an output layer. In the core convolutional layer and sampling layer structure, the two layers are alternately linked to each other and repeated several times. The neuron input value is obtained by adding the weight of the connection plus the local input weight plus the offset value and, each neuron of the output feature surface in the convolutional layer is locally connected to the output of the upper layer [9]. The detail description for CNN will follow.

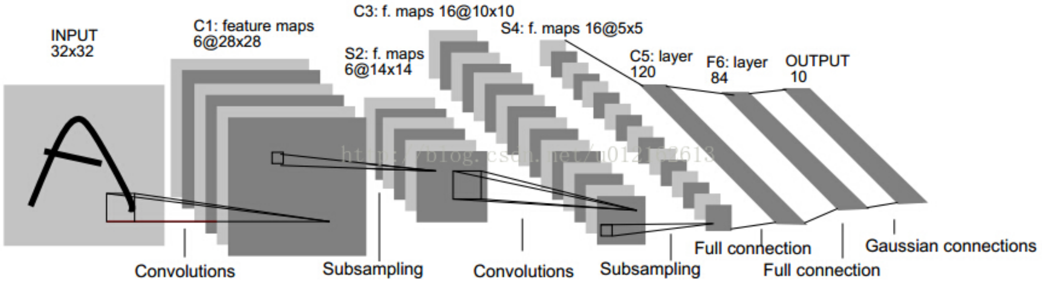


Figure 2.10: The CNN structure [9].

In the CNN network structure process [160] [9], which shown in Figure 2.10, the first to appear is the input layer. The input-layer to C1 part is a convolution layer (convolution operation), C1 to S2 is a sub-sampling layer (pooling operation), then S2 to C3 are convolutions, and C3 to S4 are sub-sampling. It can be found that both convolution and subsampling appear in pairs, followed by subsampling after convolution. S4 to F6 are fully connected. Finally, pass the classifier to the output layer.

The C layer is a feature extraction layer, and the network of this layer inputs the extraction features according to the selected region of the upper layer and the neurons of this layer. In this layer of work, the experience of wild connection plays a very important role [158]. The S layer is a feature mapping layer. Each computing layer of the network is composed of multiple feature maps. Each feature is mapped to a plane, and the weights of all neurons on the plane are equal. In addition, since the neurons on one mapping surface share the weight [161], the number of free parameters of the network is reduced, and the complexity of network parameter selection is reduced.

Specifically, the input image is convoluted by a trainable filter and an addable bias, and after the convolution, a feature map is generated at the C1 layer, and then the pixels of each group in the feature map are further summed, weighted, and biased. Set, the feature map of the S2 layer is obtained through a Sigmoid function. These maps are then filtered to obtain the C3 layer. This hierarchical structure produces S4 again like S2. Finally, these pixel values are rasterized and concatenated into a vector input to a traditional NN to get the output.

Just in 23 May 2018, Tong Liu [63] and his partners provide a strategy to recognize the emotion contained in songs by classifying their spectrograms, which contain both the time and frequency information, with a CNN. The experiments conducted on the 1000-song database indicate that the proposed model outperforms traditional machine learning method. In his work, his team used a CNN-based model to extract a spectrogram features from the song, then classify them with the CNN-based model and the result compared with the performance which provide by extracting the traditional acoustic features from the audio signals, such as Zero crossing rate, MFCCs, spectral centroid and so on as the input and classifying with SVM. The result shown that the CNN deep learning methods got nearly two times higher accuracy then the SVM one.

2.5.2 Recurrent Neural Network (RNN)

RNN is a deep learning method for solving a problem , which is that the network cant thinking or work according to the information was inputted few time before, in normal NN [162]. NN is kind of learning method which imitates the human brain [163]. A real human can think and learning things according to the previous knowledge or objects. But the traditional NN cannot train like that. It cannot predict a thing based on the previous events for the later one. But for some serialized inputs with obvious contextual features, such as predicting the playback of the next frame in the video, it is clear that such output must rely on previous inputs, which means that the network must have some “memory”. In order to give the network such memory, a special structure of the NN - the RNN came into being.

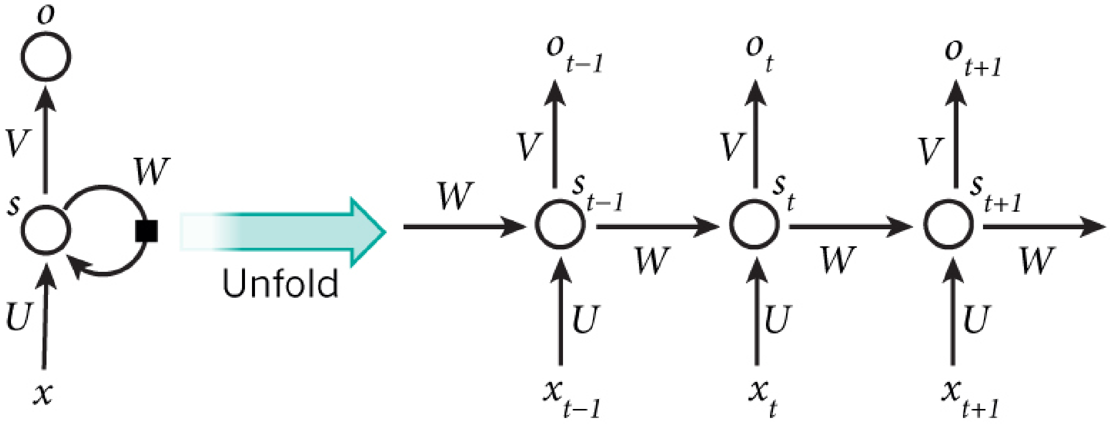


Figure 2.11: The RNN structure [10].

As the most kinds of NN, RNN got three layers in its structure as well, Input layer, hidden layer and output layer, which can be found in Figure 2.11. The thing more in the RNN is the circle with the w, which is the key for the difference to make the RNN have that “memory”. Before understand the structure, in this figure, some of parameter should be known at first. Input x_t is the moment which input to the network network. h_t represents the hidden state of time t. o_t represents the output of time t. The direct weight of the input layer to the hidden layer is represented by U, which abstracts the original input as the input of the hidden layer. The weight of the hidden layer to the hidden layer is W, which is the memory controller of the network, is responsible for scheduling memory. The weight V is hidden from the layer to the output layer, and the representation learned from the hidden layer will be abstracted again by it and used as the final output. In the left side of the figure shown the fold structure, which is the structure for all time and the unfold on is in the right side, which illustrate the structure according to the time sequence.

These are two stage when RNN runs, the forward stage and backward stage. In forward stage, the output is controlled by the current time input layer x_t and the previous hidden layer h_{t-1}, which can be found in Equation 2.8 and 2.9. The function f and g can be the traditional $tanh$ and $softmax$ or other needed function. In these equation, the secret of the "memory" can be found as that the previous time h_{t-1} provides the previous information to the current time. This movement endues the current output o_t the combine information of both current timing and previous timing.

$$h_t = f(Ux_t + Wh_{t-1}) \tag{2.8}$$

$$o_t = g(Vh_t) \tag{2.9}$$

In backward stage, each weight is updated using a gradient descent method based on error e_t between output layer and true label in each timing. The Equation 2.10 to 2.14 shown the calculation of the each gradient of weight $\triangledown U$, $\triangledown V$ and $\triangledown W$ form the last output layer backward to the each previous one.

$$\delta_t = \frac{\partial e_t}{\partial_{Ux_t + Wh_{t-1}}} \tag{2.10}$$

$$\delta_t^h = (V^T \delta_t^o + W^T \delta_{t+1}^h) \cdot * f'(Ux_t + Wh_{t-1}) \tag{2.11}$$

$$\triangledown V = \sum_t \frac{\partial e_t}{\partial V} \tag{2.12}$$

$$\triangledown W = \sum_t \delta_t^h \times h_{t-1} \tag{2.13}$$

$$\triangledown U = \sum_t \delta_t^h \times x_t \tag{2.14}$$

Therefor, the network can correct the weight for the correct prediction and the whole RNN will be trained for the mission which need to predict based on the context of a contact time series.

In the MediaEval task in 2015 [164] only had one task for people – 'dynamic emotion characterization'. But for the database, they do some changes. They connected come royalty-free music from several sources. Then set the development part with 431 clips of 45 second and 50 complete music pieces around 300 seconds for the test. Pellegrini [165] chose the RNN [166] to predict the A-V emotion prediction for the system for their sequence modelling capabilities. He completed the 260 baseline features which are given by [164] with 29 acoustic feature types which extract by ESSENTIA toolbox [167]. After 10-fold cross-validation CV setup, he got the result that valence and arousal values are correlated as r=0.626 (r: Pearson's correlation coefficients).

2.5.3 Long Short-Term Memory (LSTM)

The previous section mentioned the RNN is a effective deep learning methods for the mission which need to predict based on the context of a contact time series. But the RNN still have some weakness. The "pover" of RNN for dealing with the previous information will decrease with the growing of the distance between current timing and the timing needed [168]. For example, if RNN is used to predict the thing based on the information just beside it, it will be easy success. But if it need to predict the same thing based on the information input before long time in time sequence, it will be high rate with wrong prediction [169] [170].

However, the LSTM, which is one of the variant of RNN can solve this problem smoothly. LSTM networks are well-suited to classifying, processing and making predictions basedon time series data, since there can be lags of unknown duration between important eventsin a time series. LSTM was proposed by Hochreiter Schmidhuber [169] and improved and promoted by Alex Graves [171]. LSTM avoids long-term dependencies by deliberate design. It improve the key cell f function in RNN structure with a three gate cell layer, input gate, forget gate and output gate. These gates can filer the information from the previous timing information and current timing information input and for the output, the gates can keep the useful information stay in the layer cells [172]. The detail of the LSTM deep learning method will proposed in following chapter 4.

2.6 Music Emotion Recognition (MER) Application

After the demand for new music classification, which because the old labels cannot meet the needs of customers [28]. Emotion as the new third common label was started using in some popular commercial music website [173]. Because emotional music-based retrieval is receiving more and more attention in academia and industry. With the developing of research and theory in academia, some of the applications are used in real life of people.

For example, “i.MV” [174] [175] [176] and Mood cloud [177] [178] etc. used the music emotion analysis in the multimedia system for better use experience. For the mobile media production, the MP3 and cellphone can play the music which is most suitable for the current emotion of the audience based on the MER [179]. Even the most comfortable music which is according to the emotion of the customer can be play in a restaurant or meeting room [180].

And there also are some applications for the HCI. Imaging, there is a multimedia system can arrange the music and pictures automatically based on the emotion classification and it even can use these graph and music to design an MV for the customer [181] [182]. Of course, the most commonly used function for users is to choose a song or music that best suits emotions [183] of people. Because people always choose the music they want to listen to according to the emotional state at the time. When searching and retrieving music based on the user’s emotions, it is more accurate and reasonable [18].

Recently, the more features and more learning methods will be found and design after research. There would be more application based on MER to make life of people more convenient.

2.7 Summary

In this chapter, the literature review of the MER has been done for the research challenge. The knowledge of music signals, emotions, features, learning methods and some of the application have been reviewed. After doing the literature review of sound and audio, the knowledge helped the pre-process part in following researches. The signal can be processed based on different kind of audio signals and the literature of emotions guides the researches focus on the discrete emotion data.In features part, the approach of features extraction has been learned. The kinds of the features and the application field has been understood, such as fundamental frequency, spectrum, ZCR and MFCCs. The machine learning literature was mainly reviewed the 3 basic traditional machine learning methods and 3 popular deep learning methods for support the following researches and field of the MER are also reviewed for the applications and main direction of research.

.All the review work aim to learning and understanding the MER area and preparing for solving the research challenge which mentioned previous chapter. Chapter 3 will focus the new features design and selection according to the reviewed paper. In Chapter 4, a selected deep learning methods LSTM will be used for if it can provide better performance for the system. Chapter will show and new kind of MER system building based on a music theory element – chord.

Chapter 3

MER Based on Feature Combination and Random Forest Classification

3.1 Introduction

After doing the reviews of literature, as one of the most classic human inventions, music appeared in many artworks, such as songs, movies and theaters. It can be seen as another language, used to express the author's thoughts and emotion. In many cases, music can express the meaning and emotion emerged which is the author's hope and the audience's feeling. At the same time, music as a sound, it can express different meanings by changing sound characteristics, such as frequency, amplitude, so that the audiences have different feelings. In the famous TV show "I'm a singer", a singer who can act only by the rhythm of the song adapted, with the original tone to express different emotions. Most of the audience will be affected with such sentiments, but they are not as accurate expression of this change as professional judges. So a MER system will be helpful in this kind of situation. A piece of music with clear emotion information inside will be beneficial for the entire entertainment industry and even the arts because the artist can use this information for their design and performance.

A typical MER system is to process the music signal for audio features extraction, and then these features are used for classification based on machine learning methods that is based on the training of existing data. The system can determine what kind of emotion this music belongs to.

In this chapter, firstly, common audio features are extracted from both channels of the music recordings and then other features extracted inspired by EEG signal features analysis. Finally, RF classifier is used to efficiently combine the features to boost the overall performance.

The rest of the chapter is organized as the following. In the section 3.2 related works of this field are reviewed. The proposed MER system is introduced in section 3.3. In section 3.4, the detailed experimental results will be illustrated and the work is summarized with conclusions in last section.

3.2 Related Work

In recent years, there is a strong trend on emotion recognition from music due to the requirement of the entertainment and digital media industry. Public open databases have been produced and extensive researches have been done by researchers from all over the world.

For this chapter, the aim is to provide a kind of system which can effectively recognise the emotion from the audio, especially music, signal. After the review of lots of work, it can be found that the most significant parts of the system to affect the efficacy are feature extraction and machine learning. So, several papers for these two parts have been reviewed.

In 2009, the INTERSPEECH 2009 emotion challenge [184] has taken place in Germany. In this challenge, the organisers provide a kind of features which include the 16 kinds of LLD feature (MFCCs 1-12, ZCR, RMS, F0 and HNR). These features all extracted by open source tool openSMILE [185] and are proved effective in future.

At the most recent MediaEval task in 2016 [186], the target of the task was changed to recognise the emotions from a movie clip. This task uses the ACCEDE database to provide data for both sub-tasks. The global emotion prediction sub-task uses 9800 video clips from 160 movies. This sub-task requires the contestant to use the data to predict the AV affective region for each segment of the test set and the second sub-task is continuous emotion prediction. It requires investigators to predict the complete 30 films in which the emotional regions change over time.

In this competition, Chen and Jin [187] used the features called IS09, IS10 and IS13, which based on INTERSPEECH 2009 [184], 2010 [188] and 2013 [189] Emotion Challenge configuration extracted by OpenSMILE [185], and compared with MF-generated Bag-of-Audio-Words (BoW) and Fisher Vector Encoding (FV) features. Then they combined them with CNN extracted image features for SVR and RF to do classifications. The final submission results show that the results obtained by using FV features combine with IS09 features and CNN image feature, and using random forests as classification are higher than other feature. So the IS09 feature from openSMILE is selected as one of the features for the system.

After reviewed the paper from Robert Jenke, a kind of think has appeared. In his paper [44] which was published in 2014, he did some features selection for the EEG data signals. From these kinds of features, lots of global statistics features have been selected. For example, mean, standard deviation, Hjorth features and fractal dimension. His paper shows that these kinds of features are very effective for the EEG signal processing. These features show a lot of global statistics features of the EEG signals. In other words, the features give much effective information about the properties of the signals, which are kinds of wave. Considering the similarity between audio signals wave and EEG signal wave, these kinds of statistics features can be tried to use in the audio signal processing part. That will provide some features to show what properties the music got and the performance of the features were tested in the experiment of this chapter.

In 2010, Kim et al. [48] have done a comprehensive work in MER review. He started with the psychology research on emotion with the AV space and the perceptual considerations. In feature part, he described the lyrics feature selection and the acoustic features. Lyrics features selection was based on Affective Norms for Chinese Words (ANCW) [190], Affective Norms for English Words (ANEW) [191] and the Pleasure (Valence), Arousal and Dominance (PAD) [192] to select the effective features from the signals. For the acoustic features, he gave 5 types (Rhythm Timbre, Dynamics, Register, Harmony and Articulation) 17 features (MFCCs [193], spectral contrast, RMS energy, spectral shape, Rhythm strength and etc.). But he focused on the MFCCs to comment and in machine learning part, he mentioned the RF, SVM, NB, kNNs and etc. In the end of his paper, he gave an example of some combinations of the emotion types. For example, Yang and Lee [194] combined the both audio and lyrics for emotion classification with 12 low-level MPEG-7 descriptors and increase 2.1% (82.8% vs 80.7% with only audio) of the results and other review was for Bischoff [108] combining audio and tags (use Music-IR tasks as classification to make the tags like happy and sad [195]) by Bischoff with 240-dimensional features and 178 mood tags. Then Bischoff proves that the combination is better than the single one.

In the reviewed paper in 2011, Khan and his partners [196] compared the kNNs and SVM classification for the SER. In their experiment, 7 different emotions should be recognised based on an English speech database. Two classification methods were tested separately for predict these seven emotions. From the final result can be found that the kNNs got 91.71% while SVM got 76.57% accuracy in overall performance. However for the single emotion, these two methods Win with each other.

In 2013, the researcher Vinay et al. [197] did an experiment in SAVEE database for testing the performance of NB. The features, such as pitch, linear predicate coefficient, MFCCs [198], DELTA etc., have been extracted for classification training based on 66% data of the whole database. After machine learning process by NB with 10-fold cross validation, the output predicted 7 emotions for comparing with the true labels. Can be found from the final results, the system got 84% accuracy, which means the NB is highly effective for this system.

In 2015, Song and Dixon [199] used the SVM and RF to test the different models of MER. The 2 different databases which contain 207 songs has been used in this experiment. Then the features, which included RMS, peak, MFCCs etc., were extracted by MIRtoolbox 1.5 from 2 kinds of clips groups (30 or 60seconds and 5seconds). The result showed that the RF (40.75%) one in clips 5 seconds was better than the SVM (40.35%) one and the RF also was used in following experiment for others model test.

And also in the Chen and Jins work [187], the combination features with CNN image features and IS09 trained by RF methods gave the best result.

Apart from above work, Gao [200] used the 6552 standard low-level emotional features which were extracted by OpenSMILE to compare with his multi-feature (GMM Super Vector (GSV), positive constrain matching pursuitPCMP, multiple candidates and optimal path) with the APM database. Then he does the machine learning by SVM classifiers with RBF kernel and got good results. For the results, his working proved that his multiple features are more effective than only use the openSMILE one by 5.25% (from 74.9% to 80.15%)

In research of this chapter, another MER system is tried to build in advance of the previous system [200]. The main differences were that: 1).The system extracted the feature from two channels; 2). The system extracted extra features of the statistics which inspired by EEG analysis methods; 3). The system used the RF to combine these features together in the classification efficiently.

According to the review works done above, the IS09 audio feature was thought as the most suitable audio part feature for this recognition system and in addition, the global statistics features from EEG analysis will be tested in the research. For the machine learning part, the RF method was looking as ensemble part of this system. RF is one of the machine learning methods based on the decision tree. It contain several trees to decide the final prediction according to the classes. It can assess the importance of variables when deciding a category and it also keep the high accuracy even the data is not balance or lost some information. Because this research is based on a four emotion tagging music database, the RF as an effective classification with multiple components will suit for the database. However, there is also a comparison part with two traditional and related methods (kNNs and NB) in the research for proving the performance of RF.

3.3 Emotion Recognition System

Based on preliminary analysis of the music data, two channels of data are used for feature selection. A better system will be tried to build for the musical emotion recognition with better features and better classifiers.

3.3.1 System Overview

Figure 3.1illustrates the overview of the proposed system. Both channels of the music signals are processed separately and typical audio features are extracted from them. In addition, some statistics features used for EEG analysis are also extracted. Then these features are combined together as the overall features for the classification. Finally RF classifier is chosen to ensemble all these features together and to predict the emotion categories of the music.

3.3.2 IS09 Audio Feature Extraction

There are many features existed for audio signal analysis. Here some of the common emotion related features are selected based on "The INTERSPEECH 2009 Emotion Challenge" [184], in which these features can be very accurately reflect the most emotional characteristics of the audio signal. These features include 12 functionals and 16 LLD. Totally, there are 384 features selected. In the following, these features will be described briefly. The detailed information can be found in [184]. Then the configuration of the INTERSPEECH 2009 has been used in the features extraction open-source toolkit: OpenSMILE [201] for the features of INTERSPEECH 2009 which named as IS09, which is

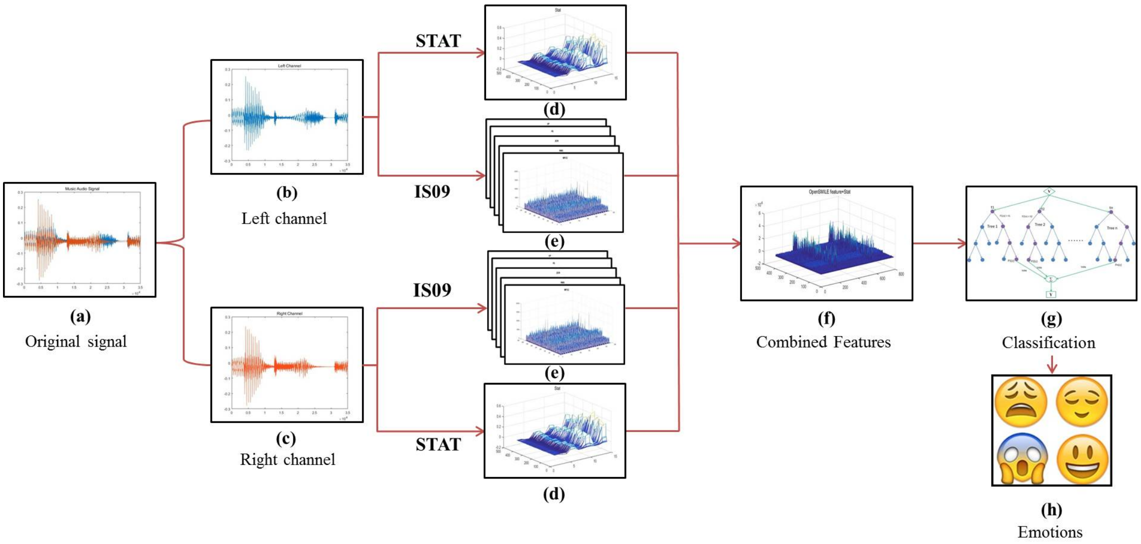


Figure 3.1: The overview of the MER system. (a). original signal, (b). left channel, (c). right channel, (d). statistics features, (e). typical audio feature, (f). two channel and all features combination, (g). random forest classifier) (h).emotion prediction.

used as main features for the low level features part.

The 16 LLD include 5 kinds of features which will shown follow and illustrated in Figure 3.2. Can been found form the Figure 3.2 that these 5 kinds of features are complete difference in amplitude and demotions. Even they are extracted from same signal. RMS Energy [202] is based on the amplitude of the peak value of the signal to calculate the power of a signal, which shows the energy carried by the signal. It is one of the common audio feature representation.In MFCCs [198], Mel frequency is a major concern for the human hearing characteristic frequency. Because humans often unconsciously for a certain period of frequency is very sensitive, if the analysis of such frequency will greatly improve accuracy. MFCCs is a linear mapping of the audio spectrum to Mel spectrum, then the conversion factor cepstral frequency domain when available. ZCR of a signal means the ratio that the signal crosses from one end of the axis 0 to the other side of the axis. It is usually expressed by a signal of a predetermined value of 0, or as a filter are searching for a rapidly changing signals. The F0 [184] means the fundamental frequency, in a complex wave is a periodic waveform the lowest sum of frequency. In music, it is usually the lowest pitch notes. The Voiceprob(VP) means that the voicing probability computed from the Aggregate Channel Features (ACF).

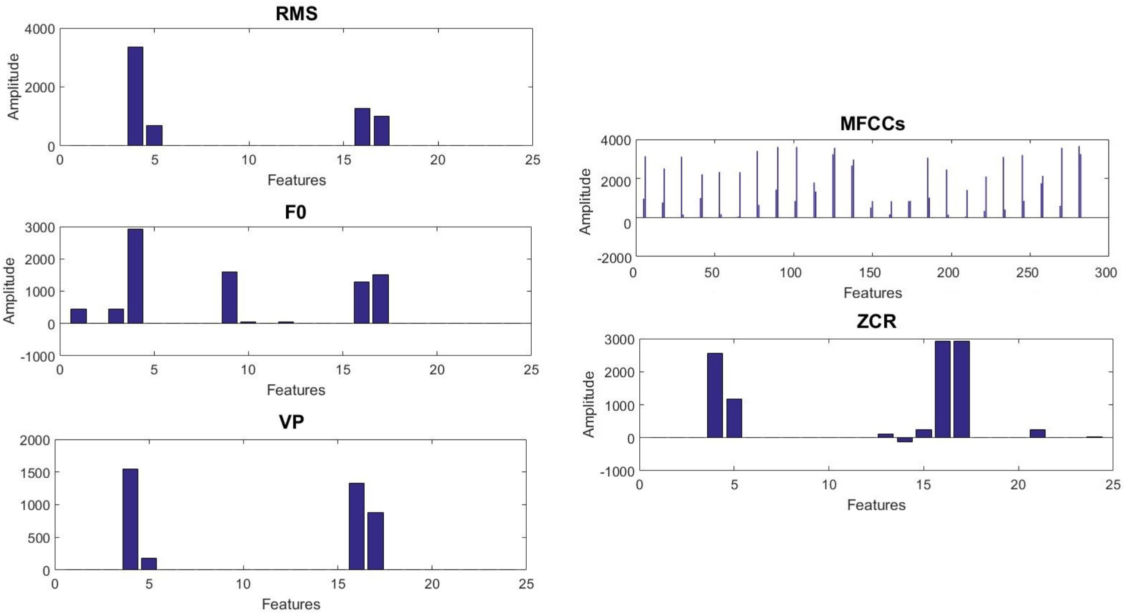


Figure 3.2: Five common audio features selected from IS09.

3.3.3 The Global Statistics Feature Inspired by EEG Signal Analysis

In the emotion recognition feature extraction, there are some interesting feature extraction methods [44] by extracting features from the brain wave (e.g. EEG) signals.

If the people is stimulated external, such as music and voice, the emotion of mind will changed with the EEG signal changing. So these features were extracted in this research and find out whether this set of features will enhance the musical emotion recognition accuracy or not.

a. Basic Statistics Features (STAT)

Through the reviews of the related work, it found that most of the research much focus on the more representative and more specific features of the signals. For example, single MFCCs feature can help to training the machine about the details of sound production progress. It really works well in most time. But the common and macro features also can contribute their usefulness for improve the recognition. According to [44], the basic statistics features (STAT) of EEG signals works well in his work. So after considered the similarity between EEG signal and an audio signal, these characteristics are the basic statistical characteristics of the signal in the time domain. The figure of STAT feature can be found in Figure 3.3.

For the statistical characteristics in STAT, there are 7 different characteristics features.

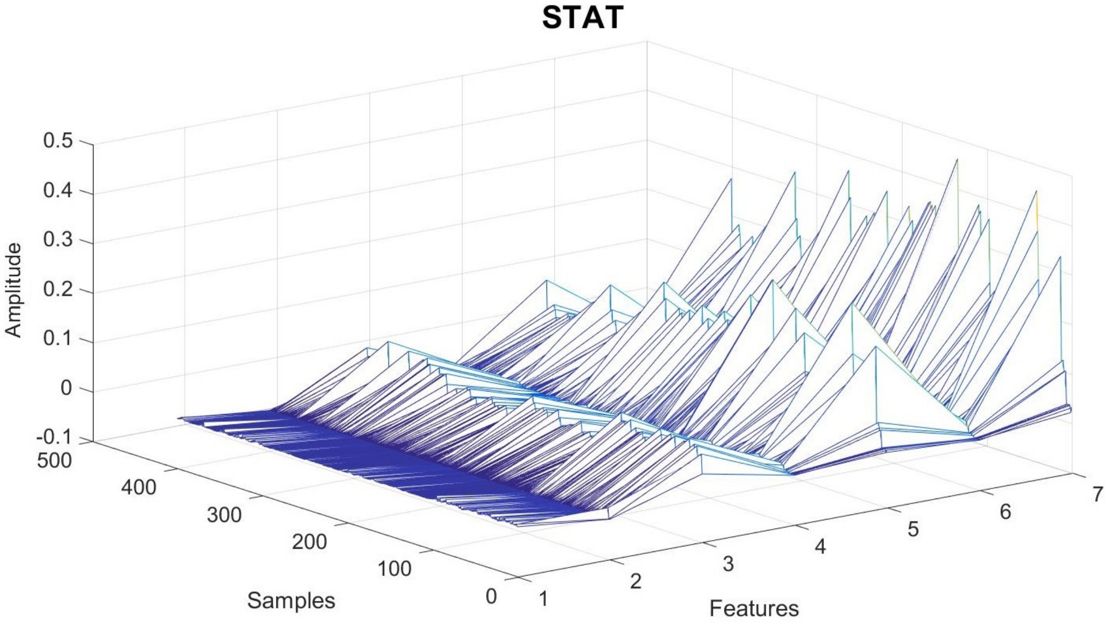


Figure 3.3: Statistic Feature.

Their equations shown as follow from Equation 3.1 to 3.7. In these formula of features, T is the time sample number of then recording. t means the time series. ξ means the vector in single audio channel recording and $\xi(t)$ means the vector changes with the time changing.

Power:

$$P_\xi = \frac{1}{T}\sum_{t=1}^{T}|\xi(t)|^2 \tag{3.1}$$

Mean:

$$\mu_\xi = \frac{1}{T}\sum_{t=1}^{T}\xi(t) \tag{3.2}$$

Standard deviation:

$$\sigma_\xi = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\xi(t)-\mu_\xi)^2} \tag{3.3}$$

First difference:

$$\delta_\xi = \frac{1}{T-1}\sum_{t=1}^{T-1}|\xi(t+1)-\xi(t)| \tag{3.4}$$

Normalized first difference:

$$\overline{\delta} = \frac{\delta_\xi}{\sigma_\xi} \tag{3.5}$$

Second difference:

$$\gamma_{\xi} = \frac{1}{T-1}\sum_{t=1}^{T-2}|\xi(t+2)-\xi(t)| \tag{3.6}$$

Normalized second difference:

$$\bar{\gamma} = \frac{\gamma_{\xi}}{\sigma_{\xi}} \tag{3.7}$$

b. Other Four Kinds of Statistical Features

For the same thinking with STAT, other 4 kinds of statistical features have been proposed. Hjorth features is another kind of statistical features for signal processing which based on Hjorth's [203]introduction. The Fractal Dimension (FD) [44] is used to describe and measure the complex objects which is hard to describe by normal dimension because of the irregular shape or volume in geometry.Higher Order Crossings (HOC) is in order to use the high-pass filters, which called Herein, to zero-mean time series for oscillatory pattern of the signals. The PSD shows that the features of the different frequency power in time dimension. The researchers can obtain some statistical properties of the signal, such as the tone and pitch of the sound, through power spectrum analysis.

3.3.4 Classification

The final step of whole system will be the machine learning part. Considering that the huge amount of methods can be used in the system, the most effective classification methods will be select after the performance testing between some alternative options.

a. k-Nearest Neighbours (kNNs)

kNNs [148] classification algorithm is one of simplest data mining classification techniques. The so-called kNNs is mean to say is that it can be used for each sample kNNs to represent.

The core idea of kNNs algorithm is that if a sample of k in the feature space most adjacent to the sample belongs to a category most, the sample also falls into this category, and the category having the sample characteristics [149]. The method of determining the classification decision based solely on the nearest one or several categories of samples to determine the category to be sub-sample belongs. kNNs method when the category decision, with only a very small amount of adjacent samples related [150]. Since kNNs method is mainly limited by the surrounding adjacent samples, rather than discrimination class field method to determine the category for, so for overlap or class field of more sample set is to be divided, the kNNs method is than other methods more suitable.

b. Naïve Bayes (NB) Classifier

Bayesian classification algorithm is a classification of statistics, it is a kind of use of probability and statistics knowledge classification algorithms. In many cases, NB [204] classification algorithm with decision trees and neural network classification algorithm is comparable, the algorithm can be applied to large databases, and the method is simple, accurate classification rate of speed.

NB classifier is a very simple classification algorithm, called it NB classifier is because of this approach is really thinking very simple, NB ideological foundation is this: For a given to be classified items appear to solve the probability of each category under the conditions of this appears, the largest of which, it believes this to be classified items belong to which category.

c. Random Forest (RF)

RF [205] as a mainstream machine learning, it can handle high-dimensional data, and can be used to assess the importance of each category according to the final results of forecast accuracy. Most importantly, he can estimate the missing part of the sample, and provides high accuracy results. This approach combines machine learning ’baging’ ideas and features. It is possible to sample the original sample and the formation of a new training set. Then they were randomly training for different tree. After a few times of packet classification training, get the final result by voting the decision tree.

3.4 Experiment Results

The following experiments are to evaluate the proposed system for emotion recognition from music recordings. APM database [200] was used this experiment. In this study, firstly Matlab extracted one-dimensional audio signal, the signal and two channels separately. Use OpenSMILE extracted five characteristics of the signal, RMS energy, MFCCs, ZCR, F0 and Voice Prob and extract STAT. It features a total of 12 set last two channels together. RF classifier using machine learning, and get the final accuracy of the confusion matrix correct rate.

The experimental results will be compared with the results in previous work [200] because the same APM database was used.

3.4.1 APM Database

“APM Music” is the largest production music library in the industry as a production music company. It contains more than 40 libraries, more than 475,000 tracks and CDs and he almost contains almost every type and style of music. Therefore, this article chose the APM database [200],which based on this huge music data, as this experiment database.

In order to make sure the effect of the features during the real situation, an experiment based on APM MER database be started. APM music database is a music data in wav format. The database including 4 kinds of emotion: happy, sad, relax and fear. Each of the emotion has 100 audio samples with around 35 seconds long.

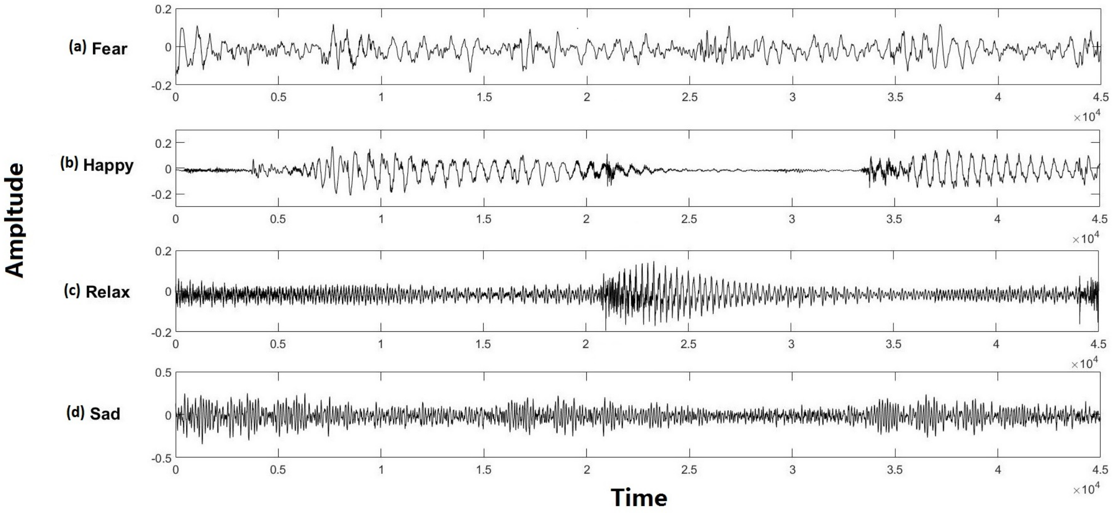


Figure 3.4: Four audio signals for 4 different emotions.(take continuous 45,000 sampling points from the music recording for approximately one second).

3.4.2 Experimental Setting

At the beginning of the experiment, the audio data is processed as a set of data, and feature extraction is performed using OpenSMILE software tool. However, by observing the audio data, it was found that two different channels of audio may have different effects on the emotion recognition. Divided into two channels were tested.

The system uses software to extract music emotion OpenSMILE part features. The two channels of the audio signals are separated and input to OpenSMILE software, and then IS09 was selected as the configuration

The second step, the obtained feature file was split into two parts, and experimented using Weka software for machine learning and classification.

At the same time, for the statistics features, this experiment also carried out feature selection. At the beginning of the feature selection. statistics features has been selected five alternative features for test. The features ware used MATLAB to extract individually and combined with each others or IS09 features when needed.

Then, the classifier running setting is set as 5 times 2-fold cross-validation same as previous work [200]. Finally, the prediction results obtained with the actual results were combined with confusion matrix accuracy of analysis to get the overall accuracy.

At last, all of the features were used to do machine learning in Weka for prediction and got the result for comparison.

3.4.3 Performance Measurement

The performance is measured based on the classification rate of the prediction on the testing database. The accuracy is defined by the following Equation 3.8 [206] (prediction positive(P), prediction negative(N), true positive (TP), true negative (TN), accuracy (A)):

$$A = \frac{TP+TN}{P+N} \tag{3.8}$$

a. Individual Feature Experiments

Firstly, IS09 features, STAT and other 4 kinds of statistic features which mentioned previous, are tested individually by machine learning methods kNNs, NB and RF classifiers.

The recognition accuracy is shown in Table 3.1. By comparing their results, it can be found that the RF method much higher than other two classifiers in each feature selected. For example, the result of RF working on the STAT features higher 10% than other two. Therefor the RF was selected as the proposed machine learning method and for the features, 6 individual features are compared in this table. The IS09 features shows the best result as 77.6% and for statistic features, STAT features provide the highest accuracy as

68.6%. Therefor the IS09 and the STAT features were selected to combine for better accuracy.

Table 3.1: Accuracy of individual features and Combined features using three different machine learning methods.

	Features	kNNs(%)	NB(%)	RF(%)
Individual features	IS09	68.3	73.3	**77.6**
	PSD	46.6	51.6	50.9
	Hjorth	44.1	47.9	46.1
	FD	45.9	44.9	42.6
	HOC	56.4	52.4	59.9
	STAT	54.6	58.6	**68.6**
Combined features	IS09+PSD	65.8	73.1	76.3
	IS09+Hjorth	64.3	72.8	75.8
	IS09+FD	64.8	72.8	76.8
	IS09+HOC	65.3	72.3	76.6
	All features	70.3	73.6	81.3
	IS09+STAT	70.8	74.1	**83.8**

b. Combined Feature Experiments

Table 3.1 also shows the comparison between the results of the Combination of statistic features and INTERSPEECH 2009 Emotion Challenge features (IS09) reveals that the best of other features, which is STAT. This may be due to the statistic feature based signal characteristics that maximize the preservation of the original audio signal information. While the other statistic features in the extraction of the loss of the key information. And for the combination part, the best result due by the IS09 combine with STAT, which is alliance between giants and shown the highest in combination features as 83.8%. The reason for all features combination is not very good maybe is the useless features from other features effect the final result.

For testing the performance of IS09 audio features and combination features, an experiment base on APM database for both IS09 features and openSMILE standard features (containing 6552 Low-Level features), which were used in previous work [200], has been done. Table 3.2. show the results of both two kind of features working alone and combine with STAT. As can be seen from the results, the IS09 features has highly effective than the standard one, and the STAT features can improve both of them.

For development proving the STAT features work, an experiment based on another database has taken in the same time. The DEAM database [207] is a new database which contains the three databases using in MediaEval challenge tasks from 2013 to 2015 for emotion analysis on music. The details of the database can be found in [207]. There are total 1802 songs shaped in 45 seconds and the songs are given with the A-V standard deviation label. For the data pretreatment, according to the A-V label of the database, four different basic emotions(angry, happy, relax and sad) have been labeled for the song based on four quadrants of A-V coordinate System. Because of that there only are 200 sad and relax songs in database, 800 songs, with 200 songs for emotions for balance of the weight of training samples, have been random selected.

Then an experiment which used the IS09 and STAT features with RF method have been done. The results of the test can be found in Table 3.2. It shows that IS09 still works well in this new database and the STAT feature can also effectively improve the performance.

Table 3.2: Comparison between IS09 Combination features and standard openSMILE features in APM database and DEAM database with RF method.

Database	Features	Accuracy(%)
APM Database	Standard	56.6
	IS09	**77.6**
	Standard+STAT	61.1
	IS09+STAT	**83.8**
DEAM Database	IS09	68.1
	IS09+STAT	**70.0**

c. Parameter Chosen

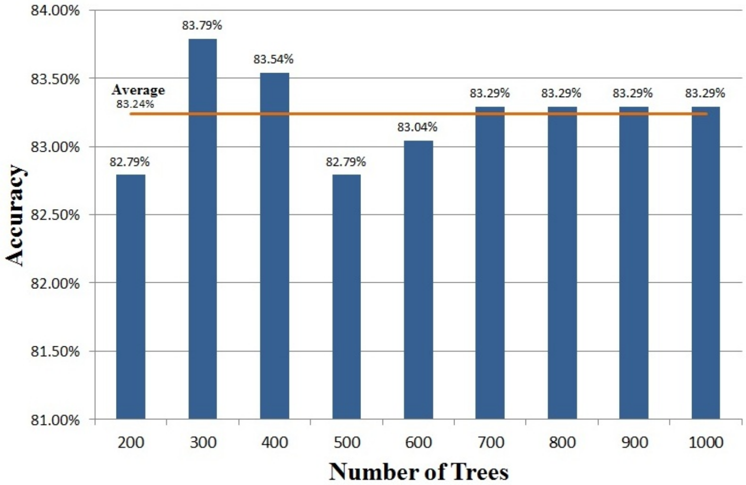


Figure 3.5: The performance of RF with different tree numbers setting.

When the features were sent in to Weka, RF was selected as a machine learning classifier. At the beginning of this section, the number of trees in the RF is set to the default value as 100, and the default numfeature value is set to 0. Then the parameter chosen testing started. Numfeature value has been set as a number by experienced as 200 as usual. So the value of the tree as a test item for a number of tests and found the best set value. From the Table 3.3 and Figure 3.5, it can be seen that the best results the experiment got is the accuracy for the combination feature with IS09 and STAT when tree number set to 300 as 83.8%. So, decision trees are set to 300 and the numFeature(the number of attributes to be used in random selection) is changed to 200 as normal.

Table 3.3: The results of using RF in different tree numbers setting.

Tree number	200	300	400	500	600	700	800	900	1000
Accuracy (%)	82.8	**83.8**	83.5	82.8	83.0	83.3	83.3	83.3	83.3

d. Final Results Comparison with Existing Results

After the completing the feature extraction, parameter setting and machine learning, the final results of experiment are illustrated in the Table 3.4. In this work, only a few of these audio features were selected and the accuracy was 76.3%. Two channels have got the same performance although the confusion matrix are different. The combined features from two channels makes the accuracy of 77.6%. Although the STAT did not perform well with 68.6% only, the combined performance is 83.8% that is better than all the results on this database.

Table 3.4: Comparison with existing results on APM database.

Method	Feature	Accuracy (%)
SVM with RBF kernel [200]	Standard	74.9
	Standard+PCMP	77.4
	GSV+PCMP +multiple candidates +optimal path	80.2
Proposed method	IS09 (left channel)	76.3
	IS09 (right channel)	76.8
	IS09 (2 channels)	77.6
	STAT(2 channels)	68.6
	IS09 (2 channels)+STAT	**83.8**

The results of comparison with previous work [200] are illustrated in Table 3.4. In previous work on APM database, OpenSMILE was used to extract 6552 Standard Low Level Emotional Features(Standard) with an accurate rate of 74.9%. The detailed results are illustrated in Table 3.4 as follow. Further, PCMP feature [208] was used in combining with Standard Low Level Emotional Features(Standard) form OpenSMILE and the accuracy is improved to 77.4%. In addition, he used multiple techniques (GSV+PCMP+multiple candidates+optimal path) to get the best accuracy rate of 80.2%.

Although features from statistics property are not very good as single, its compensation in feature space to the audio feature has made contributions on the performance improvement. The result of the accuracy increases by 3.6% and the selected features in this system achieved better performance. The proposed system achieved the best results for the overall performance.

3.5 Summary

In this chapter, an automatically musical emotion recognition system was proposed base on the new features combination by IS09 selected traditional feature and selected STAT statistic features and RF is chosen as the machine learning method according to the data and methods comparison test. 16 LLD include 5 kinds of features and 7 different characteristics features has been selected and combined from the APM database. The music signals split by 2 channels before extracting from the music signal. Then, the 768 dimensions IS09 LLD features and 14 STAT statistics features were extracted from the signal and combined for the machine learning. For the machine learning, considering the advantage of the decision tree kind machine learning methods, RF methods was selected for this system. The selected features combination classed by the RF under 300 tree numbers setting for the final classification predictions. The comparison result with the existing work can be seen from Table 3.4,, the same use OpenSMILE to do feature extraction, the feature selection one (IS09) exceed all standard feature by 2%. After the combining with statistics features, the IS09 + STAT combination features produce the best results over performance of previous system [200] ratio higher 3. 2% with RF machine method. This shows that this approach of feature selection and machine learning can indeed improve the accuracy of musical emotion recognition and properly setting the reasonable parameters will effectively improve the accuracy.

In addition, in this chapter, the research was shown that the machine learning model RF model was chosen as the system classifier. Since the classification target of the APM database is four different types of discrete, the RF model as a high-precision classifier in the classical model is very effective for this classification task with a large amount of input variable data. Because the RF model can make a weighted assessment of each variable when making classification decisions. Thereby selecting the most likely classification category. It can be seen from the comparison of the research in chapter 3. The effect of

RF in this kind of emotion recognition task is much higher than the other two. Because of the experimental discovery on the algorithm model, the machine learning model method became next research goal after completing the research on data characteristics.

Chapter 4

Emotion Detection Based on Dynamic Deep Learning Models

4.1 Introduction

The music signals has 2 different domains features and information, just like Figure 4.1, the time domain and the frequency domain. Time domain and frequency domain is the most common and most effective approach to illustrate the property of an audio signal.

The time dimension is a dimension that exists in the world, and the amount of information it contains is enormous. Every event or behavior that is happening will intersect with it. This also means that a time dimension analysis of the music signal can yield a lot of useful information. As a typical time-series signal, music signal only studies its information in the frequency domain will lose useful resources for dressing up [66]. Moreover, because of the close relationship between music and music, analyzing its continuous dynamic characteristics is very helpful to explain the information carried by the signal. Therefore, researching and analyzing the information in the time domain of the signal will be very helpful for improving the emotional recognition of music.

Frequency domain, especially in radio and communications systems use more high-speed digital applications will encounter in the frequency domain. The most important properties of the frequency domain is: It's not true, but a mathematical construct. Time Domain is the domain only exists objectively, and the frequency domain is a follow specific rules of mathematics category.

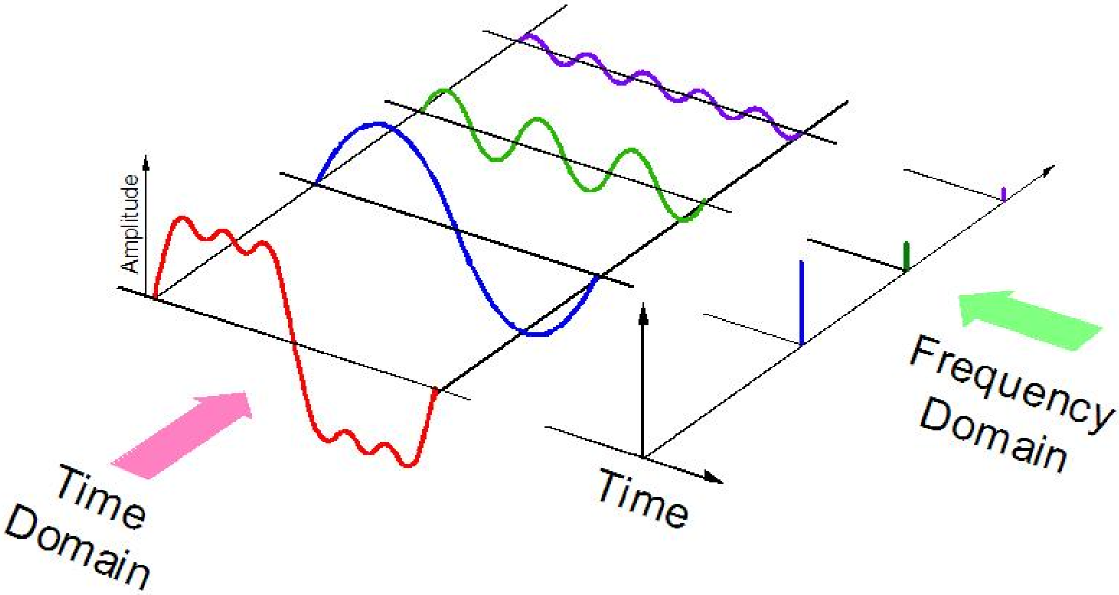


Figure 4.1: Time and frequency domain[11].

In the previous chapter, it is provided a effective features for music signal and the most significant contribution is for the frequency part. For considered the property of the music signal, the single domain analysis is not satisfied enough as a research. The time continuity analysis should be set as the next research point for the project aim.

After done the literature review for the currently effective and advanced methods for research the dynamic continuity signal, one of the most popular and reasonable deep learning method – LSTM [169] was been chosen as the dynamic deep learning model for the music emotion detection.

4.2 Related Works

For improvement of the recognition affection, the single frequency domain thinking is not enough to cover all the information the audio signal contained. The time domain and the relationship between two neighboring frames should be consider as well. As a emotional music, the single frame is meaningless when it is alone. But if large amount of it come together and arrangement according to different time, it became a dynamic continuously data and it will show us the total different meaning of the frames.

For providing the dynamic feature of the signal, a dynamic model is necessary for research. A dynamic model is a mathematical expression that describes the law of changes in system quantities over time. Generally it is expressed by differential equations or difference equations. Describe system characteristics related to the time and sequence of operations, events that affect changes, sequences of events, the environment of events, and the organization of events. It is generally expressed in mathematical equations containing continuous or discrete time variables.

For using the dynamic model to solve the continuously signals process thinking, the researchers in all over the world build lots of model and use them in their experiment research.

In March 2017, Haytham M. Fayek et al. [209] use the deep multi-layered neural networkcomposed of several fully-connected, convolutional or recurrent layers ingests a target frame (solid line) concatenated with a number of context frames (dotted line)to predict the posterior class probabilities corresponding to the target frame. He build a SER system for emotion recognition on Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [210]. In this system, Fayek use the DNN and ConvNet to process the number of concatenated frames jointly to predict a class label and use the LSTM-RNNs as the dynamic classifier of the multi-layer to deal with the different length context frames of speech signals. The result shown that the system on frame-based got the 64. 78% accuracy which is around 12% higher than the SVM and single DNN.

In the October of 2017, Ye Ma et al. [211], who does the research in Tsinghua University, use the Deep Bidirectional LSTM (DBLSTM) based on multi-scale fusion with the LLD features which extract by Opensmile on the Medieval 2015 challenge database [164] for the motion in the music task. His work considered that the dynamic continuous emotion information in music is because of the a clip music but not the moment of the time. So he use the LSTM model as the basic method, then use the propose the Multi-scale Context based Attention (MCA) for link the information of the previous and the later music by weights. As the result, his system shown the best performance, which is 0.285 in valance and 0.225 in arousal of the Root Mean Square Error (RMSE).

In November 2017, Stefano Pini et al. [212] gave people a example which is using the audio emotion recognition for cooperate working with the facial emotion recognition to predict the video emotions based on The Acted Facial Expressions in the Wild (AFEW) database [213] (2017 edition), The Facial Expression Recognition 2013 (FER-2013) database [214] and The eNTERFACE database[215]. In his research, he apply LSTM networks in order to capture the temporal evolution of the audio features in audio part. He made use of a LSTM network to learn the temporal dependencies between consecutive audio snippets, and represent them with the last hidden state of the network. The final result illustrate that his performance achieve an accuracy of 50. 39% and 49. 92% which is much higher than the baselines (38. 81% and 40. 47%).

And in 2018 Panagiotis Tzirakis et al. [216] from Imperial College show us how to use the CNN [217] and LSTM to build a model for recognizing the emotions from the RECOLA speech database [218]. In this model , Tzirakis extract features from the database by a 3 layers CNN. Then a LSTM two layers network is working for the predictions in order to model a temporal dynamics in the data, as to consider the contextual information in the data. . As the result, the performance shown that the system got the 0.787 in arousal and 0.440 for valance in Concordance Correlation Coefficient (CCC) which is higher than other methods.

It can be found from the related work in recent few years, that the LSTM is become very popular in audio emotion recognition area. Because it has the effective provide the dynamic features of the continuously signals. It applies a very significant relation and features between two following frame, which can be used to improve the system performance from the single static classifier or features.

4.3 LSTM Based Emotion Detection System

In this research, the deep learning method, which are the a kind of advanced learning methods, is used for MER. For looking for the approach to improve the performance of MER, dynamic deep learning method is selected as the proposed method. After testing the image based deep learning method – CNN, the accuracy is only half of the previous RF method. It is believed that music signals are continuity and there are a lot of relation and information during music running with time. It is dynamic. Therefore, it is selected the LSTM as the main model of the experiment system. It is used the LSTM on the APM database with the selected pre-processed features for research for the relation between different time sequence. The research provides the performance and the testing of the LSTM working on the APM database which can prove the applicability of using LSTM on this database.

4.3.1 System Overview

In this system, the classic and statistical static audio features has been extract from the original AMP database before provide as the input data to the LSTM. Then, the LSTM is using as a dynamic classifier for predict the emotion label for each song. After compare the predictions and the true labels, the performance will be calculated.

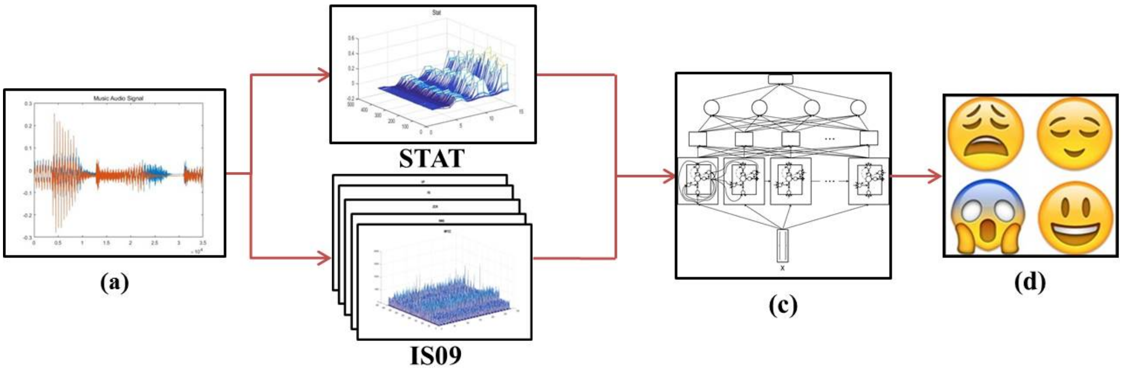


Figure 4.2: LSTM based emotion detection system from music signals. (a) Music signal, (b) Feature selection, (c) LSTM model, (d) Emotion classification.

4.3.2 Feature Extraction

a. IS09 Feature

As an affective feature bundle, Interspeech 2009 conference provided some classic and useful features. There are 384 features selected. In the following, these features will be described briefly. The detailed information can be found in [184]. Then the configuration of the INTERSPEECH 2009 has been used in the features extraction open-source toolkit: OpenSMILE [201] for the features of INTERSPEECH 2009 which named as IS09, which is used as main features for the low level features part.

As illustrated in Figure 4.3, the 16 LLD include 5 kinds of features which will shown follow. RMS Energy [202] is based on the amplitude of the peak value of the signal to calculate the power of a signal, which shows the energy carried by the signal. It is one of the common audio feature representation. In MFCCs [198], Mel frequency is a major concern for the human hearing characteristic frequency. Because humans often unconsciously for a certain period of frequency is very sensitive, if the analysis of such frequency will greatly improve accuracy. MFCCs is a linear mapping of the audio spectrum to Mel spectrum, then the conversion factor cepstral frequency domain when available. ZCR of a signal means the ratio that the signal cross from one end of the axis 0 to the other side of the axis. It is usually expressed by a signal of a predetermined value of 0, or as a filter are searching for a rapidly changing signals. The F0 [184] means the fundamental frequency, in a complex wave is a periodic waveform the lowest sum of frequency. In music, it is usually the lowest pitch notes. The VP means that the voicing probability computed from the ACF.

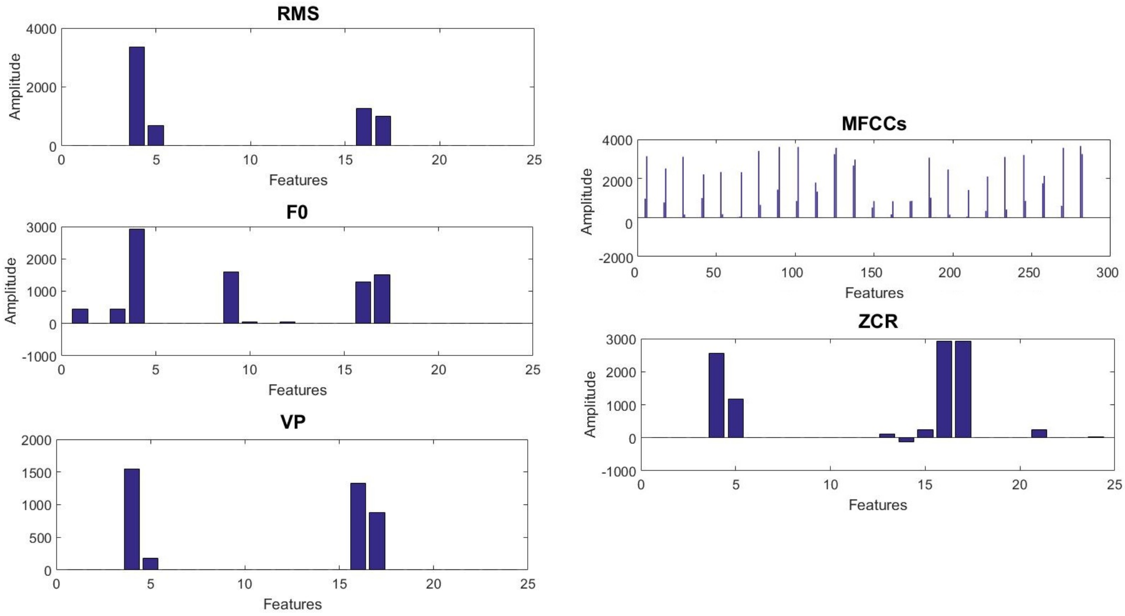


Figure 4.3: IS09 features extracted from APM database.

b. STAT Feature

Through the reviews of the related work, it found that most of the research much more focus on the more representative and more specific features of the signals. But the common and macro features also can contribute their usefulness for improve the recognition. According to [44], the basic statistics features (STAT) of EEG signals works well in his work. So after considered the similar between EEG signal and an audio signal, these characteristics as the basic statistical characteristics of the signal in the time domain.

For the statistical characteristics in STAT, which shown in Figure 4.4, there are 7 different characteristics features shown as follow: Power, Mean, Standard deviation, First difference, Normalized first difference, Second difference and Normalized second difference.

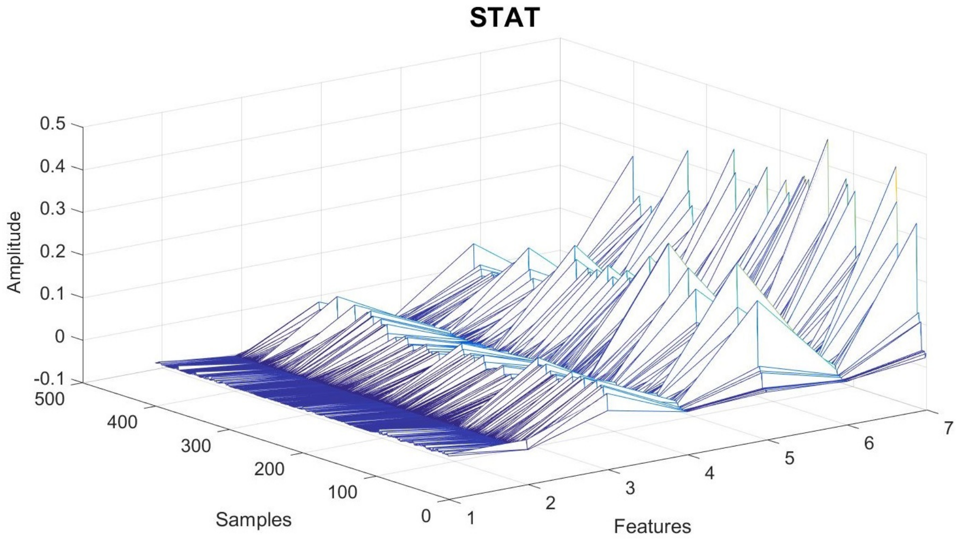


Figure 4.4: Statistic Feature for all the 400 music sample in APM database.

4.3.3 Long Shot-Term Memory (LSTM) Modelling

LSTM as the core part of the experiment system is a very effective deep learning method for the dynamic continuity signal. LSTM is a RNN and suitable for processing and prediction interval time series of important events and very long delays. LSTM is containing LSTM block (blocks) or other kind of NN, document or other materials LSTM blocks may be described as intelligent network unit, because it can memorize the indefinite length of time values, block there is a gate input to decide whether to be important to keep in mind and cannot be exported output [219].

In simple terms, for the key point, there are four function unit form right side (input) to left side (output) in LSTM layer. According to the left-most function block input situation may become, it will go through the gate on the right three input determines whether the incoming blocks left for the second input gate, if there approximate output zero, where the value of the block will not advance to the next level. The third is left forget gate, when it produces a value near zero, there will remember the value of the block to forget. The fourth is the rightmost input to output gate, he can decide whether to input in the memory block can be output. The most important thing is LSTM Forget gate, followed by Input gate, most times is Output gate. In order to minimize the training error, Gradient descent and Backpropagation Through Time (BPTT), it can be used to modify each based on the weight of the error.

For the details of LSTM model, it can be found from the Figure 4.5. It gives the over view of the LSTM model. The network structure of this LSTM model is divided into five layers of networks. The first layer is the sequence input layer, and the music signals of a single vector will be entered into the network here and each signal segment is input to the input layer in chronological order.

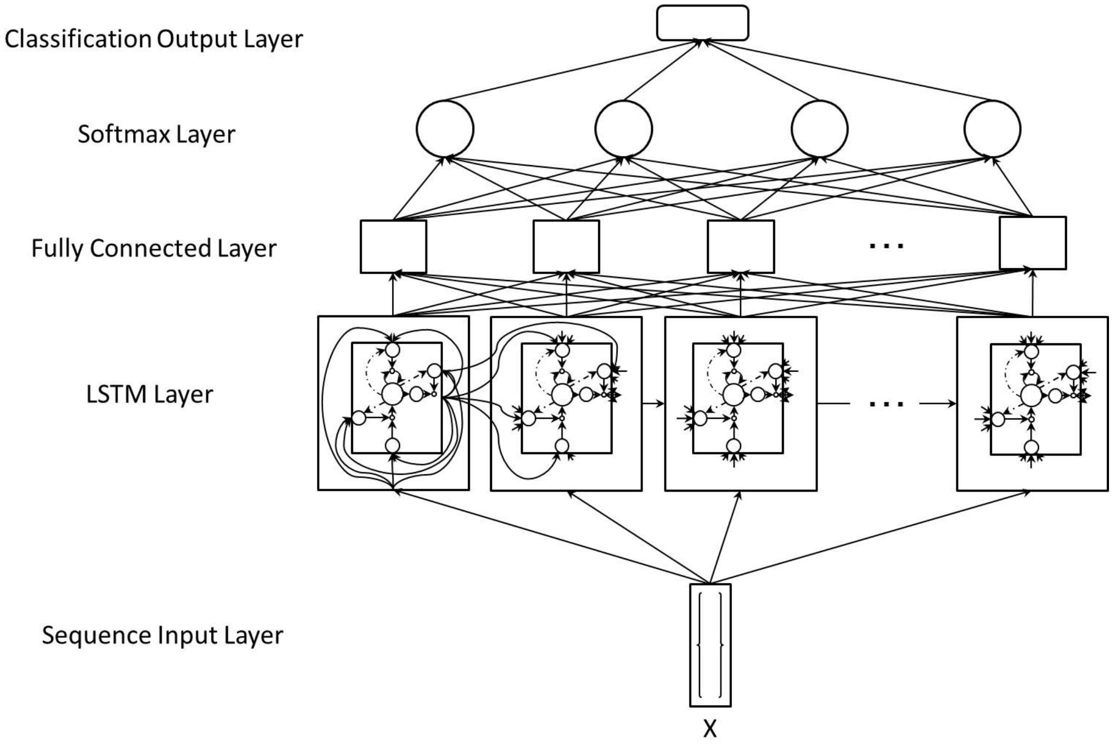


Figure 4.5: LSTM model.

The second layer is the LSTM layer and the core layer of the entire LSTM network. This layer consists of several LSTM hidden units. Each hidden unit is a repeating module. The hidden cell module structure that has been enlarged in detail can be seen from the Figure 4.6. This module achieves a filtering process for filtering useful information by discarding and adding input information. Next, this article will explain how each individual module updates and records information.

LSTM replaces the underlying layers of a traditional RNN model with a unique cell containing three different gates. In traditional RNN, the nonlinear activation function used by the model which may be just tanh or ReLU. But LSTM replaced them with such unique and complex structures, giving the entire network the ability to screen, remember and store[12].

LSTM is also a structure like all of the RNN kind model, but duplicate modules have a different structure. Which can be seen from Figure 4.8, there are four function in the bottom of the middle block. Each of them is one gate for controlling the information.

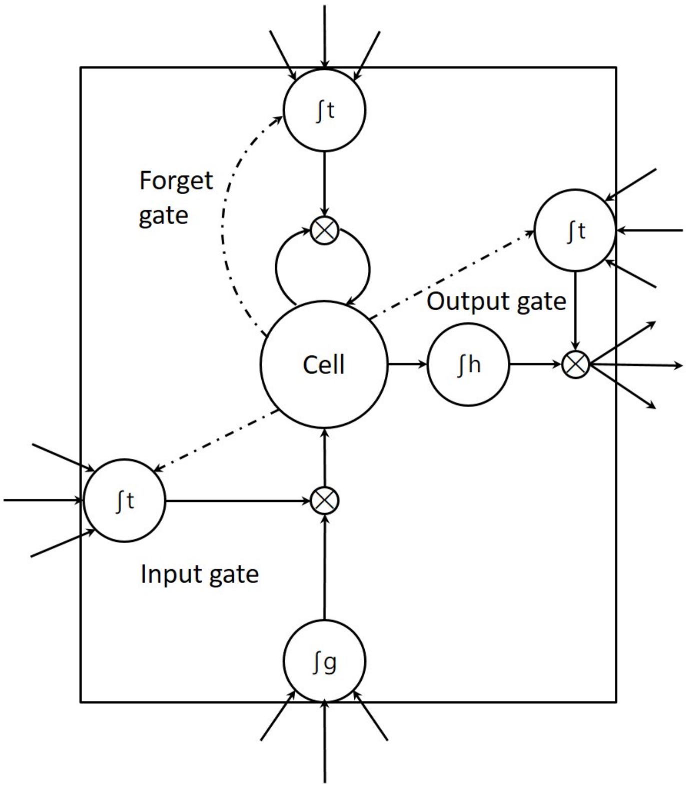


Figure 4.6: LSTM hidden unit.

In Figure 4.10, it illustrates that the main state of the information in the LSTM cell. The stat from the input timing C_{t-1} go through whole cell block to the output timing C_t. During this progress, the LSTM will change the stat of the cell based on the information controlled by the gates.

LSTM has the ability to remove or add information to the state of the cell through a well-designed structure called “gate”. A gate is an approach to make information choices pass. They contain a Sigmoid function layer (Figure 4.11) and a pointwise multiplication operation. The Sigmoid function controlled the information passing or stopping. It can block the input information go through or not based on the value it output.

The LSTM has three gates based on the Sigmoid function layer to protect and control cell status, which shown in Figure 4.6. They are input gate, forget gate and out put gate. Foget gate will check the information which inputs to the cell state. The input gate will protect the checkpoint of update the cell state. Output gate will chose the information if it is correct for the prediction.

In the forget gate, which is shown in Figure 4.12, the gate will filter the information from

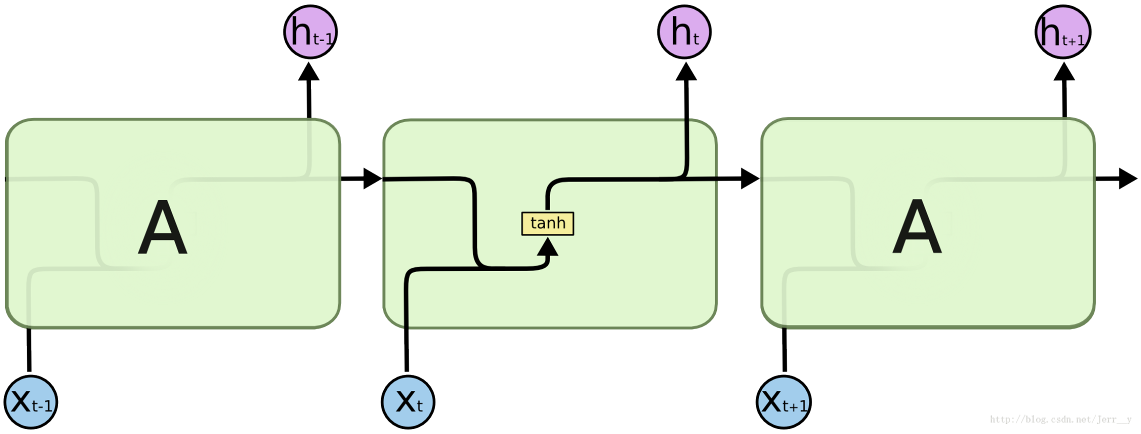


Figure 4.7: Normal RNN structure (The A block means the same structure unit in different time sequence)[12].

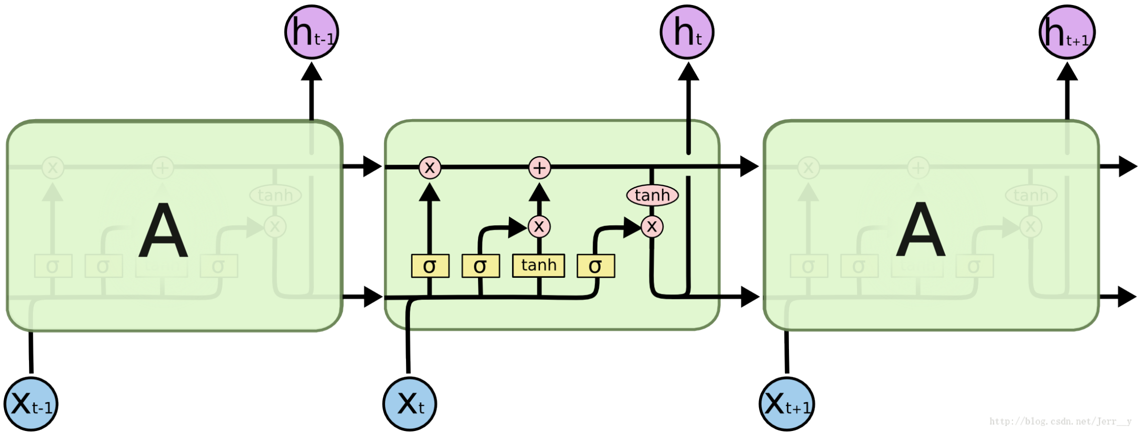


Figure 4.8: LSTM structure[12].

the last timing cell state. If the information is still useful for the system, it will be keep, either it will be removed from the cell. The gate makes the decision based on h_{t-1} and x_t and the outputs a value between 0 and 1 for each number in the cell state C_{t-1}. The value 1 will keep the information, or 0 will remove them.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{4.1}$$

The next part is the input gate's work. The tanh function in this gate will provide the new information which input from the new timing $\widetilde{C}_t$. But not all of them will update to the cell state. Another part of the gate, Sigmoid function will decide the final result: which information can smooth entry the cell i_t. Therefore, the input gate can chose the information which will add to the cell.

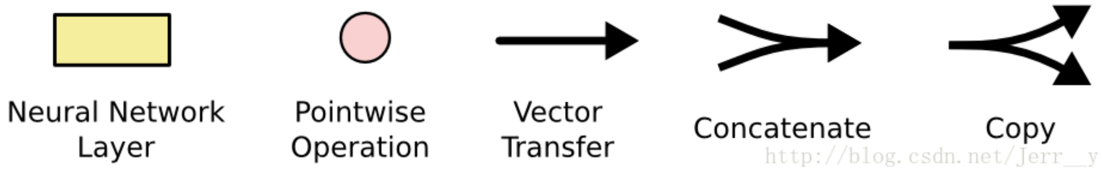


Figure 4.9: Icon explanation[12].

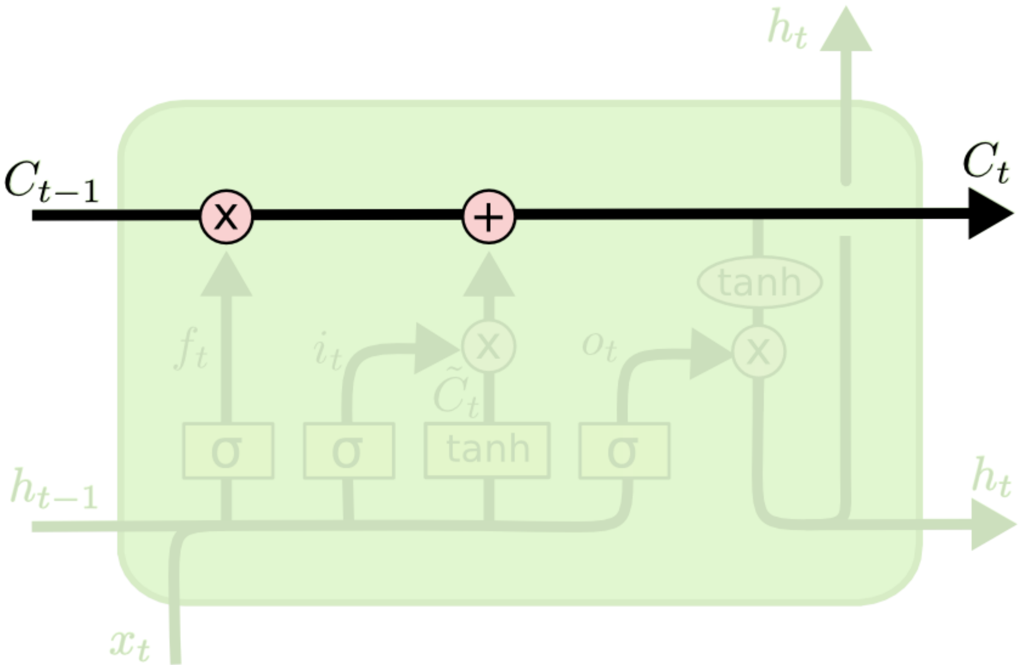


Figure 4.10: The cell statement in the LSTM[12].

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{4.2}$$

$$\widetilde{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_C) \tag{4.3}$$

Then, the cell state will update based on the output of forget gate and output of input gate. C_{t-1} will be updated to C_t (Figure 4.14). the previous state will be multiplied by f_t and discard the information which is determined to be discarded. Then add $i_t \times \widetilde{C}_t$. This is the new candidate value, which varies according to the extent to which decide to update each state.

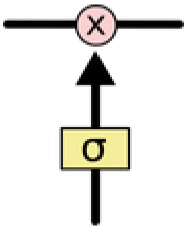


Figure 4.11: The sigmoid layer[12].

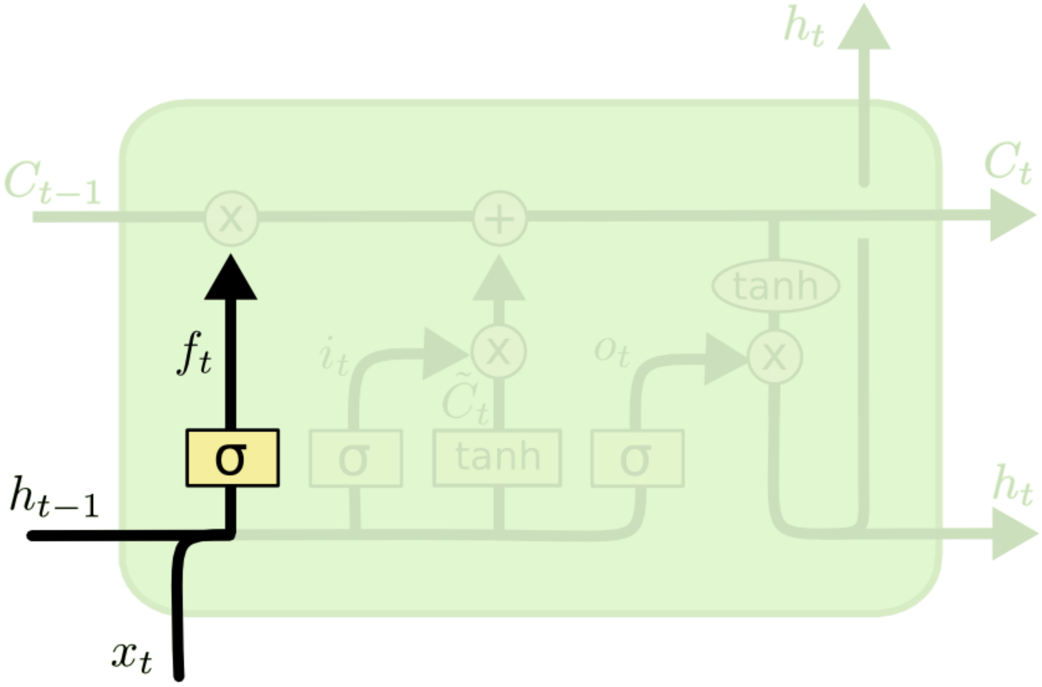


Figure 4.12: The forget gate[12].

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C}_t \tag{4.4}$$

At last, which can be found in Figure 4.15, the output gate will finally decide the information which can be outputted from the layer. Based on the C_t which is updated by the input gate, a Sigmoid function will cooperate with a tanh function to decide the final output of this timing cell state. The Sigmoid function will still filter the information like it did before and the tanh function will normalize the information from the cell state for combine with the output of Sigmoid function. At the end, the result will be output and moving to the cell state in next timing.

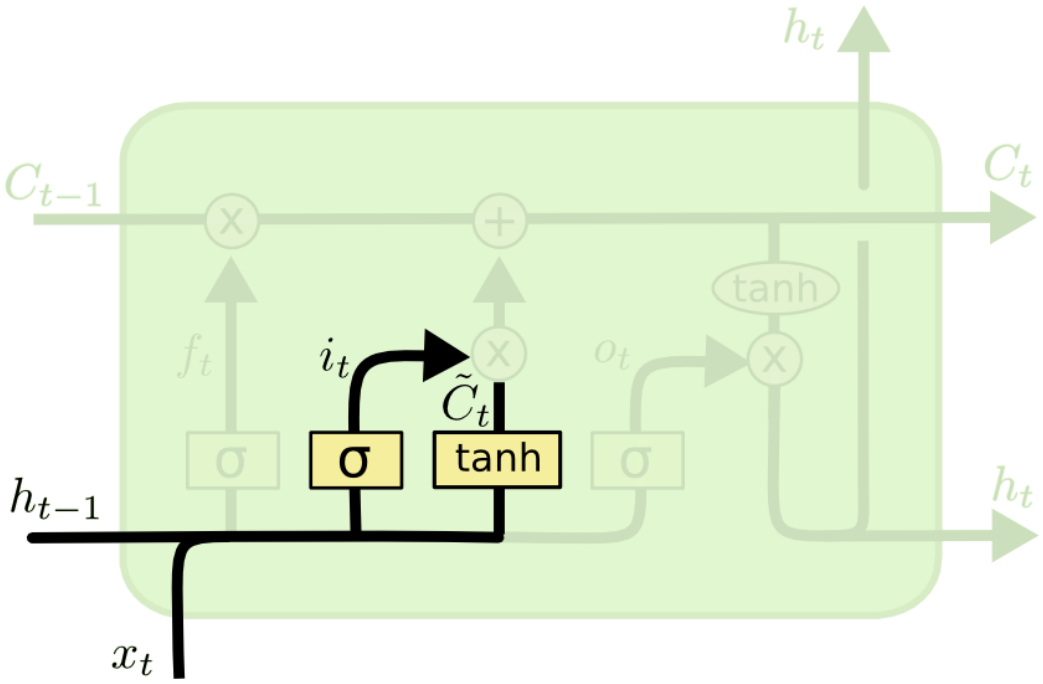


Figure 4.13: The input gate.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{4.5}$$

$$h_t = o_t \cdot tanh(C_t) \tag{4.6}$$

When the data has been processed by the LSTM layer, there will be an fully connected layer for each hidden layer. The size of the fully connected layer is same as the number of the classes.

Then, each fully connected layer send the output to the softmax layer. In this layer, each input will be calculate the probability to the each classes.

Finally, based on the comparison of the probability for each prediction result in the classification output layer, a final unique prediction result is finally obtained. After comparing this prediction with the true label, the LSTM network will re-weight each previous step according to the true label. The system hopes to get the same result as the true label after the input passes through the network with the new weight. When the system gets the optimal weighting factor, the entire network will stabilize. This final weighting factor is the basis for future prediction of the target data. So the whole LSTM model is like this.

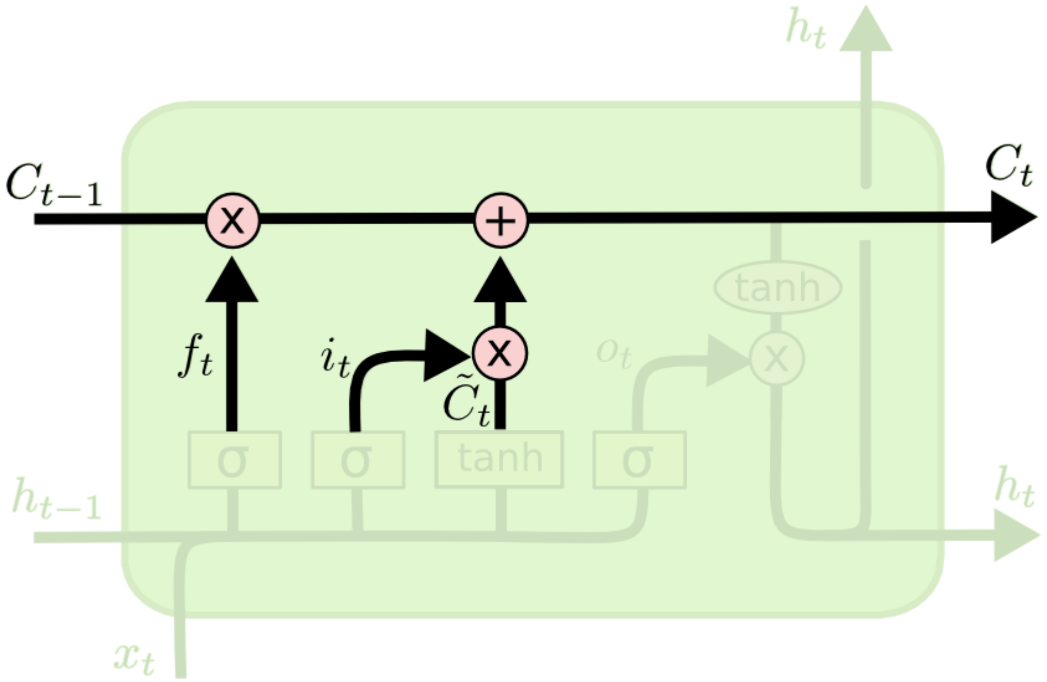


Figure 4.14: The update of cell state [12].

4.3.4 Parameter Setting

- Max Epoch

So the first test is for the max epochs. An epoch refers to the process of all data being sent to the network to perform a forward calculation and back propagation. Since an epoch is often too large for a computer to load, it is divided into several smaller batches. It is not enough to iterative train all the data during training, and it takes multiple iterations to fit the convergence [220]. In actual training, it is divided all the data into several batches, each time sending a part of the data, the gradient descent itself is an iterative process, so a single epoch update weight is not enough.

As the number of epoch increases, the number of iterations of the weight update in the NN increases, and the result gradually enters the optimal fitting state from the initial unfitting state, and finally enters the over-fitting.

Therefore, the number of epoch is very important. What is the right setting? There is no definitive answer. The number of epoch is different for different databases.

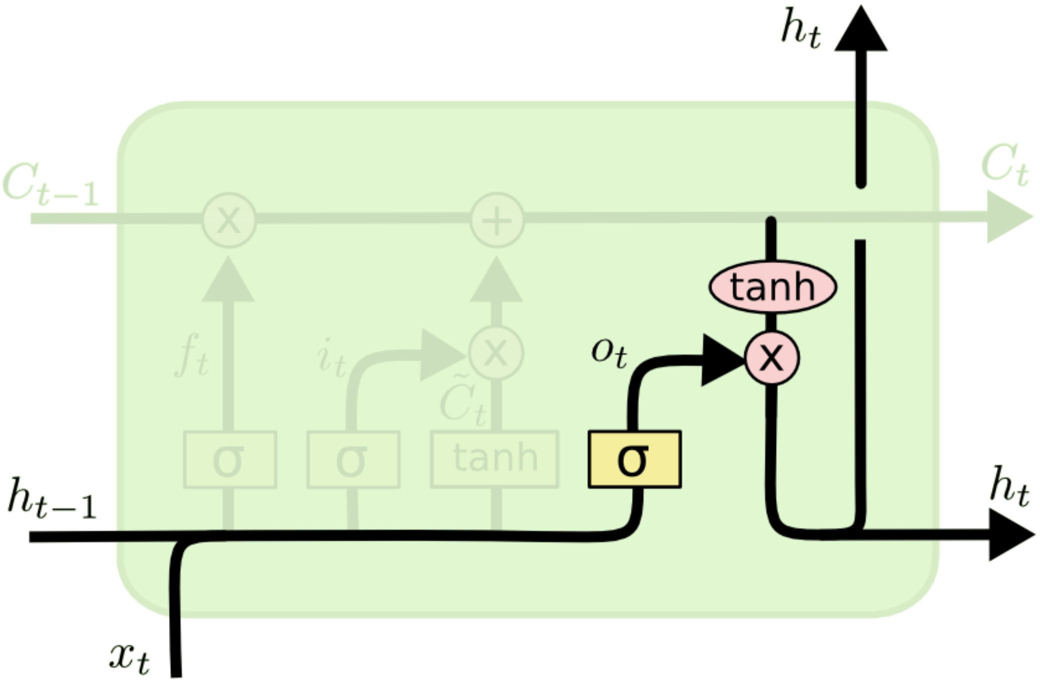


Figure 4.15: The output gate [12].

However, the size of the epoch is related to the degree of diversity of the database. The more diversified, the larger the epoch should be.

- Output Size

Then the experiment move to the output size parameter. The output size is stand to the amount of the Hidden unit size as known as the RNN cell size in LSTM.

- Mini Batch Size

The third part of test is the significant one – the mini batch size. Batch is a part of the data that is sent to the network every time, and used to evaluate the gradient of the loss function and update the weights. Batch Size is the number of training samples in each batch. The choice of Batch size is also crucial. In order to find the best balance between memory efficiency and memory capacity, the batch size should be carefully set to optimize the performance and speed of the network model. When the amount of data is small and the computer performance can be loaded, it is better to use one batch. Mini batch training accuracy is slightly lost, but greatly improved performance. Random training runs fastest, but it is difficult to reach convergence.

- Shuffle
  The shuffle has some kinds of options in LSTM working process [221].

  1. 'once' Shuffle the training and validation data once before training.
  2. 'never' Do not shuffle the data.
  3. 'every-epoch' Shuffle the training data before each training epoch, and shuffle the validation data

- Learning Rate
  At last, the test comes to the learning rate part, the significant one. The correctness of the learning rate also has a great impact on the final outcome of the LSTM. As the learning rate increases, the faster the gradient adjusts the weight of the network, the faster the gradient falls, and the shorter the convergence time, and vice versa.

4.4 Experiment

4.4.1 Data Preparation

"APM Music" is the largest production music library in the industry as a production music company. It contains more than 40 libraries, more than 475, 000 tracks and CDs and he almost contains almost every type and style of music. Therefore, this article chose the APM database [200], which is based on this huge music data, as this experiment database.

In order to make sure the effect of the features during the real situation, an MER experiment based on APM database be started. APM music database is a music data in .wav format. The database including 4 kinds of emotion: happy, sad, relax and fear. Each of the emotion has 100 audio samples with around 35 seconds long.

After researching and learning the basic principle of the LSTM deep learning method, the first things need to be done is the music signal preprocessing for the LSTM input data. According to the original signals of the APM database is too long to process as a input, the signals must be split to the small segment for the experiment. So the 44100 sampling point per second music signals was segment to the 1 second smell pieces.

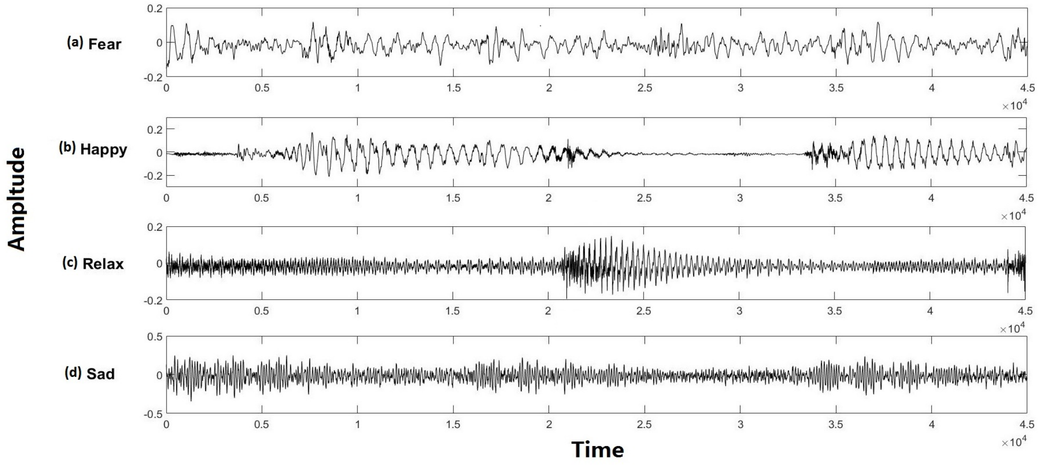


Figure 4.16: Four audio signals for 4 different emotions. (take continuous 44, 100 sampling points from the music recording for approximately one second).

For test the correction of the LSTM program based in Matlab, an experiment used the default setting, parameters and basic 12 features witch extracted from APM database has began. When the LSTM program was started working, the input data split by test and train and prepare the data to the cell type. Before running in the LSTM for the performance, some parameter should be setting.

4.4.2 Experimental Results

The first step of experiment is to define the correction of the LSTM network architecture in Matlab and the applicability of the model for the APM database. The preset database, Japanese Vowels database [222], is used for the correction test.

Based on the LSTM network introduced before, the input size is need to decide first for the input layer building. According to the dimension of input features from Japanese Vowels database, the input size will be sequenced as 12, which is the size of features sent into sequence input layer each time.

Then, the output size is setting to 100 as known as the number of the LSTM units in LSTM layer. This value based on the experience.And, because the principle of the LSTM layer working. The last element in the sequence is set as output parameter.

Next, the class of the Japanese Vowels database is 9. So the fully connected layer size is decided as 9. Which is used for mapping distributed features obtained from network training to sample mark spaces.

At last, the The softmax layer is normalized the output of the upper layer, fully connected layer, for the sample probability vector of each kind which is obtained. The classification layer will output the final prediction for each sample based on the probability.

But before running the network, there are some training options should be specified. Value 27 is input as mini-batch size which controls the speed and the extent of convergence and means the number every time running. Then set the maximum number of epochs to 100 and specify to not shuffle the data. At the same time, the learning rate has been setting as the default value 0.001.

When all options settings are finished, the network can be run for the 270 sequences train database for network training. After the network full training by the train data, the test data runs in same setting for final prediction. Finally, the accuracy is calculated.

According to the test experiment for LSTM network, the network building is useful and working well in audio dynamic sequence signals.
The experiment moves to the APM database.After using the default setting which used in Japanese Vowels database, the base lane results of APM database which is classed by LSTM shown that the LSTM model was working. Even the performance is terrible. But it is usable.

When the program is ready for work, the first task for the experiment is how long the segment should be split as for the rest of the work. The total length of music in APM database is from 30 seconds to 60 second, which means it is got a lot of choice in the length of segment. Then a testing has been set for confirm the length of the segment. There were 4 kind of segment prepare for the test: the one second segment, three second segment, half length of signals and the original signals which will use 3 traditional machine learning methods for performance . This test aim to make sure if its shorter or longer is better. It can be seen form the Table 4.1, that the result shown that the original one got the best performance and the 1 second segments worst, Which means that the longer segment will provide the better result. Considering the dynamic and continuity property of the signals and model, if the segment is too long, the dynamic information between the segment will lost too much. The three seconds split methods has been selected as the default one.

The second purpose of input preprocessed data is features selection. In this step, lots of features has been tested. Table 4.2 gives the result of the comparison of all the features which was tested. The first turn comparison is the basic STAT features and MFCCs which got similar perforce. Then the IS09 features which used in the previous work join into

Table 4.1: The accuracy(%) comparison between different length segments running in different learning methods.

Feature/Method	kNNs(5)	NB	RF
STAT(12 domension) 1s segment	48.30	43.38	49.27
STAT(12 domension) 3s segment	49.66	44.78	49.91
IS09 1s	49.87	48.63	55.66
IS09 3s	52.20	38.00	56.83
IS09 half length	62.69	70.65	77.61
IS09 Original	67.16	72.64	78.11

the 'battle' and which win obviously. For looking for the best features input, the IS09 original (384) features have been processed with PCA, normalization, hamming window overlap split and the functionals split. At last, the normalization three second IS90 win the final champion as the default input data.

Table 4.2: The accuracy(%) between different features and length segments on default setting running in LSTM.

Feature	1 second	3 seconds
12feature	29.35	26.37
MFCCs	29.85	32.77
IS09_384	33.83	41.29
IS09_pca12	32.84	40.30
IS09_mfcc(288)	25.87	38.81
IS09_mfcc_pca12	26.37	37.81
IS09_384_normal	42.37	**59.70**
IS09_384_window_normal	39.80	46.77

After this it is done a lot of tests for the parameter test for looking for the best suitable parameters which using in this database. There are totally 7 parameters which can be edit in the LSTM model: input size, output size, number of the class, max epoch, mini batch size, Shuffle and the learning rate. In all of these parameters, the number of class has been confirm as 4, because APM database got 4 class of emotions and the input size decided by the size of the input data. It is confirmed as well. For other parameters, several experiments have been set for the suitable value of the tested parameter on same default values on other parameters.

The first parameter got tested is max epoch. Table 4.3 illustrate that the effective is shown fluctuations with the increasing of the epoch size. From the lowest size 100 to the reasonable max size 600, the value on 300 provide the best result in this turn.

The next one is the output. It can be found in Table 4.4 that the results are similar on the

Table 4.3: The accuracy(%) test on epoch parameter.

LSTM/Epoch	100	150	200	300	400	500	600
IS09 384_normal(3sec)	55.72	59.70	60.20	**63.18**	56.72	57.71	62.69

different step values. But the best result in a span of 300 test is the 250 value output size provided.

Table 4.4: The accuracy(%) test on output parameter.

LSTM /Output	50	100	150	200	250	300	350
IS09 384_normal(3sec)	59.70	63.18	64.18	62.69	**65.67**	65.17	63.68

After output test, the most significant parameter mini batch size is been tested. Table 4.5 illustrate that the value of the mini batch size under 50 gives the similar result. But if the value over 50, the result will be fast fall. After focus test the mini batch size value 5 to 20, the best result is provided by the 20 mini batch size.

Table 4.5: The accuracy(%) test on mini batch size parameter.

LSTM /Mini batch size	5	10	15	16	17	18
IS09 384_normal(3sec)	56.22	53.73	61.94	54.17	56.22	53.48
LSTM /Mini batch size	19	20	21	50	100	200
IS09 384_normal(3sec)	53.23	**65.67**	58.21	51.24	46.27	24.38

For providing the complete train data feature in every epoch for network training, the 'Shuffle' value is set to 'never'.The results of comparison with discarding the data are shown in Table 4.6. It can be found that using the full data will be better for this database.

After finish the parameters testing, the best performance are provided by using LTSM to classify the normalized IS09 feature which extracted from the three second split segments with the following parameters which is shown in Table 4.7 and the final result also illustrate in this table as 65. 67%.

4.4.3 Comparison and Discussion

After comparing the performances with the previous work and other works using the AMP database, it can be found from Table 4.8. The final performance is not as good as other two methods even using the same features. Considered the LSTM is a dynamic deep learning methods and using the layers to feed back the information in next time point, and the music in this database have not enough obvious association, the result is worse than

Table 4.6: The accuracy(%) test on shuffle parameter.

LSTM /Shuffle	never	once	every-epoch
IS09 384_normal(3sec)	**65.67**	29.85	31.84

Table 4.7: The final accuracy(%) with parameters.

Parameter	Input	Output	Epoch
Parameter Value	384	250	300
Minibach	Learning rate	Shuffle	Numclass
20	0.001	Never	4
IS09_normal_3sec(Accuracy)	**65.67**		

other 2 previous works. Because the RF will focus more on the feature level information which provide more effective information.

Table 4.8: The accuracy(%) comparison with other works in same database.

Method	Feature	Accuracy
SVM with RBF kernel	GSV+PCMP+multiple candidates+optimal path	80.2
RF	IS09+STAT	83.8
LSTM	IS09+STAT	**65.7**

4.5 Summary

In is chapter, the segmentation music signal is used to extract the featrues in this system. Then the LSTM deep learning method was provide to classification them. After experiment for looking for the best features and parameter combination, the system look forward to use the dynamic continuous information between the different time frame segments for more effective emotion recognition.

Different from the traditional discrete classifier, the purpose of the research in chapter 4 is to improve the performance of the MER system based on the dynamic persistence characteristics of the music signal of the database, and the LSTM network model which has great advantages for information persistence and long-term dependence. Because the music signal is a long-term continuous time series data, there will be a great connection between the context of the data information. Proper use of the dynamic information of the signal will greatly help the emotional recognition task. At the beginning of the research, based on the data feature conclusions of the previous chapter, the original audio signal was selected and segmented. Because the LSTM model, as a variant model of RNN, effectively uses the context information relationship of a complete dynamic continuous signal for training. So the data must be processed as a chronological data format. In addition, due to the deep learning network model of LSTM, reasonable arrangement of various parameters in the network will greatly improve the recognition accuracy. So these two parts have become the focus of research.

A system based on the LSTM deep learning methods for the dynamic continuous signal thinking has been proposed. The system look forward to use the dynamic continuous information between the different time frame segments for more effective emotion recognition. In the signal pre-process part, the segmentation music signal is used to extract the featrues for this system. According to the LSTM network structure, the input of the data need to be split in segment based on time. The comparison experiment was taken for looking for the suitable length of the signal segment. After the IS09 and STAT extracting from the segment, the LSTM deep learning method was provide to classification them. The input features would go through the LSTM layer, fully connected layer and softmax layer for the final classification prediction. But in order to the best result, the parameter is also a significant element of deep learning network system. six most decisive parameters were tested and decided by experiment. After experiment for looking for the best features and parameter combination, the final prediction was proposed.

However, the final performance is not as good as expected. It may because the database is not suitable to the LSTM as the previous thoughts. The information between the segments is not good enough to improve the performance of recognition which compares with the RF machine learning method.

Chapter 5

Music Emotion Identification Through Chords

5.1 Introduction

After the first two chapters on the frequency domain, time domain and continuity of music signals, most of the signal's own attribute information has been basically mined. However, music as an artificially created sound signal, the information it displays must be derived from the thoughts and emotions that the creator wants to express. Just as humans speak and write, music as a tool for transmitting emotional information must containing its own as a ”language” in the law or structure to help it display the emotions that it wants to express. Contrast the statement of human speaking, if a sentence wants to express a complete meaning. He needs to have at least two of the grammatical units of the subject, predicate and object. In the same way, the music wants to effectively feel the emotion it wants to convey, at least to satisfy the complete structure of music. In these music structures, chords are known as an important role in turning a dull monosyllabic signal into a rich and colorful piece of music. The chords, like the adjectives in the declarative sentence, set up a display stand on a basic melody skeleton to show the emotions the author wants to express.

The word “chords” are from Greek, the original meaning is the string. In music theory, it refers to the sound of two or more different pitches combined together. In European classical music and the music style influenced by it, more often refers to a combination

Figure 5.1: Music and emotions [13].

of three or more pitches, and the combination of the two pitches is described by a pitch. The chords that make up the chord are called the chords of the chord. In the basic form of chords, the lowest sound is called the “root”. The remaining sounds are named according to their pitch relationship with the root note. The composition of the chords can be played separately or simultaneously. Separately played, it is called as decomposition chords (or scattered chords). The chords are three-degree superposition and non-three-degree superposition. The chords in the Western traditional harmony are formed according to the principle of three-degree superposition. There are many types of chords. If them are distinguished according to the number of constituent tones, chords can be divided into triads, sevenths, and ninths. The triad is composed of three tones, the seventh chord is composed of four tones, and the nine chord is composed of five tones. These chords can be subdivided by the interval structure. For example, the triads can be divided into major chords, minor chords, augmented chords, and diminished chords. In addition, the chords composed of the same sounds will have different sound effects and usage modes when the lowest sounds are different.

The chord, as a such significant point in music, some of the research is also already done for it. For example, there is a research which did by Heng-Tze Cheng [223] and his group-mates. Cheng and his team built a chord recognition system. He first extracted the Pitch Class Profile (PCP) feature vector as the base featues. Then the HMMs framework with the N-gram model was used based on the chord decoding methods, which is according to the musical knowledge, to extract the longest common chord subsequence and the chord histogram as the Mid-level features. At the same time, the Low level features (such as MFCCs) were extracted for combination. At last, their new features improved the accuracy by 7% in all over the application test.

For another work, in 2015, Xinquan Zhou [224] his partners is like a good link use deep learning methods to deal with the chord detecting in 2015. In his work, he did the Constant Q Transform (CQT) and PCA on original input audio signals, which from the 317 songs in 4 different databases. Then he used the time splicing and convolution to split the signals to smaller pieces and pre training. Afer this, he used the pre-processing data with bottleneck architecture to work through the Restricted Boltzmann Machines (RBMs) and Gibbs sampling. At last, the HMMs and SVM was used as the classifier to deal the final result. The final evaluation shown that accuracy of this model is more effective than others.

However, for all of these related researches, nobody focus on the relation between the chord and the emotions directly. After reading the book “Music and Emotions” [77], an ideal appeared: which not just link the chords and the emotions together since they got this close contact. This will be the first time use the chord to detect the emotion based on the music theory. Then a system of directly mapping music and emotion based on the Theory of Musical Equilibration [225] started researching.

5.2 Emotion of the Chords

The connection between music and emotion is more important than ever, and music research is increasingly focused on understanding the complex features of this interaction. After all, for a long time, the fact that music has an emotional impact on us is one of the greatest mysteries, because fundamentally it consists only of inanimate frequencies. This is a topic that it would not think of in everyday life, which is why the indescribable aura is still in the music. There seems to be some taboo about how and why music can convey feelings - and interestingly, musicians do the same.

Although people like to describe music as an international language, science still cannot provide an explanation for explaining the nature of this language. For decades, it has left the task of solving this mystery to a small group of people: Although music psychologists are equipped with statistical software and calculators, music psychologists have so far been less successful than brain research that has been widely cited in recent decades because it solves the problem of why music can stimulate emotional reactions.

Theory of Musical Equilibration [225]called Strebetendenz-Theorie in the original German is the first psychological paradigm to create an emotional effect on music. It breaks down the music sequence into one of the most important components - harmony - and uses this material directly as the basis for its argument. Harmony is essentially a concentrated form of music, because in an instant it can reflect melody and other musical processes, otherwise it can only be depicted in a given time interval. Harmony is the psychology of music emotion. The use of selected examples in the book clearly shows that the emotional characteristics of music and sound can be systematically deconstructed, and can be reasonably proved and empirically proved.

In the book Music and Emotions by Daniela and Bernd Willimek [77], they summarize the emotional mapping of each chord by letting children listen to alternative songs and let them express the feelings they feel. The study selected six different chords as research goals.

5.2.1 C Minor Chord

The root and third notes are major thirds, and the third and fifth notes are small thirds. They are represented by the uppercase English alphabetic names of the roots. For example, DO, MI, SOL and chords are represented by C, FA, LA, and DO chords are represented by F. , drop MI, SOL, drop SI, use Eb, rise FA, rise LA, rise DOL with F. Figure 5.2 shows the C minor chords in a music example segment.

By applying Theory of Musical Equilibration, it is also possible to logically explain the effects of minor chords. The Moser Music Dictionary describes the nature of minor chords because the primary shadows of minors are obscured. When interpreting this definition from the perspective of this monism, one third of the minor is not considered to be an independent interval, but rather regarded as one-third of the “cloudy”, it has been deprived Its main tension. The Ullstein Music Dictionary describes a small chord as a professional.

When applying “Theory of Musical Equilibration” to the small chords here, it will be seen a clear result if it is replaced the psychological image of Kurth’s balance effect (ie, the impulse of music resolution) by the image of the recognized thing. In a major supplement, it is determined the desire not to change the chords, but in the minor, this sense of will now appears to be overcast and suppressed. The feeling of satisfaction is overshadowed by dissatisfaction.

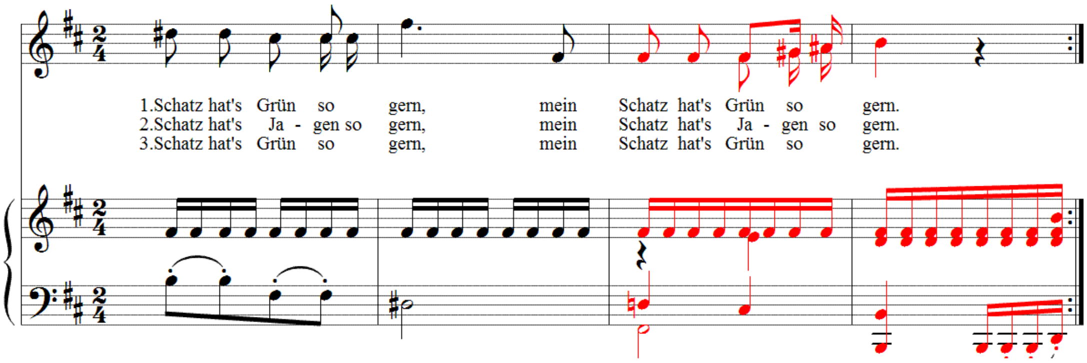

1.Schatz hat's Grün so gern, mein Schatz hat's Grün so gern.
2.Schatz hat's Ja - gen so gern, mein Schatz hat's Ja - gen so gern.
3.Schatz hat's Grün so gern, mein Schatz hat's Grün so gern.

Figure 5.2: C minor chord.

5.2.2 Minor Sixth Chord

The minor sixth chord, the first inversion of the third chord, the chord is composed of three to five major three degrees, five to the root of the pure four degrees. Figure 5.3 shows the minor sixth chord chords in a music example segment.

The example of minor 6th shows that feelings can be conveyed not only through complete chords, but also through a single interval. If minor 6th is played, the chords spontaneously produce an unusual sense of anxiety. It is worth noting that if the listener's expectations can be affected in such a approach that they no longer expect to resolve the upper pitch by dropping to the 5th of the consonant, then the sixth fear-inducing effect will disappear. For example, this can be done by playing other harmony and using the sixth as part of the consonant major chord.

In the context of the expected 5th resolution, the fear-inducing effect of minor 6th can be found. Testing this phenomenon from the perspective ofTheory of Musical Equilibration means that it is believed that it is not a balance from the sixth to the 5th, but that it recognizes the sixth desire to resolve without 5th. This raises a different question: the 5th problem that it be thinked the chord's desire can't solve can be so frustrating – it makes us feel anxious?

If it is turned to the professional literature to describe the defining characteristics of 5th, it is can be quickly found an explanation. The Moser Music Dictionary cites an unexpected

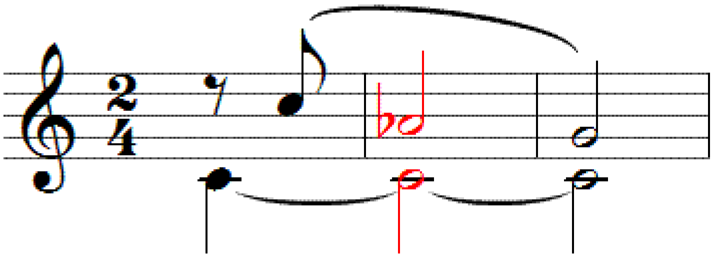


Figure 5.3: Minor sixth chord.

or dim example of a 5th sound. The Ullstein Music Dictionary calls the 5th Ghost effect.

If it is thinked of Ullstein's ghost concept and use Theory of Musical Equilibration to explain this interval, coming to the conclusion that after hearing the minor sixth degree, the listener thinks that the desire for music will not become weird. However, the process of identifying something that is not creepy is just another way of saying that it is a sense of fear and therefore the fear inducing effect of the minor sixth world.

5.2.3 Natural Minor Chord

A natural minor chord is a diatonic chord that is built by starting on the sixth degree of its relative major chord. Figure 5.4 shows the Natural minor chord in a music example segment.

In the minor dominant of minor tonic, the listener believes that the minor tonic of the dominant dominance remains unchanged. In other words, the audience recognizes the feeling of tolerating unwelcome things. Although this intention initially seems contradictory, if it is interpret the chord as an expression of courage, it will soon become reasonable. Brave people do what they would rather do: they overcome their emotions.

5.2.4 Neapolitan Sixth Chord

The Neapolitan sixth chord is generally labeled N6. As a basic material in the minor and vocabulary, it has been listed as an accent chord for almost all music textbooks for many years. To be more precise, it is considered to be a variant chord of the minor subordinate

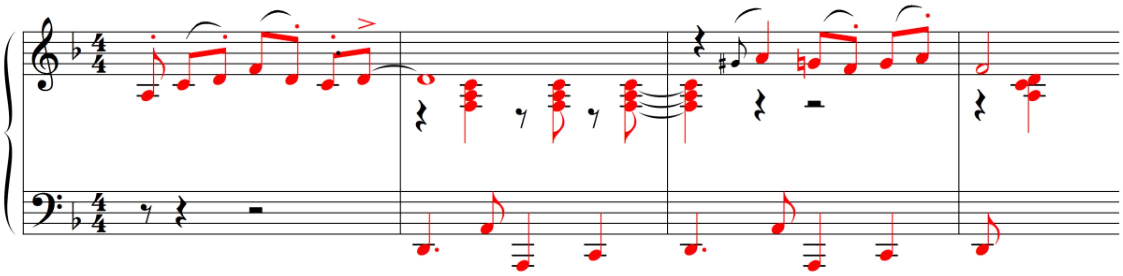


Figure 5.4: Natural minor chord.

II. This chord is generally thought to have appeared in the works of the Italian Napoli in the Baroque period. Farther away, it may be traced back to the Palestrina school in the late European Renaissance. Figure 5.5 shows the Neapolitan sixth chord in a music example segment.

The Neapolitan sixth chord has obvious painful effects. If Theory of Musical Equilibration is applied, considering the unique transformation that the chord undergoes while playing, it can only fully understand its particularly strong influence. The Neapolitan sixth chord had a different influence at the moment of its first vocalization, not after a brief moment. The reason is that it creates a sense of chaos in the audience.

This confusion can be explained as follows: When playing for the first time, The Neapolitan sixth chord sounds like a normal major chord, and the listener does not feel any unusual tone experience. If there is, they will feel Gustav Gldenstein's “a person standing upright in life”. The perception of the audience will then change, because they still feel the influence of the original key. Chords seem to be getting more and more discordant and full of increasingly strong impulses to solve. If it is insisted on the metaphor of “a person standing upright in life”, Theory of Musical Equilibration shows that a Neapolitan sixth chord symbolically turns this once-right person into a completely desperate person, and he has lost all support. sense.

Until Theory of Musical Equilibration was apply to this chord, it is can be feeled the contradiction between the sober satisfaction (equivalent to the major chord) and the clearly defined recognition and the desire to change. This inner conflict explains the remarkable effect of chords.

5.2.5 Subdominant Chord

The subdominant is the fourth pitch of the scale, which is called the subordinate because it is one level lower than the subto. Figure 5.6 shows the Subdominant chord in a music

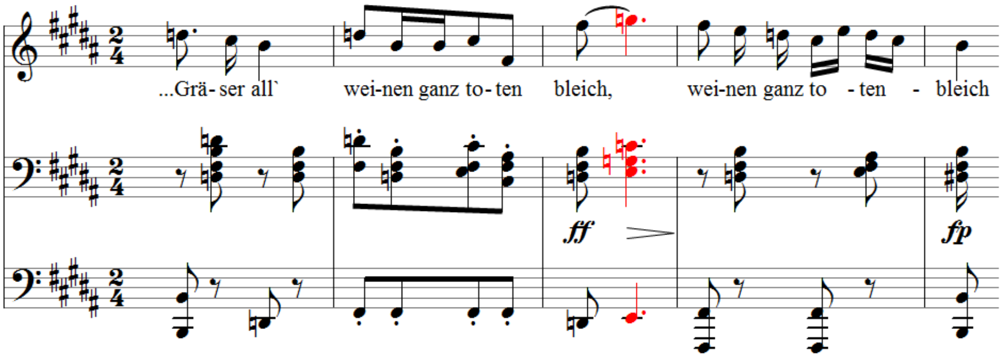


Figure 5.5: Neapolitan sixth chord.

example segment.

In classical and popular music, the main sub-dominance is used as a approach to convey relaxed and carefree emotions. In this study, passages with sub-optimal chords are described as the warmest and most friendly passages. Modulation to sub-dominant keys usually produces a sense of relaxation; instead of modulation into dominant (fifth note) scale, which adds tension.

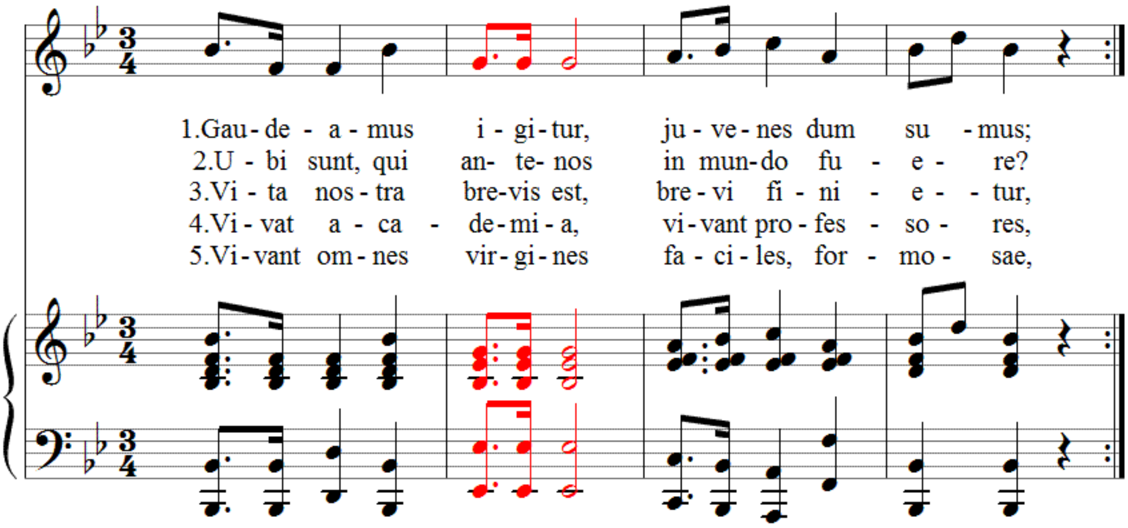

1.Gau- de - a - mus i - gi - tur, ju - ve - nes dum su - mus;
2.U - bi sunt, qui an- te- nos in mun-do fu - e - re?
3.Vi - ta nos - tra bre-vis est, bre - vi fi - ni - e - - tur,
4.Vi - vat a - ca - de - mi - a, vi - vant pro - fes - so - res,
5.Vi - vant om - nes vir- gi - nes fa - ci - les, for - mo - sae,

Figure 5.6: Subdominant chord.

When it is heard the subdominant chord, it is agreed with the desire for things that do not change, even though this feeling is very passive. From an emotional point of view, it corresponds to the satisfaction of the present and the present, which is very noteworthy because it does not seem to reveal. This mood coincides with a relaxed and happy moment, such as after an ecstasy or victory. This means that the secondary lead is also very suitable for singing songs on a cheerful occasion.

5.2.6 Augmented Chord

Augmented chord is a chord that is superimposed by three tones + major third degree intervals. The chord name is represented by the root name as aug. It can be also understanded as the third chord as a fifth degree raises the semitone in Major chord. Figure 5.7 shows the Augmented chord in a music example segment.

A typical feature of augmented chord is the ambiguity of the perceived effect of musical balance. Augmented chord has conflicting balances, which means that the emotional characteristics it causes can vary according to the theory of musical balance. In other words, when augmented chord is heard, it is cannot be clearly understanded that things haven't changed with the will: the audience is still questioning. This chord is used to convey surprise, surprise or surprise. In film music, it is a good fit to illustrate some extraordinary things happening in the plot. A series of augmented chords were played in the scene where the twins first met, until that moment they were completely unaware of each other's existence.

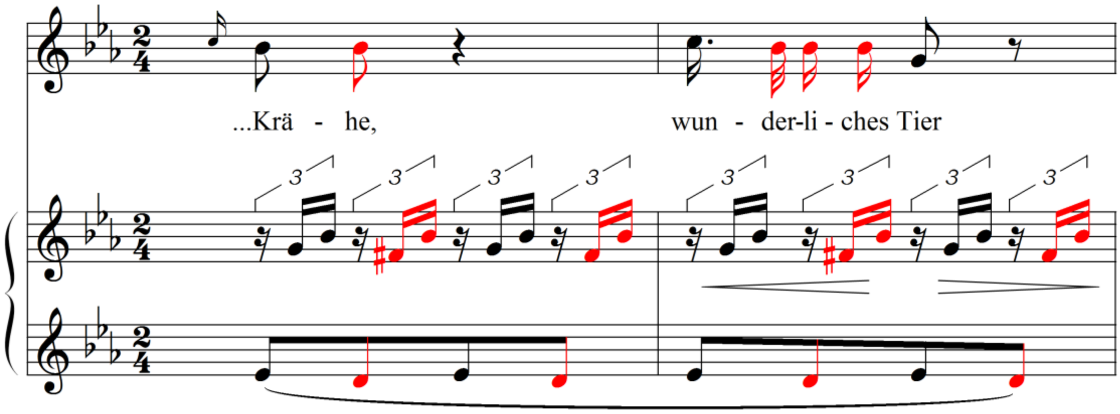


Figure 5.7: Augmented chord.

Table 5.1: The emotions in the chords.

Chords	Emotions
Minor sixth chord	Fear
C Minor chord	Sad
Natural minor chord	Exciting
Major subdominant chord	Relax
Neapolitan sixth chord	Despair
Augmented chord	Amazement

In summary, these six kinds of chords are classed in six kinds of different emotions, which can be seen in Table 5.1 and the six true label chords were selected from the piano teaching tutorial music.

5.3 Chord Identification

In this system, based on “Theory of Musical Equilibration” in the book “Music and Emotions” which designed and conducted by Daniela and Bernd Willimek [77], the different kinds of chords is the core of this system. According to the book, the data for test is extracted from the piano music teaching audio by using the Adobe Audititon.

The music segments have extracted by the chords styles in 6 types (Augmented, Minor tonic, Minor sixth, Natural minor, Neapolitan sixth and Major Subdominant chord). Each of the chord corresponds one emotion. So the whole music signal is split to single chord segments.

Then the chord segments will be sent for identification though selected recognition Method, which is correlation and Euclidean distance, after FFT and STAT feature extraction. The chord identification system overview can be found in Figure 5.8.

5.3.1 Data Preparation

Based on the study of the book “Music and Emotions”, the mapping relationship between chords and emotions is understood as the criterion for emotional classification. Therefore, how to accurately detect accurate chords has become the purpose of research. For this purpose, it is necessary to build a database with enough samples, accurate chord classification, and enough chord type to prepare for the next experiment. The details of the database shown in Table 5.2 and Figure 5.10.The flow figure of the database building is illustrate in Figure 5.9.

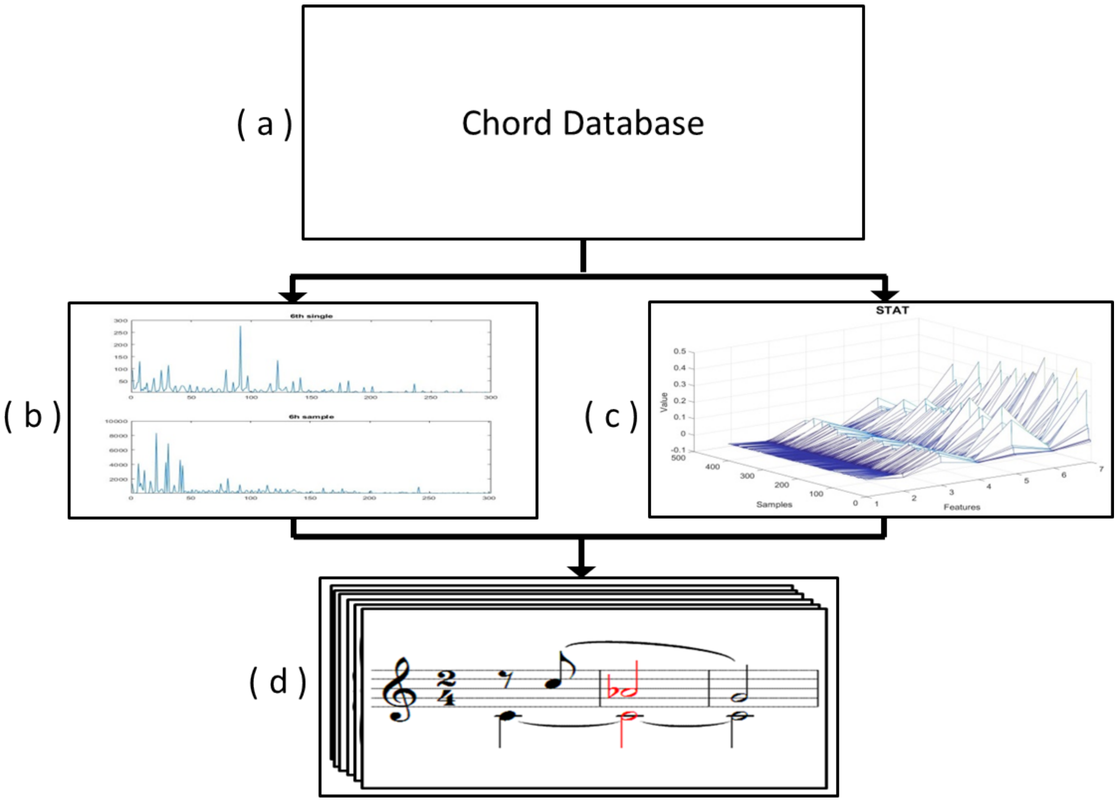


Figure 5.8: Chord identification system (a.The chord database; b.The FFT features; c.The STAT features; d.The identified chords).

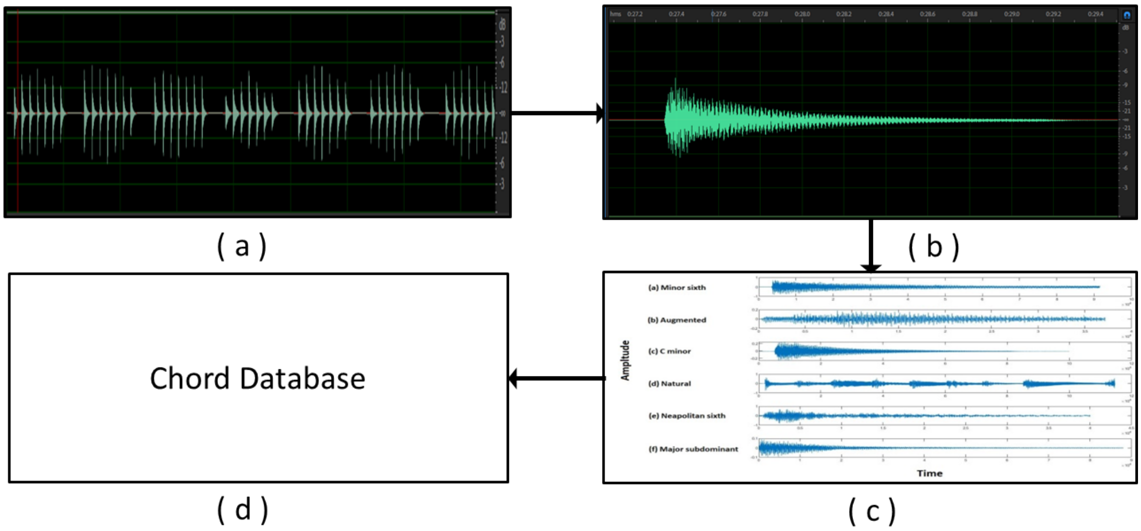
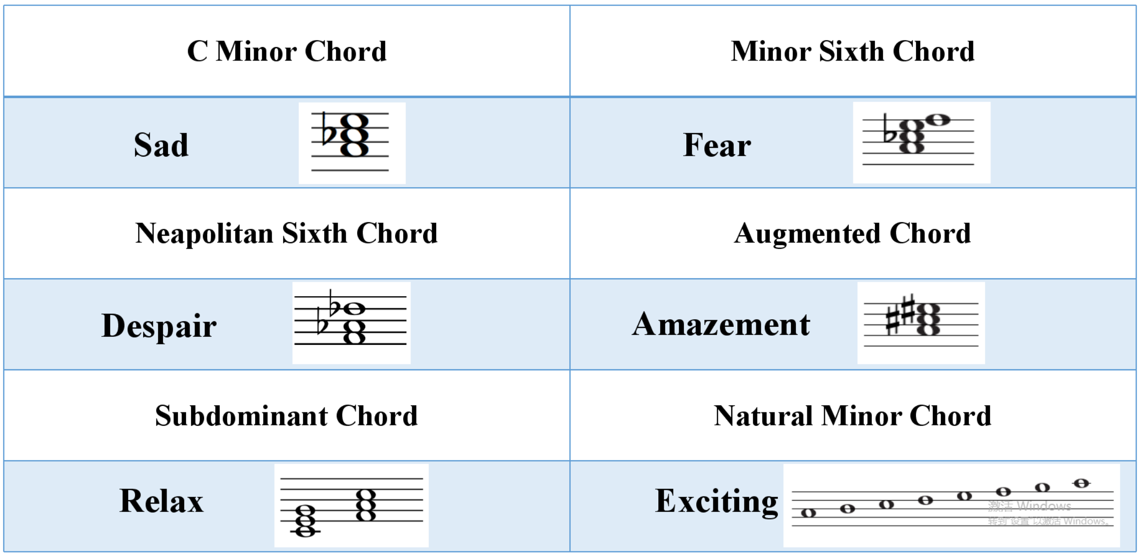


Figure 5.9: Chord database build by piano music signals(a. Original music signals; b. Chord split; c. The selected chord class).

In the first step, since chords can be expressed through various instruments, it is necessary to select a uniform instrument standard for database preparation. So the most universal piano is the best choice because it can play all the chords. Through the study of some

Table 5.2: Chords database detail.

Number of chords	6 (Augmented chord, Minor tonic chord, Minor sixth chord, Natural minor chord, Neapolitan sixth chord and Major Subdominant chord)
Number of segments per chord	10
Average sampling points per segment	78357
Types of emotion	6 (Fear, Sad, Exciting, Relax, Despair and Amazement)



C Minor Chord	Minor Sixth Chord
Sad	Fear
Neapolitan Sixth Chord	**Augmented Chord**
Despair	Amazement
Subdominant Chord	**Natural Minor Chord**
Relax	Exciting

Figure 5.10: Six kinds of chords and their corresponding emotions.

piano course teaching videos, it is found that because of the teacher's demonstration process in the teaching video, there is a clear indication of the type of chords and an accurate interpretation of the chords. A large number of piano teaching videos containing target chords were sought to prepare for the completion of the database.

In the second step, the audio portion of the found teaching screen is segmented using the software Adobe Audition shown as the work space in the software in Figure 5.11. Because the chord presentation part of piano music teaching is scattered throughout the audio file portion of the entire screen. Professional audio segmentation software is required for extraction processing Adobe Audition is a professional open audio processing software, and its functionality and extraction accuracy can be recognized. Then, through the teacher's explanation and analysis in the audio, the part of the target chord is accurately extracted and divided and saved, such as Figure 5.12.

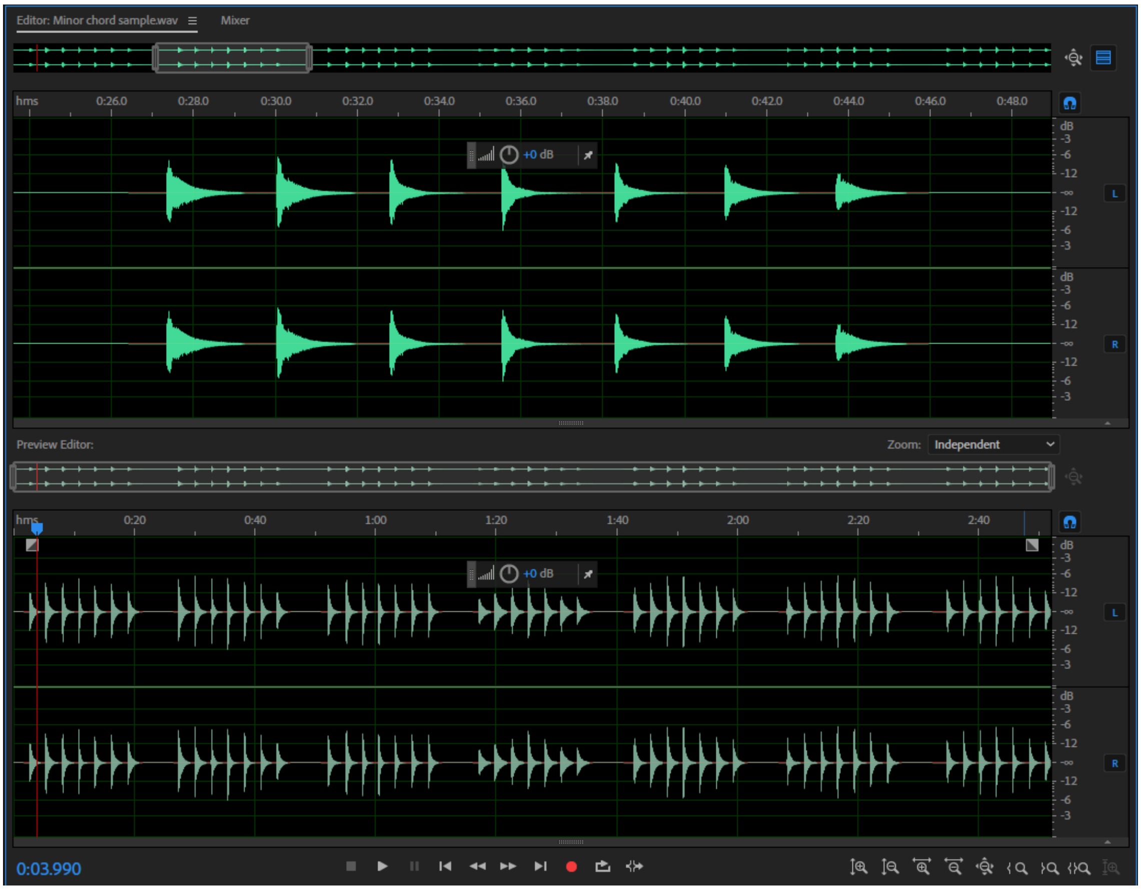

Zoom: Independent

Figure 5.11: Adobe Adudition work space.

In the third step, the extracted chords are classified and labeled. Based on the theory in the book "Music and Emotions", Augmented chord, Minor tonic chord, Minor sixth chord, Natural minor chord, Neapolitan sixth chord and Major Subdominant chord are chosen

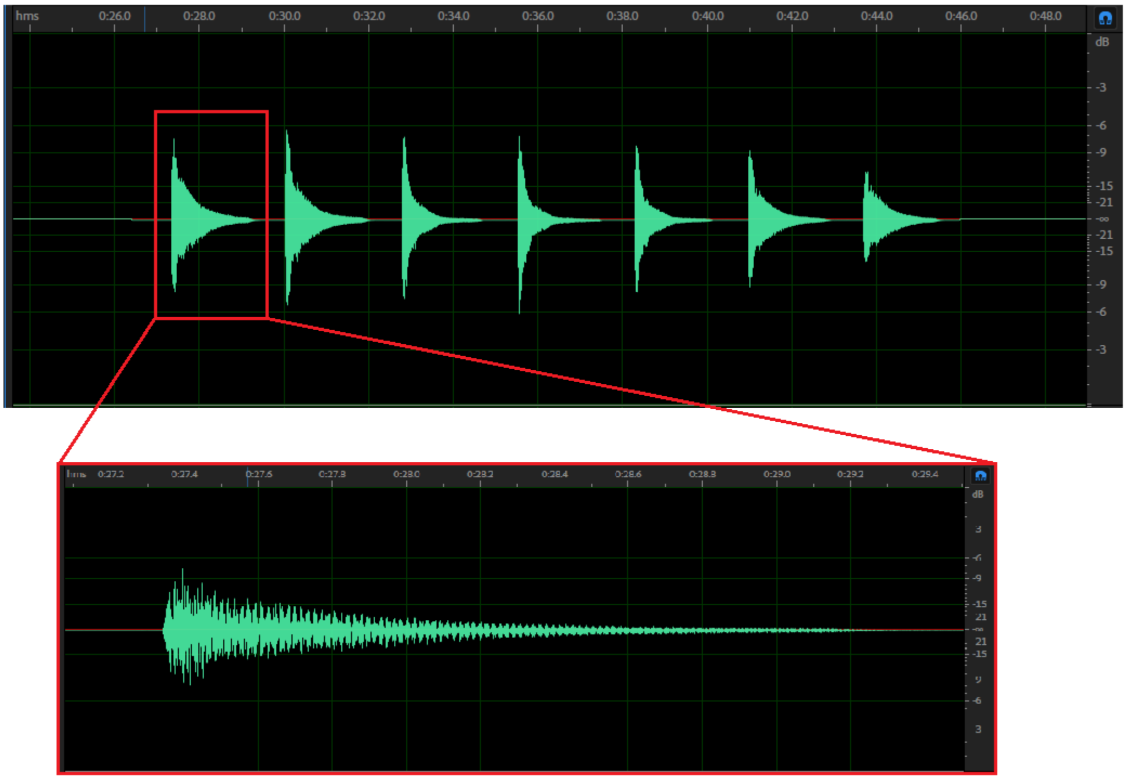


Figure 5.12: Chord extraction.

as the type of target chord. These six chords correspond to six emotions of amazement, sadness, fear, excitement, despair and relaxation. Based on these six chords, the data is divided into 10 different chord segments for each emotion shown in Figure 5.13. A sample database of 60 different chords. After completing the classification and labeling of the six chords, select the most accurate set of six different kinds of chords as the real label. At this point, the experimental data is ready to end.

5.3.2 Chord Recognition Method

When the data has sample data and real tags, the goal of the study shifts to how the actual tag data has been tested to match whether the sample data in the database matches. Because chords are a set of one-dimensional audio data, the problem to be solved is to choose what method to find a similar one to the real tag in a bunch of audio data. There are many approaches to compare one-dimensional data, the most common of which is the Euclidean distance detection method. In mathematics, the Euclidean distance is the distance between two points in Euclidean space, which can be found in Formula 5.1. Using this distance, the Euclidean space becomes the metric space. In this method, the Euclidean distance calculation is performed directly on the true label vector and all sample vectors.

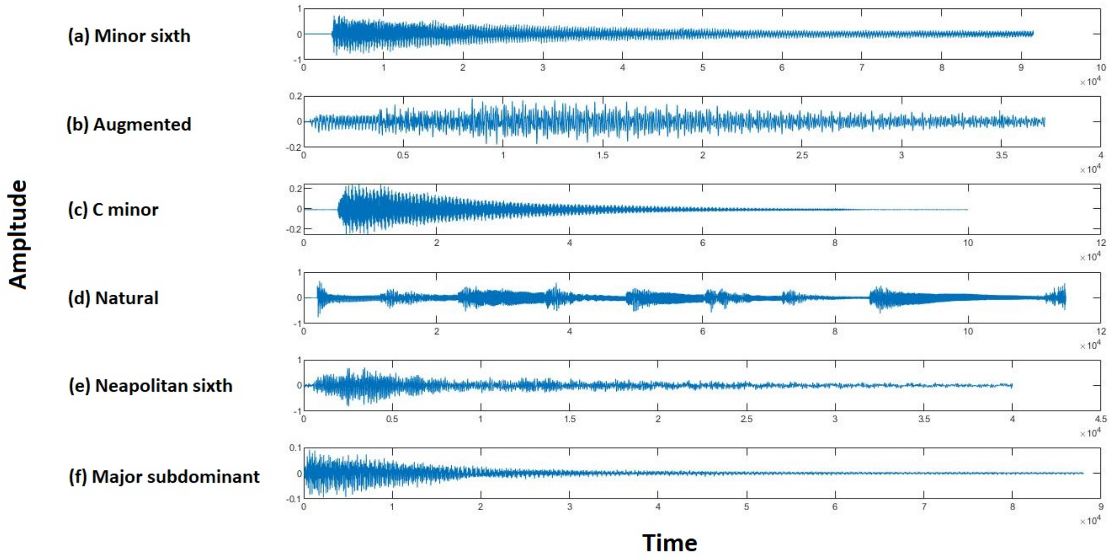


Figure 5.13: Six kinds of selected chords.

Then, compare the distance between each sample vector and the real label vector, and observe which vector has the shortest distance, and consider the classification of which real label the sample belongs to. This type of detection is very effective for comparing two vectors with the same or similar orientation.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{5.1}$$

But not every vector has a similar spatial position, so another set of detection methods is needed to match it, thus reducing the error of detection. This method is correlation detection. The correlation coefficient is the first statistical indicator designed by the statistician Carl Pearson and is the amount of linear correlation between the variables studied, usually expressed by the letter r. Due to the different research objects, there are many approaches to define the correlation coefficient. The Pearson correlation coefficient is more commonly used, can be found in Formula 5.2. The correlation coefficient is a statistical indicator used to reflect the closeness of the correlation between variables. The correlation coefficient is calculated by the difference method. It is also based on the dispersion of the two variables and their respective averages. The two differences are multiplied to reflect the degree of correlation between the two variables. The linear single correlation coefficient is studied in this study. The Pierce correlation coefficient was chosen as the criterion for detected correlations.

$$r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}} \tag{5.2}$$

5.4 Experiments

At the beginning of the experiment, for analyze the shape, distribution and features of the six different chords, six benchmark labels are entered into Matlab for observation. At the beginning, spectrograms were considered a good approach to observe and analyze audio data.The segment of music is used with chord to do the spectrograms for find the features. Then the correct spectrum and spectrogram are illustrated for the music chord. It is aim to find the features of each different chord. The CNN features extraction methods and classification was considered to be used on the spectrum as well in the beginning of the experiment. However, because the recognition method based on the Theory of Musical Equilibration is first time to be used. Two more basic features and methods were decided to be applied. The spectrums for the chords which based By Discrete Fourier transform (DFT) have been got. It can be easily found from the spectrums, the specific chord have the specific frequency feature. By observing the spectrogram, it is found that the frequency characteristics of the audio data can better show the difference between the different and the county. So the FFT of the target audio signal becomes the next step.

FFT have been done before interpolated 200 points in chord feature. Then 1000 points features have been extracted from the both chord and sample FFT. After the distance calculation between the chord and sample music, the results show the strange relation between all the features. Might be that have not done the normalization after extract 1000 FFT points.

But when the work will move to the database testing, it was found that the AMP database and all data are not suitable for the methods testing. Because the piano or single instrument is playing the chord, the data got were that the music is extracted from the piano tutorial video as the basic database which is mentioned previous.

Following, there is a problem: How can it be made sure that the different kinds of samples are actually different? Six comparisons of different kinds of chord have been taken. The graphs intuitive display the similar between the same kind of chord and the difference between the other chords. Figure 5.14, 5.15, 5.16, 5.17, 5.18 and 5.19 illustrate that the 6 kinds of chords have different frequency distributed and different frequency strength after 1000 sampling point FFT, Which means that the unique frequency combination in chords can distinguish the different kind of chords. So the selected FFT has been chosen as the one of the significant features in chord identification.

After completing the production and preparation of the database, the experiment went to the key part.In the next experiment, the audio data FFT of each true label is calculated

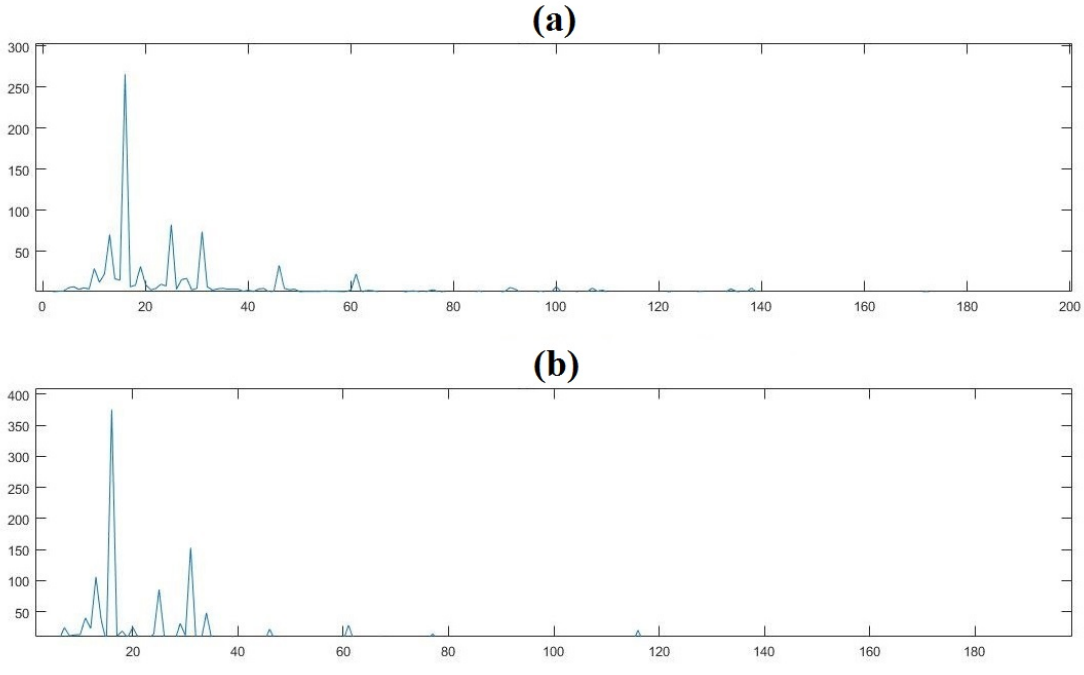


Figure 5.14: The comparison figure between True FFT and Sample FFT of Augmented Chord.(a. benchmark Augmented chord; b. sample Augmented chord)

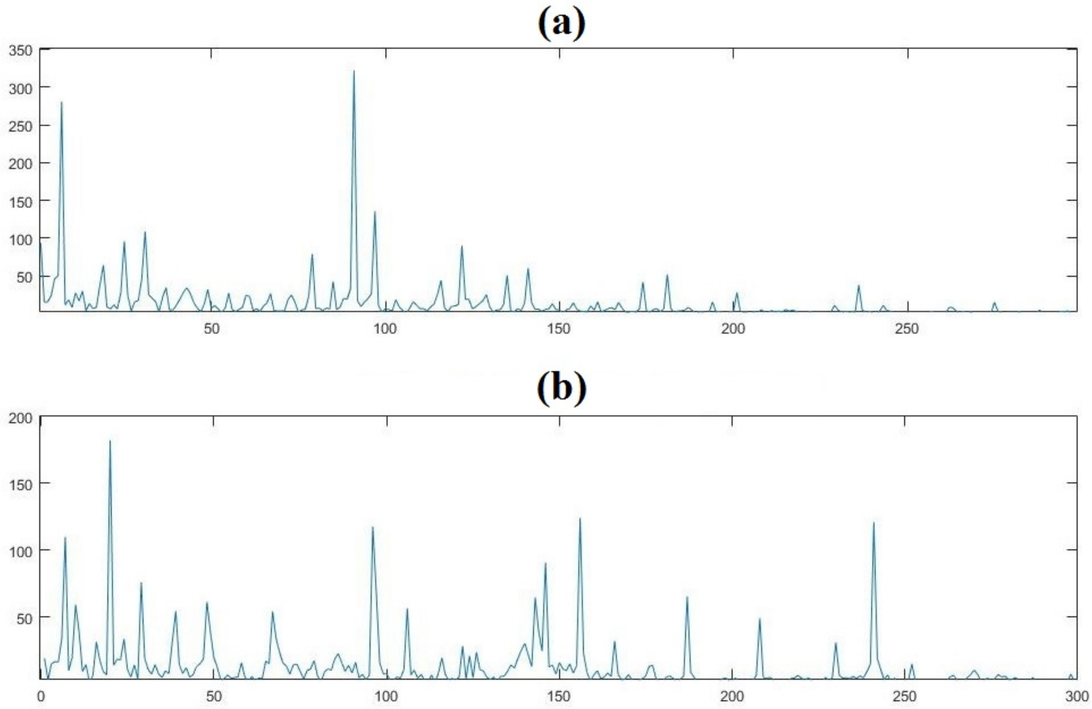


Figure 5.15: The comparison figure between True FFT and Sample FFT of Minor sixth Chord.(a. benchmark Minor sixth chord; b. sample Minor sixth chord)

with the EMF of each sample data and the result size is compared.It can be seen from Table 5.3 that the matching results of the three chords of Augmented, C Minor and Minor 6th are much better than the other three.Taking into account the limitations of FFT and

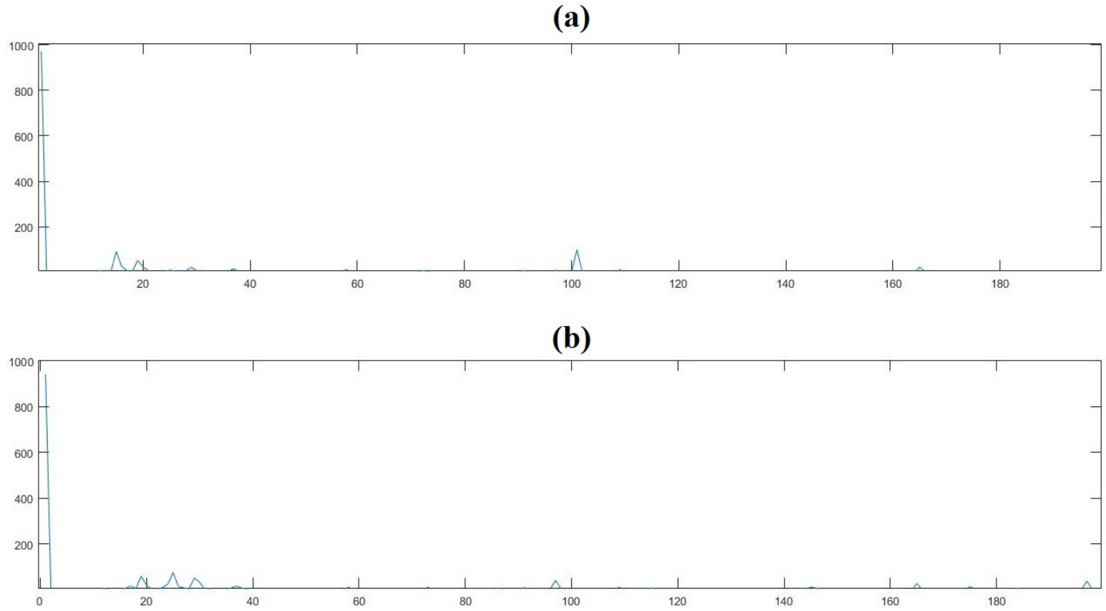


Figure 5.16: The comparison figure between True FFT and Sample FFT of C Minor Chord.(a. benchmark C Minor chord; b. sample C Minor chord)

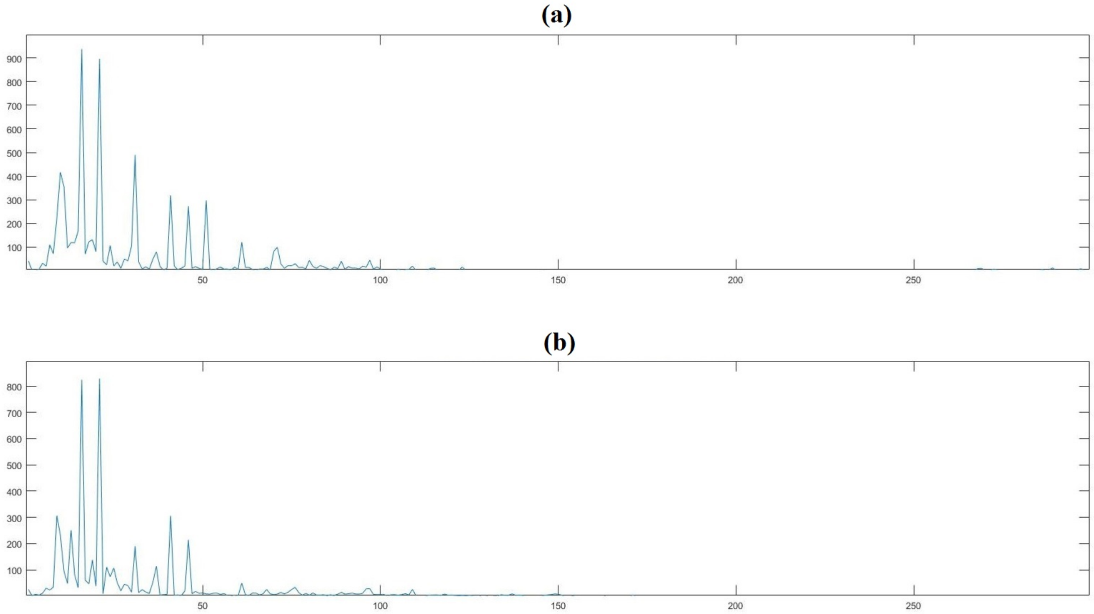


Figure 5.17: The comparison figure between True FFT and Sample FFT of Natural Minor Chord.(a. benchmark Natural Minor chord; b. sample Natural Minor chord)

Euclidean distance calculations, the STAT feature is once again selected as a new feature extraction to cover more efficient information selection.

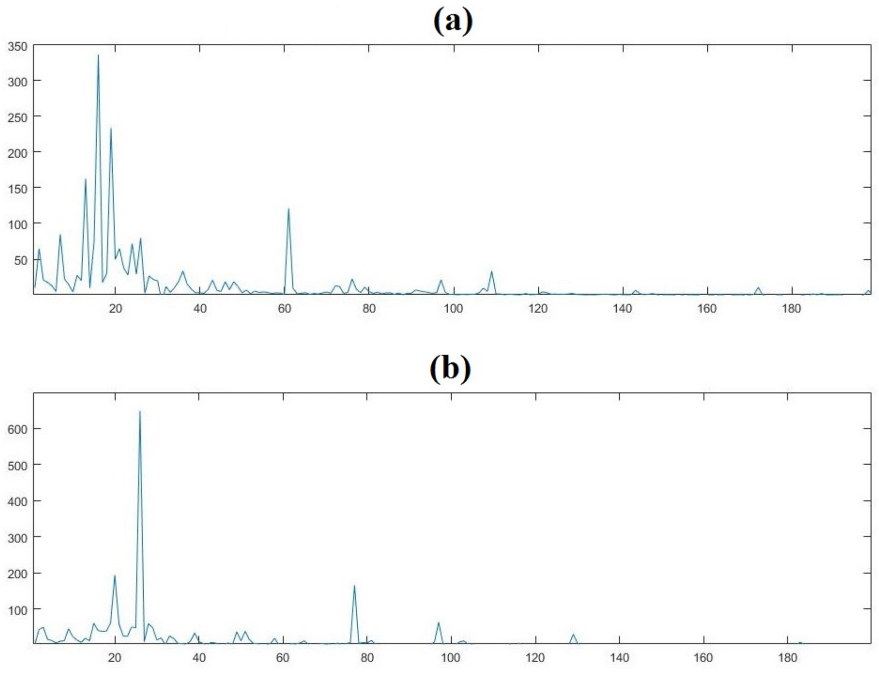


Figure 5.18: The comparison figure between True FFT and Sample FFT of Neapolitan Chord.(a. benchmark Neapolitan chord; b. sample Neapolitan chord)

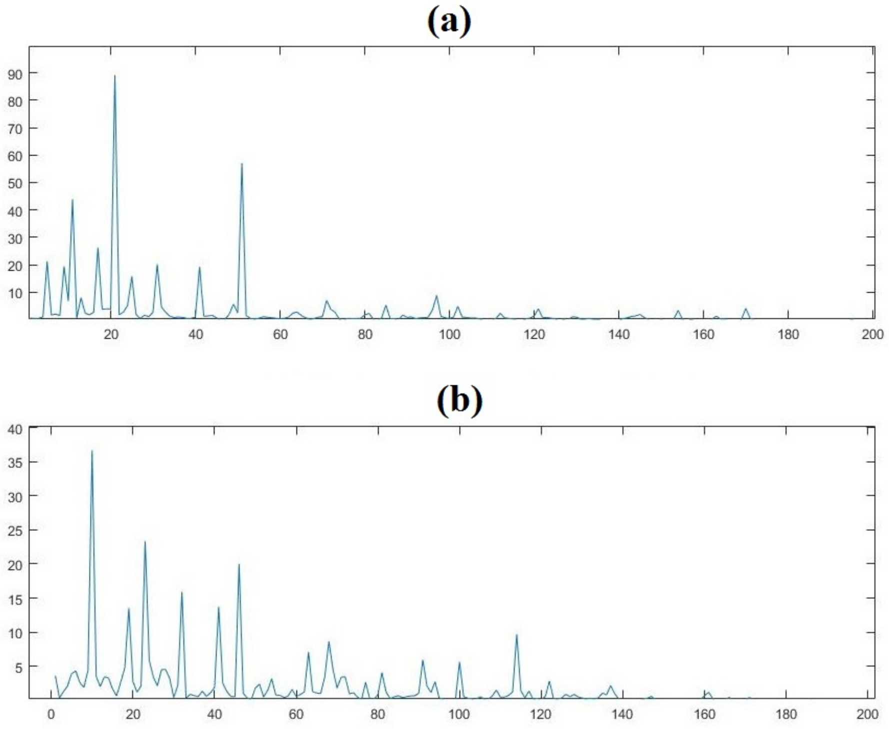


Figure 5.19: The comparison figure between True FFT and Sample FFT of Subdominant Chord.(a. benchmark Subdominant chord; b. sample Subdominant chord)

Table 5.3: The Euclidean distance comparison test with FFT features in Chord database (The meaning of abbreviation : Aug - Augmented chord, m - Minor chords, m6 - Minor 6th chord, Sub-d - Major subdominant chord).

Sample\True	Aug	m	m6	Natural	N6	Sub-d
Aug	**0.870**	1.547	1.761	1.603	1.269	1.861

Table 5.3 continued from previous page

Aug	**0.826**	1.519	1.519	1.544	1.243	1.856
Aug	**0.864**	1.640	1.846	1.395	1.331	1.839
Aug	**0.752**	1.560	1.823	1.564	1.288	1.866
Aug	1.681	1.858	2.019	1.908	1.552	**1.379**
Aug	1.572	1.762	1.960	1.811	1.441	**1.272**
Aug	**0.836**	1.550	1.811	1.558	1.196	1.744
Aug	**0.791**	1.538	1.806	1.603	1.266	1.825
Aug	**1.181**	1.776	1.877	1.728	1.506	2.017
Aug	**0.825**	1.531	1.818	1.582	1.283	1.854
m	1.726	**0.115**	1.757	1.746	1.542	1.910
m	1.724	**0.170**	1.748	1.728	1.539	1.897
m	1.728	**0.122**	1.744	1.752	1.542	1.900
m	1.746	**0.130**	1.755	1.755	1.550	1.902
m	1.728	**0.108**	1.745	1.756	1.544	1.899
m	1.724	**0.149**	1.748	1.731	1.537	1.895
m	1.733	**0.145**	1.752	1.736	1.540	1.904
m	1.851	**0.714**	1.871	1.837	1.241	1.657
m	1.873	**0.761**	1.885	1.852	1.247	1.662
m	1.833	**0.638**	1.849	1.831	1.266	1.704
m6	2.000	1.818	**1.109**	1.988	1.967	2.155
m6	2.171	2.423	**1.960**	2.267	2.352	2.540
m6	2.054	1.906	**1.101**	2.051	2.051	2.218
m6	2.024	2.214	**1.567**	2.183	2.195	2.414
m6	1.830	1.857	**1.437**	1.921	1.887	2.146
m6	2.009	1.826	**1.094**	1.982	1.984	2.184
m6	1.879	1.927	**0.971**	1.978	1.969	2.158
m6	2.054	1.925	2.068	**1.483**	1.900	1.979
m6	1.814	1.570	1.883	1.650	**1.518**	1.956
m6	2.605	**2.431**	2.556	2.666	2.615	2.860
Natural	1.450	2.004	2.116	1.643	**1.103**	1.438
Natural	1.398	1.899	2.135	1.628	**0.853**	1.606
Natural	1.301	1.910	2.067	1.432	**1.124**	1.747
Natural	1.868	1.845	1.985	**0.438**	1.738	1.937
Natural	1.429	1.932	2.133	1.638	**0.981**	1.372
Natural	1.455	1.945	2.153	1.585	**1.157**	1.659
Natural	1.436	1.985	2.112	**1.125**	1.525	1.685

Table 5.3 continued from previous page

Natural	1.315	1.874	2.125	1.666	**0.988**	1.849
Natural	1.515	1.886	2.149	1.542	**0.895**	1.916
Natural	1.485	1.920	2.096	**0.842**	1.148	1.585
N6	**1.444**	2.008	2.044	1.867	1.500	2.154
N6	1.571	1.490	1.834	1.655	**1.298**	1.839
N6	**0.906**	1.562	1.872	1.522	1.146	1.911
N6	1.999	2.007	2.089	**1.448**	1.672	2.083
N6	**1.515**	2.081	2.196	1.807	1.581	2.267
N6	2.009	1.826	1.994	1.585	**1.157**	2.184
N6	1.879	1.927	1.545	**1.125**	1.525	2.158
N6	2.054	1.925	2.068	1.666	**0.988**	1.979
N6	1.814	1.570	1.883	1.650	**1.518**	1.956
N6	2.605	**2.431**	2.556	2.666	2.615	2.860
Sub-d	1.865	1.732	1.980	1.742	1.123	**0.521**
Sub-d	1.985	1.866	1.934	**1.286**	1.843	2.089
Sub-d	**1.933**	2.517	2.134	1.942	2.283	2.641
Sub-d	1.833	**0.638**	1.849	1.831	1.266	1.704
Sub-d	1.455	1.945	2.153	1.585	**1.157**	1.659
Sub-d	1.301	1.910	2.067	1.432	**1.124**	1.747
Sub-d	1.572	1.762	1.960	1.811	1.441	**1.272**
Sub-d	1.851	**0.714**	1.871	1.837	1.241	1.657
Sub-d	1.873	**0.761**	1.885	1.852	1.247	1.662
Sub-d	1.429	1.932	2.133	1.638	**0.981**	1.372

The confusion matrix in Table 5.4 gives more clear illustration which is that the Augmented chord, C Minor chord and the Minor 6th chord have better identification accuracy than other 3 members.

Move the eyes to the STAT features.It can be seen from the Table 5.5 that in the case

Table 5.4: FFT Euclidean distance confusion matrix.

True \ Sample	Aug	m	m6	Natural	N6	Sub-d
Aug	8	0	0	0	2	1
m	0	10	1	0	1	3
m6	0	0	7	0	0	0
Natural	0	0	1	3	2	1
N6	0	0	1	7	5	3
Sub-d	2	0	0	0	0	2

of using the STAT feature selection, the matching accuracy of the latter three chords is greatly improved, but the matching accuracy of the Augmented chord is completely destroyed.This may be due to the principle of extracting STAT features, resulting in the loss of information in many Augmented chords.

Table 5.5: The Euclidean distance comparison test with STAT features in Chord database (The meaning of abbreviation : Aug - Augmented chord, m - Minor chords, m6 - Minor 6th chord, Sub-d - Major subdominant chord).

Sample\True	Aug	m	m6	Natural	N6	Sub-d
Aug	**0.244**	0.254	0.201	0.793	0.781	0.491
Aug	0.265	0.242	**0.153**	0.737	0.736	0.536
Aug	0.509	0.272	**0.244**	0.570	0.538	0.761
Aug	0.376	0.206	**0.126**	0.642	0.646	0.642
Aug	**0.176**	0.297	0.233	0.837	0.827	0.439
Aug	**0.140**	0.317	0.252	0.862	0.859	0.403
Aug	0.346	0.244	**0.194**	0.709	0.683	0.596
Aug	0.373	0.207	**0.124**	0.642	0.647	0.639
Aug	0.325	0.224	**0.166**	0.718	0.708	0.581
Aug	0.271	0.225	**0.151**	0.744	0.743	0.532
m	0.504	**0.065**	0.166	0.629	0.675	0.724
m	0.528	**0.097**	0.172	0.587	0.638	0.756
m	0.408	**0.037**	0.121	0.699	0.733	0.625
m	0.309	**0.157**	0.175	0.803	0.827	0.508
m	0.369	**0.089**	0.141	0.746	0.776	0.577
m	0.478	**0.041**	0.140	0.634	0.678	0.701
m	0.465	**0.028**	0.132	0.643	0.685	0.688
m	0.515	**0.087**	0.161	0.593	0.641	0.744
m	0.482	**0.053**	0.135	0.620	0.664	0.709
m	0.317	**0.142**	0.160	0.788	0.812	0.521
m6	0.431	0.114	**0.094**	0.623	0.639	0.669
m6	0.358	0.289	**0.189**	0.686	0.717	0.621
m6	0.361	0.108	**0.085**	0.704	0.724	0.590
m6	0.296	0.257	**0.164**	0.737	0.761	0.555
m6	0.293	0.231	**0.138**	0.734	0.759	0.550
m6	0.421	0.122	**0.102**	0.635	0.644	0.657
m6	0.349	0.136	**0.055**	0.681	0.707	0.595

Table 5.5 continued from previous page

m6	0.815	0.440	0.453	**0.290**	0.378	1.061
m6	0.603	0.271	**0.269**	0.457	0.495	0.832
m6	0.400	0.211	**0.183**	0.670	0.645	0.625
Natural	0.901	0.525	0.543	**0.312**	0.409	1.152
Natural	0.983	0.662	0.645	**0.025**	0.269	1.242
Natural	0.982	0.683	0.658	**0.090**	0.263	1.239
Natural	0.990	0.674	0.656	**0.033**	0.253	1.248
Natural	1.000	0.673	0.662	**0.145**	0.214	1.265
Natural	0.986	0.625	0.563	**0.041**	0.162	1.357
Natural	0.913	0.684	0.624	**0.154**	0.266	1.117
Natural	0.977	0.613	0.659	0.421	**0.365**	1.266
Natural	0.984	0.513	0.615	**0.095**	0.266	1.215
Natural	0.922	0.698	0.684	**0.149**	0.217	1.369
N6	1.143	0.923	0.886	0.443	**0.309**	1.392
N6	1.023	0.679	0.676	**0.101**	0.254	1.277
N6	0.834	0.525	0.525	0.361	**0.272**	1.080
N6	0.943	0.616	0.613	0.246	**0.217**	1.198
N6	0.881	0.564	0.556	0.311	**0.306**	1.145
N6	0.921	0.541	0.812	0.216	**0.206**	1.148
N6	0.912	0.617	0.691	**0.318**	0.336	1.266
N6	1.117	0.624	0.600	0.358	**0.265**	1.369
N6	1.089	0.511	0.628	0.485	**0.267**	1.132
N6	0.965	0.933	0.599	**0.266**	0.302	1.356
Sub-d	0.286	0.660	0.631	1.247	1.233	**0.002**
Sub-d	0.268	0.664	0.626	1.241	1.231	**0.079**
Sub-d	0.263	0.657	0.620	1.237	1.226	**0.069**
Sub-d	0.265	0.667	0.622	1.249	1.327	**0.065**
Sub-d	0.280	0.665	0.637	1.269	1.258	**0.052**
Sub-d	0.265	0.672	0.613	1.299	1.270	**0.066**
Sub-d	0.271	0.641	0.700	1.278	1.361	**0.052**
Sub-d	0.281	0.637	0.662	1.296	1.266	**0.037**
Sub-d	0.265	0.665	0.679	1.315	1.215	**0.070**
Sub-d	0.285	0.679	0.622	1.326	1.255	**0.074**

The confusion matrix in Table 5.6 illustrate more details about this results.All the kinds of chord shown the nice identification accuracy except the Augmented one.Especially the C Minor chord and the Subdominant chord shown the perfect results.

Table 5.6: STAT Euclidean distance confusion matrix.

True \ Sample	Aug	m	m6	Natural	N6	Sub-d
Aug	2	0	0	0	0	0
m	0	10	0	0	0	0
m6	8	0	9	0	0	0
Natural	0	0	1	8	3	0
N6	0	0	0	2	7	0
Sub-d	0	0	0	0	0	10

At the same time, the correlation coefficient detection method is proposed to be more effective than the Euclidean distance method.Therefore, the results in Table 5.7 use the Pierce correlation coefficient to calculate the correlation of STAT.From the results, it can be seen that the results of the correlation method are basically similar to the results of the Euclidean distance.It may be better than the Euclidean distance method in the matching classification of some chords, but it is basically approximate accuracy.

Table 5.7: The Correlation by Pearson correlation coefficient comparison test with STAT features in Chord database (The meaning of abbreviation : Aug - Augmented chord, m - Minor chords, m6 - Minor 6th chord, Sub-d - Major subdominant chord).

Sample\True	Aug	m	m6	Natural	N6	Sub-d
Aug	0.988	**0.993**	0.992	0.839	0.826	0.944
Aug	0.983	0.995	**0.997**	0.862	0.843	0.929
Aug	0.941	0.982	**0.984**	0.915	0.917	0.865
Aug	0.968	0.994	**0.997**	0.894	0.879	0.902
Aug	**0.993**	0.991	0.991	0.823	0.806	0.953
Aug	**0.996**	0.990	0.990	0.813	0.790	0.960
Aug	0.973	0.991	**0.992**	0.871	0.867	0.915
Aug	0.969	0.994	**0.997**	0.894	0.878	0.902
Aug	0.979	0.994	**0.995**	0.867	0.857	0.922
Aug	0.984	0.995	**0.997**	0.858	0.841	0.933
m	0.967	**0.999**	0.997	0.887	0.865	0.909
m	0.960	**0.998**	0.997	0.902	0.881	0.896
m	0.979	**1.000**	0.997	0.860	0.836	0.932
m	0.991	**0.994**	0.990	0.817	0.791	0.958
m	0.985	**0.998**	0.994	0.841	0.815	0.945

Table 5.7 continued from previous page

m	0.969	**1.000**	0.998	0.885	0.863	0.912
m	0.971	**1.000**	0.998	0.882	0.859	0.915
m	0.961	**0.998**	0.997	0.900	0.879	0.899
m	0.967	**0.999**	0.998	0.891	0.869	0.908
m	0.989	**0.996**	0.992	0.824	0.798	0.955
m6	0.966	0.997	**0.997**	0.892	0.876	0.906
m6	0.966	0.987	**0.994**	0.880	0.843	0.900
m6	0.980	0.998	**0.998**	0.861	0.840	0.931
m6	0.980	0.992	**0.996**	0.857	0.822	0.925
m6	0.982	0.994	**0.997**	0.856	0.823	0.929
m6	0.967	0.997	**0.997**	0.888	0.873	0.908
m6	0.978	0.998	**0.999**	0.874	0.850	0.923
m6	0.867	0.950	**0.955**	0.978	0.968	0.764
m6	0.913	0.977	**0.977**	0.944	0.926	0.836
m6	0.962	**0.992**	0.990	0.879	0.872	0.908
Natural	0.858	0.943	0.949	**0.982**	0.975	0.749
Natural	0.761	0.875	0.885	**1.000**	0.983	0.629
Natural	0.739	0.859	0.870	**0.999**	0.980	0.604
Natural	0.749	0.867	0.877	**1.000**	0.984	0.616
Natural	0.770	0.878	0.889	**0.995**	0.995	0.636
Natural	0.785	0.965	**0.995**	0.993	0.984	0.717
Natural	0.750	0.852	0.842	**0.993**	0.971	0.767
Natural	0.420	0.898	0.812	**0.998**	0.965	0.665
Natural	0.720	0.893	0.866	**0.998**	0.988	0.690
Natural	0.715	0.877	0.878	**0.999**	0.980	0.615
N6	0.578	0.726	0.738	0.959	**0.976**	0.417
N6	0.760	0.876	0.884	**0.998**	0.991	0.631
N6	0.836	0.922	0.925	0.965	**0.981**	0.728
N6	0.801	0.901	0.907	0.985	**0.994**	0.679
N6	0.839	0.923	0.931	0.977	**0.984**	0.720
N6	0.784	0.916	0.992	0.979	**0.999**	0.785
N6	0.842	0.931	0.848	**0.967**	0.919	0.660
N6	0.698	0.717	0.516	0.992	**0.990**	0.790
N6	0.655	0.699	0.698	0.942	**0.980**	0.800
N6	0.795	0.920	0.799	0.926	**0.997**	0.699
Sub-d	0.979	0.923	0.914	0.626	0.598	**1.000**

Table 5.8: STAT correlation confusion matrix.

True \ Sample	Aug	m	m6	Natural	N6	Sub-d
Aug	2	0	0	0	0	0
m	2	10	3	0	0	0
m6	6	0	7	1	0	0
Natural	0	0	0	9	4	0
N6	0	0	0	0	6	0
Sub-d	0	0	0	0	0	10

Table 5.7 continued from previous page

Sub-d	0.982	0.923	0.917	0.628	0.596	**0.998**
Sub-d	0.982	0.925	0.918	0.631	0.599	**0.999**
Sub-d	0.982	0.928	0.915	0.636	0.599	**0.995**
Sub-d	0.980	0.915	0.930	0.632	0.579	**0.992**
Sub-d	0.989	0.925	0.920	0.630	0.600	**0.992**
Sub-d	0.992	0.927	0.967	0.619	0.600	**0.999**
Sub-d	0.975	0.925	0.925	0.629	0.575	**0.998**
Sub-d	0.982	0.928	0.915	0.700	0.600	**0.992**
Sub-d	0.982	0.915	0.928	0.655	0.580	**0.992**

It also can been found in Table 5.8, the confusion matrix of Table 5.7.The result of identification accuracy is similar as the Euclidean distance by STAT features.The Augmented Chord identification is terrible but others are effective.

At last, the chords which extracted from the same music signal will be classification for the main chord class in that music based on the majority voting results. Then according to the Theory of Musical Equilibration, the emotion in this music should be recognized. The flow figure is Shown in Figure 5.20. The results of the identification accuracy is illustrated in Table 5.9.Can be found that the accuracy of STAT feature reaches 75%. However the FFT feature only achieves 58.3%.

5.5 Summary

Table 5.9: The identification accuracy of both STAT and FFT chord feature.

Chord Features	Accuracy(%)
STAT	**75.0**
FFT	58.3

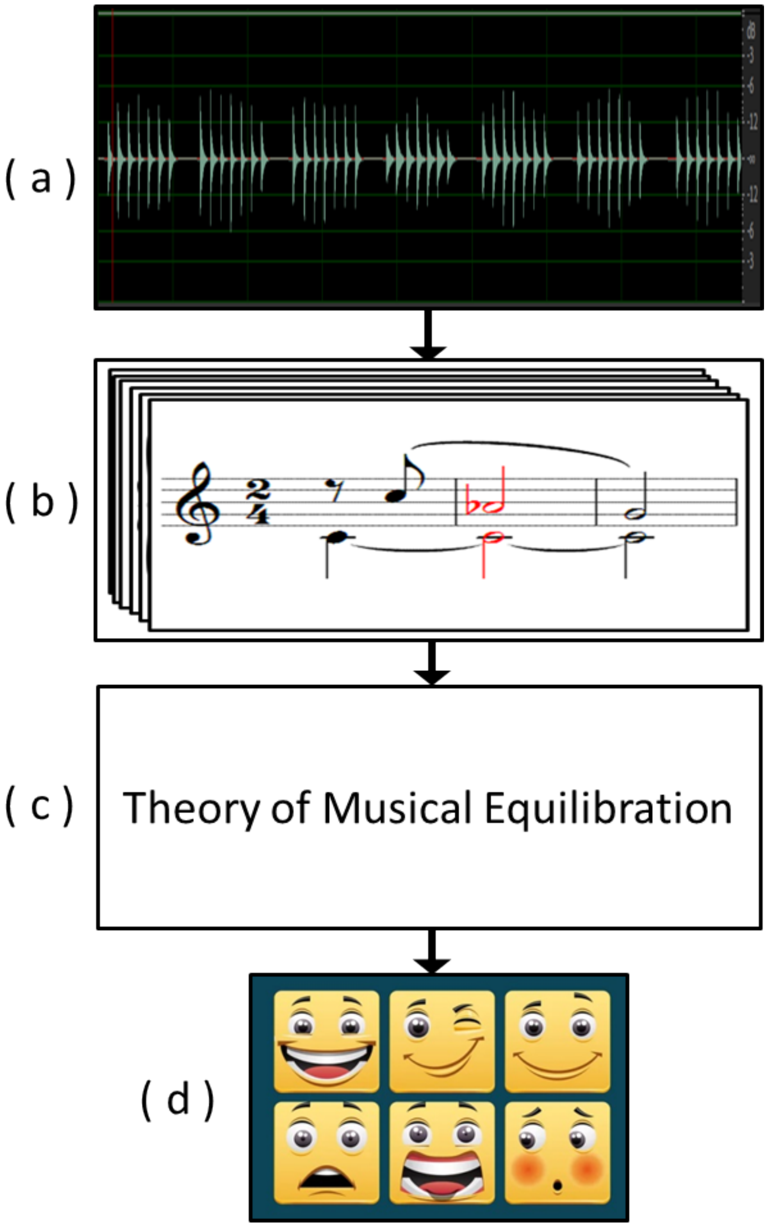


Figure 5.20: Emotion classification based on the chord. (a. Original music signals; b. The identified chords; c. The Theory of Musical Equilibration; d. The emotions [14])

The research in this chapter propose a new approach to use the chord theory – Theory of Musical Equilibration building a system for recognise emotions from music signals. In summary, the research in this chapter is based on Theory of Musical Equilibration in the music theory book "Music and Emotions" which designed and conducted by Daniela and Bernd Willimek [77], in order to identified the chords which is selected as the emotional carrier. This research start as building a new chord database which use the Adobe audition to extract the six different kinds of chords clip from the piano chord teaching audio. Then the FFT features based on the 1000 points sampling pre-process data and STAT features were extracted for the selected features from the database. In order to identified the classification of the chords in database, the selected features of theses chords would compare with the benchmark chord by Euclidean distance and correlation. After calculating and comparison the Euclidean distance and correlation, the results shown the STAT features work well in most of chords except the Augmented chord. The STAT features resulting in almost wrong results in Augmented chord. On the other hands, the FFT features provide the best results in Augmented chord, Minor chord and Minor 6th chord. However, the other three chord basically was not work on it. The final result shown that the STAT features achieves 75% accuracy in identification performance, which is 18% higher than FFT feature.

Chapter 6

Conclusion and Future Works

6.1 Conclusion

MER is an interesting part of the field of AI. After years of experimentation and improvement by researchers, how to improve the components of the system more effectively and improve the recognition rate has become the top priority of the research. Among the many research directions, data features and machine learning algorithm models are the most important part of the whole system and the two parts that have the greatest impact on the system. As an important component to describe and display the characteristics of the whole data, data features play an important role in the emotion recognition system. Data characteristics are as important to data as height, weight and gender are the same for a human. Reasonable use of data features can be very efficient and simple to describe a specified type of data. So an excellent feature can play a decisive role if it can be used in an appropriate recognition system. At the same time, the machine learning algorithm model is also not negligible. If the data feature is an important indicator of the data, the machine model algorithm model is a guide of features using for machine. An excellent algorithm model is a reasonable operation for the known conditions of different recognition tasks. The algorithm model is different when it works on different databases; different recognition targets and different identification requirements. The core of this research is how to properly use or improve a suitable algorithm model. Similarly, through literature reading, it is also an effective method to analyze the characteristics of the data to be identified and the identity of the target itself, and then use these analysis results to transform and simplify the recognition process.

In this thesis, three research directions are explored for the combination of data characteristics, machine learning models and music theory. In terms of data characteristics, new feature selections and combinations are used and combined with the RF machine learning model, the performance of the system is effectively improved. In terms of machine learning models, the most advanced deep learning model was chosen as the research direction. The LSTM, which has a very significant effect on dynamic continuous signal recognition, was tested as a research target. Research has been carried out from feature processing suitable for models to model parameter selection. Although the recognition rate is not ideal, the LSTM model has a very thorough understanding. In music theory, the relationship between chords and emotions is used as the core of the system. The system implements the identification of chords. Thus, the relationship between chord and emotion can be used to infer the emotion that music containing chords should have. Next is the details of the conclusions for three aspects of the study.

6.1.1 Feature Combination and RF Classification

As can be seen from the study in chapter 3, as a first step in the MER system, an excellent feature can help the performance of the system. System performance has improved significantly after using two different features including selected classical audio feature features and statistical features. In the study of chapter 3, because the goal is to identify and classify music signals, the selected features can embody various characteristics of the audio signal and in the comparison of the no-select feature and the selected feature, it is shown that the feature selection on feature extraction, which focuses on the target of the recognition task, will greatly improve the accuracy of the system. The importance of feature selection is thus evident and in the study, it was found that the accuracy of using only one of the features alone is much lower than the combination of the two features. This shows that the binding feature contains more information that is helpful for identification. It's as if it is wanted to classify a fruit from a bunch of fruits, it's easier to distinguish between the taste and the shape than just using one. However, it cannot be said that the more features, the better. Because the repeated part or the useless information part of the various features will also increase when the features increase. Duplicate and useless information can reduce the accuracy of the system. Therefore, a reasonable combination of complementary data features will improve system performance.

6.1.2 Dynamic Deep Learning Models

For Chapter 4, from the final result, using the LSTM deep learning network model to extract data based on basically the same features, the recognition efficiency of the system is not better than using the RF machine learning model. According to the analysis of the experimental results at the end of the study, there may be two reasons for the results. First, due to the special requirements of the LSTM network model for input data attributes, there are deficiencies in data preprocessing. For example, the input data may be too long in length, or a more appropriate windowed segmentation method should be employed. In addition, it is also possible that feature extraction should not be performed in advance, resulting in the destruction of dynamic time-related information contained in the data itself. Second, the parameter settings are not adjusted to the most perfect value. Since parameter adjustments are affected by the input data pattern and network structure, different inputs require different parameters. Therefore, in the case that the input data may have drawbacks, the parameter settings may not be adjusted to the optimal value. Third, since the attributes of the original data itself belong to a continuous repeating type of music signal, it is possible that such a signal does not fit perfectly with the LSTM network as a voice signal. It is possible to erroneously learn incorrect weighting when the network trains continuously repeating data information.

6.1.3 Music Emotion Identification Through Chords

The inspiration for Chapter 5 of research comes from a book on music theory about the relationship between chords and emotions. In the book, the author puts forward a series of conclusions that are closely related to chords and emotions according to the theory of the influence of music on human emotions and these conclusions are combined with the fragments of famous music works as an example, and the investigation of the feelings of children after listening to the works confirms the authenticity of this conclusion. Based on this conclusion, an idea that can judge the musical emotion based on the type of chord in the music is proposed by me. If this idea proves to be feasible, it will be of great help in identifying the emotions in current music.

It can be seen from the experimental results that two different features can basically identify the exact chord type and, although the two features have their own advantages and disadvantages, they can complement each other. However, since the dimension of the FFT feature is too high, the dimension reduction process is required before the combination can be performed. Two different calculation comparison methods also showed similar results, indicating that the experimental results are reliable.

By summarizing the research of each chapter on each goal, it is found that the goals and objectives set in the initial stage of the research have been basically completed, and some studies have shown a certain contribution to the field of MER. In the aspect of extracting and combining data features, the new features can improve the accuracy of emotion recognition in music signals. At the same time, the conclusions and results of this part have greatly improved and helped the research of the latter two parts. In the study of dynamic deep learning models, the study of the temporal continuity and correlation of signal data has helped the understanding of the LSTM model network and in practice, it is also not recognized that the data of each dynamic time continuous signal can be effectively recognized by LSTM. Finally, in the study of the part of musical emotion identification in chords, music theory has proved to be an important part of the musical emotion recognition system. Reasonable using music theory to assist in identifying emotions in music will effectively improve the accuracy of recognition. Although this part of the study has not been finalized until the end of the Ph. D.

6.2 Future Works

While scrutinizing the results of these successful studies, there are still some shortcomings and flaws in these studies that cannot be ignored. These shortcomings will be future improvements and research directions. In terms of data characteristics, the rapid development of AI and mathematical theory will bring more new directions and possibilities for feature extraction every day. Effective use of these new knowledge and theories to extract more outstanding features will be the next step. Since this part of the research was my initial work, by reading the latest literature, it is found that more and more new, more effective, and more streamlined features have sprung up. Combining current new technologies and theories, discovering new and enhanced features of the MER system will be highlighted in the future.

For deep learning, the new darling of AI, in so many deep learning models, how to choose and use them reasonably will be the main research direction in the future. Of course, before that, the problems mentioned in the previous section that exist in the LSTM model study will be solved first. New features and segmentation combinations [226] will be tried. For example, Interspeech conference provide more extracted features for audio signals recent year in different approaches [227]. There is a new feature extracted based on the musical texture and expressive techniques, which is proposed by Renato Panda [54] last year and some methods for features fusion [228] may help as well in combination part. Also, the research will try to improve the network structure and composition of the LSTM model, making it more suitable for the APM database. For example, although the single CNN and LSTM both shown the results are not good enough. If combine these two deep learning methods together, using the image process think. The results will be better [229] [230]? Or just a improvement of the LSTM for music sequence [231]. At the same time, the new database, such as ISMIR database [226], will also be used to experiment with the effects of different types of dynamic continuous time signals on the LSTM network. It is hoped that an effective MER system based on the LSTM model can be completed after research and improvement.

For the MER based on chord, CNN features extraction methods will be test in the database and CNN will also be tried as the classifier for the spectrum features. At the same time, for the music theory, it was found that some part of the theory is really help for the MER from the research [232]. Music has its own set of theoretical rules just like any language in the world. Understanding and combining these theories to do research will definitely find an effective approach to improve the emotional recognition of music. Research on the recognition of emotions in music through chords will continue. The next step will be studying how to identify chord segments in a complete piece of music and through experiments, the recognition target of the chord emotion recognition method is emotional classification or emotional regression. It is hoped that the final emotion recognition method using chords can assist the conventional MER method to improve the recognition rate. Later, other music theory will also be studied as an alternative auxiliary theory of musical emotion recognition.

Finally, the ultimate goal of all research is to benefit human society and advance human civilisation. My research is no exception. When my research results are basically mature, these research techniques will be used in social practice. Musical emotion recognition system has great practical value in human society. It can effectively solve some problems, make people's life more convenient and happy, and promote the development of science and technology in the field of AI [50]. For example, the basic function of the MER is to automatic class the new music in website and label them by their styles; For the medicine field, the research may use to recognised and filter the sad music for the depressed patient or anger music for Mania patient; and it can be used to detect some key points of the strongest emotion in the music. Then extract them for other application, such as EEG research.It is hoped that my research can benefit people.

Bibliography

[1] Freepik. Colored background of abstract sound wave free vector. https://www.freepik.com/free-vector/colored-background-of-abstract-sound-wave_1112283.htm#term=sound&page=1&position=5. Accessed DEC 5, 2018.

[2] ALDECAstudio / Fotolia. Speakers hesitate or make brief pauses filled with sounds like ”uh” or ”uhm” mostly before nouns. https://www.sciencedaily.com/releases/2018/05/180514151926.htm. Accessed DEC 5, 2018.

[3] bloginonline.com. Outer ear middle ear inner ear. https://bloginonline.com/outer-ear-middle-ear-inner-ear/. Accessed Nov 11, 2018.

[4] Paige Tutt. Emotional patterns dredging and drowning your life in negativity that you can break today. https://www.rebelcircus.com/blog/emotional-patterns-dredging-drowning-life-negativity-can-break-today/3/. Accessed Dec 20, 2017.

[5] Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734, 2005.

[6] Klaus R Scherer, Vera Shuman, Johnny Fontaine, and Cristina Soriano Salinas. The grid meets the wheel: Assessing emotional feeling via self-report. 2013.

[7] Roger Jang. 5-3 zero crossing rate (zcr). http://mirlab.org/jang/books/audiosignalprocessing/basicFeatureZeroCrossingRate.asp?title=5-3$%$20Zero$%$20Crossing$%$20Rate$%$20($%$B9L$%$B9s$%$B2v)&language=all. Accessed Mar 10, 2018.

[8] Laszlo Kozma. k nearest neighbors algorithm (knn). *Helsinki University of Technology*, 2008.

[9] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[10] Denny Brity. Recurrent neural networks tutorial, part 1 introduction to rnns. http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/. Accessed Dec 4, 2017.

[11] Soo Yong Lim. Education for electromagnetics: Introducing electromagnetics as an appetizer course for computer science and it undergraduates [education column]. *IEEE Antennas and Propagation Magazine*, 56(5):216–222, 2014.

[12] Colah. Understanding lstm networks. http://colah.github.io/posts/2015-08-Understanding-LSTMs/. Accessed April 4, 2018.

[13] Julia. Music and emotions. https://www.shutterstock.com/g/carbouval, 2016.

[14] Kar. Emoticons. https://www.shutterstock.com/g/carbouval.

[15] Colwyn Trevarthen. Origins of musical identity: Evidence from infancy for musical social awareness. *Musical identities*, pages 21–38, 2002.

[16] Jens Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[17] Thomas D Rossing and Peter W Stephens. The science of sound. *The Physics Teacher*, 29:64–64, 1991.

[18] Patrik N Juslin and John A Sloboda. *Music and emotion: Theory and research.* Oxford University Press, 2001.

[19] Nils Lennart Wallin, Björn Merker, and Steven Brown. *The origins of music*. MIT press, 2001.

[20] J Peter Burkholder and Donald Jay Grout. *A History of Western Music: Ninth International Student Edition*. WW Norton & Company, 2014.

[21] Friedrich Blume. *Renaissance and baroque music: a comprehensive survey*. WW Norton, 1967.

[22] Anne Draffkorn Kilmer, Richard L Crocker, and Robert Reginald Brown. *Sounds from silence: Recent discoveries in ancient Near Eastern music*. Bit Enki Records, 1976.

[23] Ian Cross. Music, cognition, culture, and evolution. *Annals of the New York Academy of sciences*, 930(1):28–42, 2001.

[24] John Morgan O'Connell and Salwa El-Shawan Castelo-Branco. *Music and conflict*. University of Illinois Press, 2010.

[25] Carroll E Izard. *Human emotions*. Springer Science & Business Media, 2013.

[26] Kate Hevner. Expression in music: a discussion of experimental studies and theories. *Psychological review*, 42(2):186, 1935.

[27] William James. What is an emotion? *Mind*, 9(34):188–205, 1884.

[28] Yi-Hsuan Yang and Homer H Chen. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):40, 2012.

[29] Jacques Ferber and Gerhard Weiss. *Multi-agent systems: an introduction to distributed artificial intelligence*, volume 1. Addison-Wesley Reading, 1999.

[30] Wendy B Rauch-Hindin. *Artificial Intelligence in Business, Science, and Industry: Fundamentals*. Prentice-Hall New Jersey, 1986.

[31] Cynthia Breazeal and Brian Scassellati. Robots that imitate humans. *Trends in cognitive sciences*, 6(11):481–487, 2002.

[32] David D Luxton. Artificial intelligence in psychological practice: Current and future applications and implications. *Professional Psychology: Research and Practice*, 45(5):332, 2014.

[33] Mariusz Flasiński. *Introduction to artificial intelligence*. Springer, 2016.

[34] Deborah Tannen and Cynthia Wallat. Interactive frames and knowledge schemas in interaction: Examples from a medical examination/interview. *Social psychology quarterly*, pages 205–216, 1987.

[35] Vimla L Patel, Edward H Shortliffe, Mario Stefanelli, Peter Szolovits, Michael R Berthold, Riccardo Bellazzi, and Ameen Abu-Hanna. The coming of age of artificial intelligence in medicine. *Artificial intelligence in medicine*, 46(1):5–17, 2009.

[36] Gordon McCalla. The fragmentation of culture, learning, teaching and technology: implications for the artificial intelligence in education research agenda in 2010. *International Journal of Artificial Intelligence in Education*, 11(2):177–196, 2000.

[37] Michael Wollowski, Todd Neller, and James Boerkoel. Artificial intelligence education. *AI Magazine*, 38(2):5, 2017.

[38] Steve Woolgar. Why not a sociology of machines? the case of sociology and artificial intelligence. *Sociology*, 19(4):557–572, 1985.

[39] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 435–442. ACM, 2015.

[40] Richard Jiang, Anthony TS Ho, Ismahane Cheheb, Noor Al-Maadeed, Somaya Al-Maadeed, and Ahmed Bouridane. Emotion recognition from scrambled facial images via many graph embedding. *Pattern Recognition*, 67:245–251, 2017.

[41] Sriparna Saha, Shreyasi Datta, Amit Konar, and Ramadoss Janarthanan. A study on emotion recognition from body gestures using kinect sensor. In *2014 International Conference on Communication and Signal Processing*, pages 056–060. IEEE, 2014.

[42] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian. Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):3030–3043, 2018.

[43] Olivier Lahaie, Roch Lefebvre, and Philippe Gournay. Influence of audio bandwidth on speech emotion recognition by human subjects. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 61–65. IEEE, 2017.

[44] Robert Jenke, Angelika Peer, and Martin Buss. Feature extraction and selection for emotion recognition from eeg. *IEEE Transactions on Affective Computing*, 5(3):327–339, 2014.

[45] Suwicha Jirayucharoensak, Setha Pan-Ngum, and Pasin Israsena. Eeg-based emotion recognition using deep learning network with principal component based covariate shift adaptation. *The Scientific World Journal*, 2014, 2014.

[46] Değer Ayata, Yusuf Yaslan, and Mustafa Kamaşak. Emotion recognition via random forest and galvanic skin response: Comparison of time based feature sets, window sizes and wavelet approaches. In *Medical Technologies National Congress (TIPTEKNO), 2016*, pages 1–4. IEEE, 2016.

[47] S Jerritta, M Murugappan, R Nagarajan, and Khairunizam Wan. Physiological signals based human emotion recognition: a review. In *Signal Processing and its Applications (CSPA), 2011 IEEE 7th International Colloquium*, pages 410–415. IEEE, 2011.

[48] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *Proc. ISMIR*, pages 255–266. Citeseer, 2010.

[49] Dan Liu, Lie Lu, and Hong-Jiang Zhang. Automatic mood detection from acoustic music data. 2003.

[50] Yi-Hsuan Yang and Homer H Chen. *Music emotion recognition*. CRC Press, 2011.

[51] Wang Muyuan, Zhang Naiyao, and Zhu Hancheng. User-adaptive music emotion recognition. In *Signal Processing, 2004. Proceedings. ICSP'04. 2004 7th International Conference on*, volume 2, pages 1352–1355. IEEE, 2004.

[52] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H Chen. A regression approach to music emotion recognition. *IEEE Transactions on audio, speech, and language processing*, 16(2):448–457, 2008.

[53] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[54] Renato Panda, Ricardo Manuel Malheiro, and Rui Pedro Paiva. Novel audio features for music emotion recognition. *IEEE Transactions on Affective Computing*, (1):1–1, 2018.

[55] Byeong-jun Han, Seungmin Rho, Roger B Dannenberg, and Eenjun Hwang. Smers: Music emotion recognition using support vector regression. In *ISMIR*, pages 651–656, 2009.

[56] Byungsoo Jeon, Chanju Kim, Adrian Kim, Dongwon Kim, Jangyeon Park, and Jung-Woo Ha. Music emotion recognition via end-to-end multimodal neural networks. *RECSYS Google Scholar*, 2017.

[57] Miroslav Malik, Sharath Adavanne, Konstantinos Drossos, Tuomas Virtanen, Dasa Ticha, and Roman Jarina. Stacked convolutional and recurrent neural networks for music emotion recognition. *arXiv preprint arXiv:1706.02292*, 2017.

[58] Yu-Hao Chin, Yi-Zeng Hsieh, Mu-Chun Su, Shu-Fang Lee, Miao-Wen Chen, and Jia-Ching Wang. Music emotion recognition using pso-based fuzzy hyper-rectangular composite neural networks. *IET Signal Processing*, 11(7):884–891, 2017.

[59] Eric Eaton, Sven Koenig, Claudia Schulz, Francesco Maurelli, John Lee, Joshua Eckroth, Mark Crowley, Richard G Freedman, Rogelio E Cardona-Rivera, Tiago Machado, et al. Blue sky ideas in artificial intelligence education from the eaai 2017 new and future ai educator program. *AI Matters*, 3(4):23–31, 2018.

[60] Seong Ho Park and Kyunghwa Han. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*, 286(3):800–809, 2018.

[61] Elsa B Kania. *Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power*. Center for a New American Security, 2017.

[62] Junhui Li, Yang Zhao, Chenjun Sun, Xichun Bao, Qi Zhao, and Haiming Zhou. A survey of development and application of artificial intelligence in smart grid. In *IOP Conference Series: Earth and Environmental Science*, volume 186, page 012066. IOP Publishing, 2018.

[63] Tong Liu, Li Han, Liangkai Ma, and Dongwei Guo. Audio-based deep music emotion recognition. In *AIP Conference Proceedings*, volume 1967, page 040021. AIP Publishing, 2018.

[64] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou, and Ioannis Giannoukos. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2):155–177, 2015.

[65] Theodor W Adorno. *Beethoven: The philosophy of music*. John Wiley & Sons, 2018.

[66] Jonghwa Kim and Elisabeth André. Emotion recognition based on physiological changes in music listening. *IEEE transactions on pattern analysis and machine intelligence*, 30(12):2067–2083, 2008.

[67] Lucanne Magill Bailey. The use of songs in music therapy with cancer patients and their families. *Music Therapy*, 4(1):5–17, 1984.

[68] Jacek Grekow. Music emotion maps in the arousal-valence space. In *From Content-based Music Emotion Recognition to Emotion Maps of Musical Pieces*, pages 95–106. Springer, 2018.

[69] Anthony Storr. *Music and the Mind*. Simon and Schuster, 2015.

[70] William Forde Thompson. *Music, thought, and feeling: Understanding the psychology of music*. Oxford university press, 2015.

[71] Jenny Preece, Yvonne Rogers, and Helen Sharp. *Interaction design: beyond human-computer interaction*. John Wiley & Sons, 2015.

[72] Annabel J Cohen. Music as a source of emotion in film. *Music and emotion: Theory and research*, pages 249–272, 2001.

[73] Lola L Cuddy and Jacalyn Duffin. Music, memory, and alzheimers disease: is music recognition spared in dementia, and how can it be assessed? *Medical hypotheses*, 64(2):229–235, 2005.

[74] Kenneth D Miller and David JC MacKay. The role of constraints in hebbian learning. *Neural Computation*, 6(1):100–126, 1994.

[75] Takuya Fujishima. Real-time chord recognition of musical sound: A system using common lisp music. *Proc. ICMC, Oct. 1999*, pages 464–467, 1999.

[76] Aniruddh D Patel. Language, music, syntax and the brain. *Nature neuroscience*, 6(7):674, 2003.

[77] Bernd Willimek and Daniela Willimek. *Music and Emotions-Research on the Theory of Musical Equilibration (die Strebetendenz-Theorie)*. Bernd Willimek, 2013.

[78] Gang Sheng. *Vehicle noise, vibration, and sound quality*. 2012.

[79] Jacques Moeschler. Speech act theory and the analysis of conversation. *Essays in speech act theory*, 77:239–262, 2002.

[80] Suvi Saarikallio and Jaakko Erkkilä. The role of music in adolescents' mood regulation. *Psychology of music*, 35(1):88–109, 2007.

[81] Gordon C Bruner. Music, mood, and marketing. *the Journal of marketing*, pages 94–104, 1990.

[82] Lucy Green. *How popular musicians learn: A way ahead for music education*. Routledge, 2017.

[83] David Brian Williams and Peter Richard Webster. *Music Technology*. Academic Press, 1996.

[84] Neville H Fletcher and Thomas D Rossing. *The physics of musical instruments*. Springer Science & Business Media, 2012.

[85] Ben Gold, Nelson Morgan, and Dan Ellis. *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons, 2011.

[86] Alexander Lerch. *An introduction to audio content analysis: Applications in signal processing and music informatics*. Wiley-IEEE Press, 2012.

[87] JM Blackledge. Digital signal processing: Mathematical and computation methods: Software development and applications. *London: Horwood Publishing Limited*, 2006.

[88] Peter W Alberti. The anatomy and physiology of the ear and hearing. *Occupational exposure to noise: Evaluation, prevention, and control*, pages 53–62, 2001.

[89] Christopher Platt and Arthur N Popper. Fine structure and function of the ear. In *Hearing and sound communication in fishes*, pages 3–38. Springer, 1981.

[90] Steven Errede. The human ear hearing, sound intensity and loudness levels. *UIUC Physics*, 406, 2002.

[91] Hallowell Davis and Sol Richard Silverman. *Hearing and deafness*. Holt, Rinehart & Winston of Canada Ltd, 1970.

[92] John R Franks. Hearing measurement. *Occupational Exposure to Noise: Evaluation, Prevention and Control. Geneva: World Health Organisation*, pages 183–231, 2001.

[93] Elaine Fox. *Emotion science: An integration of cognitive and neuroscience approaches*. Palgrave Macmillan, 2008.

[94] Stanley Schachter and Jerome Singer. Cognitive, social, and physiological determinants of emotional state. *Psychological review*, 69(5):379, 1962.

[95] Peter Goldie. Emotion. *Philosophy Compass*, 2(6):928–938, 2007.

[96] Lisa Feldman Barrett, Michael Lewis, and Jeannette M Haviland-Jones. *Handbook of emotions*. Guilford Publications, 2016.

[97] Klaus R Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[98] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[99] John A Sloboda and Patrik N Juslin. Psychological perspectives on music and emotion. 2001.

[100] Robert Plutchik. *Emotions and life: Perspectives from psychology, biology, and evolution.* American Psychological Association, 2003.

[101] Robert Plutchik. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984:197–219, 1984.

[102] Robert E Thayer. *The biopsychology of mood and arousal*. Oxford University Press, 1990.

[103] Charles E Osgood. The nature and measurement of meaning. *Psychological bulletin*, 49(3):197, 1952.

[104] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[105] Cyril Laurier, Mohamed Sordo, Joan Serra, and Perfecto Herrera. Music mood representations from social tags. In *ISMIR*, pages 381–386, 2009.

[106] Zhongzhe Xiao, Emmanuel Dellandréa, Liming Chen, and Weibei Dou. Recognition of emotions in speech by a hierarchical approach. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference*, pages 1–8. IEEE, 2009.

[107] Hatice Gunes and Maja Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, 1(1):68–99, 2010.

[108] Kerstin Bischoff, Claudiu S Firan, Raluca Paiu, Wolfgang Nejdl, Cyril Laurier, and Mohamed Sordo. Music mood and theme classification-a hybrid approach. In *ISMIR*, pages 657–662, 2009.

[109] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[110] Alicja A Wieczorkowska. Towards extracting emotions from music. In *Intelligent Media Technology for Communicative Intelligence*, pages 228–238. Springer, 2005.

[111] Yi-Hsuan Yang, Chia-Chu Liu, and Homer H Chen. Music emotion classification: a fuzzy approach. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 81–84. ACM, 2006.

[112] Yazhong Feng, Yueting Zhuang, and Yunhe Pan. Popular music retrieval by detecting mood. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 375–376. ACM, 2003.

[113] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen. A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):448–457, Feb 2008. doi:10.1109/TASL.2007.911513.

[114] Tao Li and Mitsunori Ogihara. Content-based music similarity search and emotion detection. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference*, volume 5, pages V–705. IEEE, 2004.

[115] Densil Cabrera et al. Psysound: A computer program for psychoacoustical analysis. In *Proceedings of the Australian Acoustical Society Conference*, volume 24, pages 47–54, 1999.

[116] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, Jul 2002. doi:10.1109/TSA.2002.800560.

[117] Dimitri P Solomatine and Durga L Shrestha. Adaboost. rt: a boosting algorithm for regression problems. In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference*, volume 2, pages 1163–1168. IEEE, 2004.

[118] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[119] Ashish Sen and Muni Srivastava. *Regression analysis: theory, methods, and applications*. Springer Science & Business Media, 2012.

[120] A. Hanjalic and Li-Qun Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, Feb 2005. doi:10.1109/TMM.2004.840618.

[121] Emery Schubert. Measurement and time series analysis of emotion in music. 1999.

[122] Mark D Korhonen, David A Clausi, and M Ed Jernigan. Modeling emotional content of music using system identification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(3):588–599, 2005.

[123] Jeff A Bilmes et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126, 1998.

[124] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623, 2003.

[125] Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE transactions on audio, speech, and language processing*, 17(4):582–596, 2009.

[126] Yi-Lin Lin and Gang Wei. Speech emotion recognition based on hmm and svm. In *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference*, volume 8, pages 4898–4901. IEEE, 2005.

[127] Ronald Newbold Bracewell and Ronald N Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986.

[128] Siqing Wu, Tiago H Falk, and Wai-Yip Chan. Automatic speech emotion recognition using modulation spectral features. *Speech communication*, 53(5):768–785, 2011.

[129] Yuan-Pin Lin, Chi-Hong Wang, Tzyy-Ping Jung, Tien-Lin Wu, Shyh-Kang Jeng, Jeng-Ren Duann, and Jyh-Horng Chen. Eeg-based emotion recognition in music listening. *IEEE Transactions on Biomedical Engineering*, 57(7):1798–1806, 2010.

[130] Lijiang Chen, Xia Mao, Yuli Xue, and Lee Lung Cheng. Speech emotion recognition: Features and classification models. *Digital signal processing*, 22(6):1154–1160, 2012.

[131] Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn Schuller, Cate Cox, Ellen Douglas-Cowie, and Roddy Cowie. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. 9th Interspeech 2008 incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia*, pages 597–600, 2008.

[132] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9-10):1162–1171, 2011.

[133] Chien Shing Ooi, Kah Phooi Seng, Li-Minn Ang, and Li Wern Chew. A new approach of audio emotion recognition. *Expert systems with applications*, 41(13):5858–5869, 2014.

[134] Xiaoling Yang, Baohua Tan, Jiehua Ding, Jinye Zhang, and Jiaoli Gong. Comparative study on voice activity detection algorithm. In *Electrical and Control Engineering (ICECE), 2010 International Conference*, pages 599–602. IEEE, 2010.

[135] Steven B Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *Readings in speech recognition*, pages 65–74. Elsevier, 1990.

[136] Sigurdur Sigurdsson, Kaare Brandt Petersen, and Tue Lehn-Schiøler. Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music. In *ISMIR*, pages 286–289, 2006.

[137] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974.

[138] James R Williamson, Thomas F Quatieri, Brian S Helfer, Rachelle Horwitz, Bea Yu, and Daryush D Mehta. Vocal biomarkers of depression based on motor incoordination. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 41–48. ACM, 2013.

[139] Dmitri Bitouk, Ragini Verma, and Ani Nenkova. Class-level spectral features for emotion recognition. *Speech communication*, 52(7-8):613–625, 2010.

[140] Konstantin Markov, Motofumi Iwata, and Tomoko Matsui. Music emotion recognition using gaussian processes. In *Working Notes Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, October 18-19, CEUR-WS. org, ISSN 1613-0073*, 2013.

[141] Anna Aljanaki, Frans Wiering, and Remco C Veltkamp. Mirutrecht participation in mediaeval 2013: Emotion in music task. In *Working Notes Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, October 18-19, CEUR-WS. org, ISSN 1613-0073*, 2013.

[142] Chris Cannam, Michael O. Jewell, Christophe Rhodes, Mark Sandler, and Mark d'Inverno. Linked data and you: Bringing music research software into the semantic web. *Journal of New Music Research*, 39(4):313–325, 2010.

[143] Olivier Lartillot and Petri Toiviainen. A matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects*, pages 237–244, 2007.

[144] Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classifiers and fmri: a tutorial overview. *Neuroimage*, 45(1):S199–S209, 2009.

[145] Robert Hecht-Nielsen. Theory of the backpropagation neural network. In *Neural networks for perception*, pages 65–93. Elsevier, 1992.

[146] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[147] Gerald Tesauro. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–69, 1995.

[148] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[149] Godfried Toussaint. Geometric proximity graphs for improving nearest neighbor methods in instance-based learning and data mining. *International Journal of Computational Geometry & Applications*, 15(02):101–150, 2005.

[150] Oliver Sutton. Introduction to k nearest neighbour classification and condensed nearest neighbour data reduction. *University lectures, University of Leicester*, 2012.

[151] Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. Avec 2011–the first international audio/visual emotion challenge. In *Affective Computing and Intelligent Interaction*, pages 415–424. Springer, 2011.

[152] Hongying Meng and Nadia Bianchi-Berthouze. Naturalistic affective expression classification by a multi-stage approach based on hidden markov models. In *Affective computing and intelligent interaction*, pages 378–387. Springer, 2011.

[153] Tin Kam Ho. Random decision forests. In *Document analysis and recognition, 1995., proceedings of the third international conference*, volume 1, pages 278–282. IEEE, 1995.

[154] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[155] Tom M Mitchell. Logistic regression. *Machine learning*, 10:701, 2005.

[156] Eamonn Keogh. Naive bayes classifier. *Accessed: Nov*, 5:2017, 2006.

[157] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[158] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

[159] Kunihiko Fukushima. Neocognitron–a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *NHK Broadcasting Science Laboratory Report*, (15):p106–115, 1981.

[160] Feiyan Zhou, Linpeng Jin, and Jun Dong. Review of convolutional neural networks research. *Chinese Journal of Computers*, 40(6):1229–1251, 2017.

[161] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, and Gerald Penn. Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference*, pages 4277–4280. IEEE, 2012.

[162] Danilo P Mandic and Jonathon Chambers. *Recurrent neural networks for prediction: learning algorithms, architectures and stability*. John Wiley & Sons, Inc., 2001.

[163] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990.

[164] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. Emotion in music task at mediaeval 2015. In *MediaEval*, 2015.

[165] Thomas Pellegrini and Valentin Barrière. Time-continuous estimation of emotion in music with recurrent neural networks. In *MediaEval 2015 Multimedia Benchmark Workshop (MediaEval 2015)*, pages pp–1, 2015.

[166] Eduardo Coutinho, Felix Weninger, Björn Schuller, and Klaus R Scherer. The munich lstm-rnn approach to the mediaeval 2014" emotion in music'" task. In *MediaEval*. Citeseer, 2014.

[167] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José R Zapata, and Xavier Serra. Essentia: An audio analysis library for music information retrieval. In *ISMIR*, pages 493–498. Citeseer, 2013.

[168] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[169] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[170] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth annual conference of the international speech communication association*, 2014.

[171] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. Hybrid speech recognition with deep bidirectional lstm. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop*, pages 273–278. IEEE, 2013.

[172] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*, 2012.

[173] Paul Lamere. Social tagging and music information retrieval. *Journal of new music research*, 37(2):101–114, 2008.

[174] Shiliang Zhang, Qingming Huang, Qi Tian, Shuqiang Jiang, and Wen Gao. i. mtv: an integrated system for mtv affective analysis. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 985–986. ACM, 2008.

[175] Shiliang Zhang, Qi Tian, Shuqiang Jiang, Qingming Huang, and Wen Gao. Affective mtv analysis based on arousal and valence features. In *Multimedia and Expo, 2008 IEEE International Conference*, pages 1369–1372. IEEE, 2008.

[176] Shiliang Zhang, Qi Tian, Qingming Huang, Wen Gao, and Shipeng Li. Utilizing affective analysis for efficient movie browsing. In *Image Processing (ICIP), 2009 16th IEEE International Conference*, pages 1853–1856. IEEE, 2009.

[177] Cyril Laurier and Perfecto Herrera. Mood cloud: A real-time music mood visualization tool. *Proceedings of the Computer Music Modeling and Retrieval*, 2008.

[178] Cyril Laurier, Mohamed Sordo, and Perfecto Herrera. Mood cloud 2.0: Music mood browsing based on social networks. In *Proceedings of the 10th International Society for Music Information Conference (ISMIR 2009), Kobe, Japan*. Citeseer, 2009.

[179] Sasank Reddy and Jeff Mascia. Lifetrak: music in tune with your life. In *Proceedings of the 1st ACM international workshop on Human-centered multimedia*, pages 25–34. ACM, 2006.

[180] Michael S Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2(1):1–19, 2006.

[181] Tien-Lin Wu and Shyh-Kang Jeng. Probabilistic estimation of a novel music emotion model. In *International Conference on Multimedia Modeling*, pages 487–497. Springer, 2008.

[182] Chin-Han Chen, Ming-Fang Weng, Shyh-Kang Jeng, and Yung-Yu Chuang. Emotion-based music visualization using photos. In *International Conference on Multimedia Modeling*, pages 358–368. Springer, 2008.

[183] Braja G Patra, Dipankar Das, and Sivaji Bandyopadhyay. Music emotion recognition system. In *Proceedings of the international symposium frontiers of research speech and music (FRSM-2015)*, pages 114–119, 2015.

[184] Björn Schuller, Stefan Steidl, Anton Batliner, et al. The interspeech 2009 emotion challenge. In *INTERSPEECH*, volume 2009, pages 312–315, 2009.

[185] Florian Eyben, Felix Weninger, Martin Wöllmer, and Björn Schuller. open-source media interpretation by large feature-space extraction.

[186] Emmanuel Dellandréa, Liming Chen, Yoann Baveye, Mats Sjöberg, Christel Chamaret, and ECD Lyon. The mediaeval 2016 emotional impact of movies task. In *Proc. of the MediaEval 2016 Workshop, Hilversum, Netherlands*, 2016.

[187] Shizhe Chen and Qin Jin. Ruc at mediaeval 2016 emotional impact of movies task: Fusion of multimodal features.

[188] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A Müller, Shrikanth S Narayanan, et al. The interspeech 2010 paralinguistic challenge. In *InterSpeech*, volume 2010, pages 2795–2798, 2010.

[189] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. 2013.

[190] Yajie Hu, Xiaoou Chen, and Deshun Yang. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *ISMIR*, pages 123–128, 2009.

[191] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology, University of Florida, 1999.

[192] Albert Mehrabian and James A Russell. *An approach to environmental psychology.* the MIT Press, 1974.

[193] Luca Mion and Giovanni De Poli. Score-independent audio features for description of music expression. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):458–466, 2008.

[194] Dan Yang and Won-Sook Lee. Disambiguating music emotion using software agents. In *ISMIR*, volume 4, pages 218–223, 2004.

[195] Peter Knees, Elias Pampalk, and Gerhard Widmer. Artist classification with web-based data. In *ISMIR*, 2004.

[196] Muzaffar Khan, Tirupati Goskula, Mohmmed Nasiruddin, and Ruhina Quazi. Comparison between k-nn and svm method for speech emotion recognition. *International Journal on Computer Science and Engineering*, 3(2):607–611, 2011.

[197] A Vinay and Anu Mehra. Vocal emotion recognition using naïve bayes classifier.

[198] Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *ISMIR*, 2000.

[199] Yading Song and D Simon. How well can a music emotion recognition system predict the emotional responses of participants? In *Sound and Music Computing Conference (SMC)*, pages 387–392, 2015.

[200] Boyang Gao. *Contributions to music semantic analysis and its acceleration techniques*. PhD thesis, Ecole Centrale de Lyon, 2014.

[201] Florian Eyben, Martin Woellmer, and Bjoern Schuller. the munich open speech and music interpretation by large space extraction toolkit. 2010.

[202] Kirill Sakhnov, Ekaterina Verteletskaya, and Boris Simak. Approach for energy-based voice detector with adaptive scaling factor. *IAENG International Journal of Computer Science*, 36(4):394, 2009.

[203] Bo Hjorth. Eeg analysis based on time domain properties. *Electroencephalography and clinical neurophysiology*, 29(3):306–310, 1970.

[204] T Mitchell. Generative and discriminative classifiers: naive bayes and logistic regression, 2005. *Manuscript available at http://www. cs. cm. edu/˜ tom/NewChapters. html.*

[205] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[206] Stephen V Stehman. Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, 62(1):77–89, 1997.

[207] Anna Alajanki, Yi-Hsuan Yang, and Mohammad Soleymani. Benchmarking music emotion recognition systems. *PLOS ONE*, pages 835–838, 2016.

[208] Boyang Gao, Emmanuel Dellandréa, and Liming Chen. Music sparse decomposition onto a midi dictionary of musical words and its application to music mood classification. In *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop*, pages 1–6. IEEE, 2012.

[209] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60–68, 2017.

[210] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335, 2008.

[211] Ye Ma, XinXing Li, Mingxing Xu, Jia Jia, and Lianhong Cai. Multi-scale context based attention for dynamic music emotion prediction. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1443–1450. ACM, 2017.

[212] Stefano Pini, Olfa Ben Ahmed, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara, and Benoit Huet. Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 536–543. ACM, 2017.

[213] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(3):34–41, 2012.

[214] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer, 2013.

[215] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas. The enterface05 audio-visual emotion database. In *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference*, pages 8–8. IEEE, 2006.

[216] Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W Schuller. End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5089–5093. IEEE, 2018.

[217] Sander Dieleman and Benjamin Schrauwen. End-to-end learning for music audio. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference*, pages 6964–6968. IEEE, 2014.

[218] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops*, pages 1–8. IEEE, 2013.

[219] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2017.

[220] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. An empirical exploration of recurrent network architectures. In *International Conference on Machine Learning*, pages 2342–2350, 2015.

[221] MathWorks. trainingoptions. https://www.mathworks.com/help/deeplearning/ref/trainingoptions.html;jsessionid=d4461a144108c887d682d3d1de64. Accessed Dec 20, 2017.

[222] MathWorks. Sequence classification using deep learning. https://www.mathworks.com/help/deeplearning/examples/classify-sequence-data-using-lstm-networks.html. Accessed Dec 3, 2017.

[223] Heng-Tze Cheng, Yi-Hsuan Yang, Yu-Ching Lin, I-Bin Liao, Homer H Chen, et al. Automatic chord recognition for music classification and retrieval. In *2008 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1505–1508. IEEE, 2008.

[224] Xinquan Zhou and Alexander Lerch. Chord detection using deep learning. In *Proceedings of the 16th ISMIR Conference*, volume 53, 2015.

[225] Bernd Willimek. die strebetendenz-theorie.. *Tonkünstlerforum Baden-Württemberg*, (29).

[226] Jia Dai, Shan Liang, Wei Xue, Chongjia Ni, and Wenju Liu. Long short-term memory recurrent neural network based segment features for music genre classification. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5. IEEE, 2016.

[227] Björn W Schuller, Stefan Steidl, Anton Batliner, Peter B Marschik, Harald Baumeister, Fengquan Dong, Simone Hantke, Florian Pokorny, Eva-Maria Rathner, Katrin D Bartl-Pokorny, et al. The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats. *Proceedings of INTERSPEECH, Hyderabad, India*, 5, 2018.

[228] Daniel Krause and Konrad Kowalczyk. Multichannel fusion and audio-based features for acoustic event classification. In *Audio Engineering Society Convention 145*. Audio Engineering Society, 2018.

[229] Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. Learning representations from eeg with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*, 2015.

[230] Ivan Himawan, Michael Towsey, and Paul Roe. 3d convolution recurrent neural networks for bird sound detection. 2018.

[231] Adrien Ycart, Emmanouil Benetos, et al. A study on lstm networks for polyphonic music sequence modelling. ISMIR, 2017.

[232] Patrik N Juslin and John Sloboda. *Handbook of music and emotion: Theory, research, applications*. Oxford University Press, 2011.