Title	深層学習法を用いた人間の聴覚特性に基づく音声感情認識
Author(s)	PENG, ZHICHAO
Citation	
Issue Date	2020-09
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/16997
Rights	
Description	Supervisor:赤木　正人，先端科学技術研究科，博士

Doctoral Dissertation

Speech emotion recognition based on human auditory characteristics using deep learning methods

Zhichao Peng

Supervisor: Masato Akagi

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Information Science

September 2020

Abstract

The coming era of the Internet of Everything provides huge development opportunities for the field of human-robot interaction. Speech is the most natural and convenient way for communication between humans and robots. Emotion information from speech can effectively help robots understand the speaker's intentions in natural human-robot interaction. Therefore, speech emotion recognition (SER) is one of the hotspots in current research, which can play an essential role in all human-robot interaction scenarios such as education, medical care, service, etc.

Identifying emotions from speech requires to extract discriminative and robust features that can effectively represent the emotion of speech. However, the traditional acoustic features have problems with weak emotional discrimination and poor noise robustness. The human auditory system can easily perceive the emotional states of speech even in a noisy environment, so this study is to explore auditory representations of computational auditory models and deep learning methods to improve the performance of categorical and dimensional emotion recognition.

Due to the complexity of the human auditory system, the process of speech signal processing is not completely clear, nor which the auditory model can better simulate the human auditory system. Recent psychoacoustic experiments show that temporal modulation cues play an important role in speech perception and contain multi-dimensional spectral-temporal information. Therefore, this study proposes a 3D convolutional neural network (3D CNN) architecture for categorical emotion recognition. In this architecture, 3D CNN is used to extract the discriminative auditory representations from temporal modulation cues by joint spectral-temporal feature learning. The experimental results show that the joint spectral-temporal auditory representations can be extracted using 3D CNN from temporal modulation cues. The results demonstrate that the performance of emotion recognition based on joint spectral-temporal representation can exceed the recognition accuracy compared to that of the state-of-the-art methods.

The high-level auditory representation sequence extracted from 3D CNN is segmented into non-overlapping subsequences, and then LSTM is used to capture the segment-level temporal dependence of subsequences in the previous study. These discontinuous segment-level features cannot fully reflect the dynamic changes of emotions. In addition, existing studies on the attention model only focus on the salient regions of emotion but ignore the continuity of cognition. Inspired by cognitive behavior, this study proposes an attention-based sliding recurrent neural network (ASRNN) to effectively model auditory representation sequence by mimicking the auditory attention to capture salient emotion regions. In the ASRNN model, a high-level feature representation is obtained continuously through a sliding window, and then a temporal attention model is used to capture salient regions of emotion representation. Moreover, a subjective evaluation experiment is designed to analyze the correlation between the temporal attention model and human auditory attention. The results of the experiments showed that this model could effectively obtain emotional information by capturing salient emotion regions using the ASRNN model. The subjective evaluation shows that the temporal attention model is basically consistent with human auditory attention in recognizing emotions.

In categorical emotion recognition, the 3D convolution is used to extract high-level auditory representation from temporal modulation cues. However, the high-dimensional data space through auditory and modulation filtering is not suitable for dimension emotion recognition. Neuroscience research shows that the cortical encoding of natural sounds entails the formation of multiple representations with different spectral and temporal resolution. Inspired by neuroscience, this study proposes a novel auditory feature, namely multi-resolution modulation-filtered cochleagram (MMCG), to capture temporal and contextual modulation cues. Considering that each modulation-filtered cochleagram in MMCG contains different temporal and contextual modulation cues, a parallel LSTM network structure is designed to model multi-temporal dependencies of MMCG and track the temporal dynamics of speech signal sequence for dimensional emotion recognition. Experimental results show that the MMCG feature can significantly improve the performance of emotion recognition compared with all evaluated features. The results also show that the parallel LSTM can track the temporal dynamics of emotion from each modulation-filtered cochleagram at different scales.

In conclusion, this dissertation investigates different auditory features and some deep learning models for categorical or dimensional emotion recognition according to different features. This study proposes 3D CNN architecture to learn joint spectral-temporal auditory representation from the temporal modulation cues and ASRNN model to capture the salient regions of emotion continuously. Experiment results proved that the proposed methods could effectively extract distinguishable spectral-temporal representations and capture the salient regions from the representation sequence. In addition, this study also proposes the MMCG feature to capture the temporal and contextual modulation cues in different resolutions, and develops a parallel LSTM to capture the temporal dynamics of the MMCG features for dimensional emotion recognition. Experiment results further prove that the proposed methods could effectively capture the temporal dynamics of emotion. The results show that the proposed deep learning models based on human auditory characteristics have achieved good performance in speech emotion recognition.

Keyword: speech emotion recognition, human auditory characteristics, multi-resolution modulation-filtered cochleagram, 3D convolution, attention-based sliding recurrent neural network

Acknowledgments

I wrote my dissertation with a big thankful heart. During the last four years at the Japan Advanced Institute of Science and Technology (JAIST), I would like to express my warmest gratitude to all those who helped me in the completion of this dissertation.

First and foremost, I would like to express my deepest appreciation to my supervisor, Prof. Masato Akagi, for his patient guidance, encouragement, and extreme support in these years. This dissertation could not be accomplished without his precious suggestions and well-directed advice. I am grateful for his encouragement and support, which accelerate my academic research in the field of speech emotion.

Secondly, I would like to express my deepest appreciation to my joint supervisor at Tianjin University, Prof. Jianwu Dang, who gave me careful guidance in scientific research and generous support in my life.

I would like to acknowledge my vice supervisor, Prof. Masashi Unoki, for his invaluable suggestions and comments in my research. The discussions with Prof. Unoki make my research progress continuously.

I thank Prof. Qinghua Hu of Tianjin University and Prof. Tatsuo Kaneko of JAIST, who help me finish the minor research project. I also thank Prof. Longbiao Wang of Tianjin University and Prof. Shogo Okada of JAIST, who gave me many precious comments for my dissertation.

I also owe my sincere gratitude to my friends and my lab mates, Dr. Zhi Zhu, Dr. Yongwei Li, Dr. Feng Li, Dr. Xingfeng Li, who gave me their help and encouragements in not only research but also everyday life.

Last but not least, I would like to thank my parents and other family members for their support, understanding, and encouragement. Most importantly, I am profoundly indebted to my wife, Hua Zeng, for her unconditional support and encouragement.

Zhichao Peng
July 07, 2020

Table of Contents

List of Figures

List of Tables

Acronym and Abbreviation

ACNN	Attentive convolutional neural network
AI	Artificial intelligence
AIOT	Artificial intelligence of things
ARNN	Attention-based recurrent neural network
ASRNN	Attention-based sliding recurrent neural network
AVEC	Audio/Vision Emotion Challenge
BPTT	Backpropagation-through-time
CCC	Concordance correlation coefficient
CN	Cochlear nucleus
CNN	Convolutional neural network
CRNN	Convolutional and recurrent neural network
DCT	Discrete cosine transform
DFT	Discrete Fourier transform
ELM	Extreme learning machine
EQ	Emotional quotient
ERB	Equivalent rectangular bandwidth
F0	Fundamental frequency
GMM	Gaussian mixture model
HMM	Hidden Markov model
HRI	Human-robot interaction
HSF	High-level statistics function
IC	Inferior colliculus
IHC	Inner hair cells
IQ	Intelligence quotient
LLD	Low-level descriptors
LSTM	Long short-term memory
MAE	Mean absolute error
MFB	Modulation filterbank
MFCC	Mel frequency cepstral coefficient
MMCG	Multi-resolution modulation-filtered cochleagram
MPCRNN	Multichannel parallel convolutional recurrent neural network
MRCG	Multi-resolution cochleagram
MSE	Mean square error

MSF	Modulation spectral feature
MSR	Modulation spectral representation
PCA	Principal component analysis
PCC	Pearson's correlation coefficient
PLP	Perceptual linear predictive
ReLU	Rectified linear unit
RNN	Recurrent neural network
SMO	Sequential minimal optimization
SNR	Signal-to-noise ratio
STFT	Short time Fourier transform
SVM	Support vector machine
SVR	Support vector regression
UA	Unweighted accuracy
V-A	Valence and arousal
VAD	Voice activity detection
WA	Weighted accuracy

Chapter 1
Introduction

1.1 Motivation

In 1985, one of the founding fathers of artificial intelligence (AI), Marvin Minsky, thought that machines should be given the ability to identify, understand, and express human emotion. He said that the question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions [1]. Since then, the researchers in the AI field have been interested in the exploration that endows an intelligent machine or robot with emotional ability. The word robot in this dissertation refers to both physical robots and virtual software agents. A real intelligent robot should have not only intelligence quotient (IQ) but also emotional quotient (EQ). This kind of robot needs to have the ability of perception, recognition, understanding, and expression of emotion, so as to realize human-robot interaction (HRI). In the era of artificial intelligence of things (AIOT), there are some new changes in HRI: first, the interaction scenario has gradually changed from offline intelligent service to online intelligent service, and a large number of service robots have been launched to meet different needs, resulting in the explosive growth of online interaction; second, the interaction mode has gradually shifted to simple, convenient and natural interaction mode.

Speech is the most convenient and natural way for communication between humans and robots. The key point of effective communication is to make robots understand speakers' true intentions. Speech contains not only linguistic information but also para-linguistic and non-linguistic information. Only using linguistic information is by no means sufficient enough for an understanding of intentions. The vocal emotion information as a kind of non-linguistic information can significantly help robots to understand speakers' true intentions. Therefore, speech emotion recognition (SER) plays an important role in robots understanding a speaker's intentions. It has a wide range of application prospects, including e-learning environments [2], intelligent game [3], humanoid service robots [4], car accidents [5], lie detection [6], robot-assisted therapy [7], empathetic social chatbot [8,9] and so on.

In SER, one of the core research issues is how to extract the emotion-salient and noise-robust features from speech signals. Most of the current studies mainly focus on traditional hand-tuned acoustic features such as prosodic features, voice quality features, and spectral features to find the salient features relevant to emotional speech [10]. Nevertheless, finding the distinguished spectral and prosodic feature set for SER is still a challenge due to the cultural differences, various expression types, context, ambient noise, etc. Moreover, when the traditional acoustic features are used for emotion recognition, the recognition performance decreases rapidly with the decrease of signal-to-noise ratio (SNR). Therefore, considering how to extract noise-robust features is an important part of SER systems. As the human auditory system is powerful in processing time-frequency signals and extracting noise-robust features, research has focused on auditory-based speech signal processing for emotion recognition by mimicking the stages of the human auditory system.

In the auditory system, sound signals are firstly analyzed by the cochlea and then are transmitted to the auditory cortex for perceiving the emotional states via the auditory pathway. The cochlea, which is the main part of the peripheral auditory system, decomposes sound signals into multi-channel acoustic frequency components along the length of the basilar membrane. Inner hair cells (IHC) detect the motion of the basilar membrane and transduce it into neural signals. Temporal amplitude envelope information is obtained from each transduced signal and travels further to the inferior colliculus (IC) at the midbrain through the auditory nerve and cochlear nucleus. Physiological studies have revealed that the processing of temporal amplitude modulation is performed in the IC for high-resolution temporal information by tuning to certain modulation frequencies [11]. Møller first observed that the mammalian auditory system has a specialized sensitivity to amplitude modulation of narrowband acoustic signals [12]. Suga showed that amplitude modulation information is maintained for different acoustic frequency channels [13]. Additionally, Chi et al. have extended the findings above to include combined spectral and temporal modulations [14]. Finally, the primary auditory cortex is responsible for the perception of sound from temporal modulation cues using the spectral-temporal receptive field of the neuron [15].

Different computational auditory models are proposed to mimic the different stages of signal processing in the auditory system. The auditory filterbank is used to simulate the time-frequency signal decomposition of the cochlear basilar membrane. The temporal-envelope extraction from the acoustic-frequency components is used to effectively simulate the mechanic-to-neural signal transduction in the IHC. The modulation filterbank (MFB) is introduced to generate high-resolution temporal-modulation cues provided by the temporal envelope and its modulation-frequency components. Recent psychoacoustic experiments showed that temporal modulation is important for speech perception and understanding [11,16–18]. Dau et al. [11] proposed an auditory perception model to simulate signal processing of the peripheral auditory system. In the perceptual model, temporal modulation cues are obtained using auditory filtering of the speech signal and modulation filtering of temporal amplitude envelope in a cascade manner. These cues contain rich spectral-temporal information to perceive variations of loudness, timbre and pitch of speech [17], which has been widely used in sound texture perception [19], speaker individuality perception [20], speech recognition [21,22], acoustic event recognition [23], and emotion recognition [24–26].

Some studies extracted modulation spectral features (MSFs) from temporal modulation cues by calculating the spectral skewness, kurtosis, and other statistical features. Those studies showed that the MSFs contribute to the perception of vocal emotion[24,25]. Wu et al. [25] showed that the MSFs perform better than the traditional acoustic features such as the Mel frequency cepstral coefficient (MFCC) and perceptual linear predictive (PLP) coefficient for SER. Zhu et al. [20] further confirmed that the MSFs contribute to the perception of vocal emotion. To reduce the computation of spectral features, however, MSFs are only calculated in each modulation channel and produce time-averaged spectral features, whereas these features lose the temporal dynamic information of speech signals. In fact, the speech signal is processed by auditory filtering to generate two-dimensional (2D) spectral-temporal representations and then processed by modulation filtering to generate three-dimensional (3D) spectral-temporal representations. These time-domain signal processing models produce more rich data than the original one-dimensional (1D) speech signal. Therefore, these auditory or modulation filtered time-domain signals are more suitable for high-level feature extraction and emotion recognition using machine learning methods, especially using deep learning methods.

Conventional approaches for SER usually extract low-level descriptors (LLDs) from speech and then recognize human emotional states using the machine learning methods such as hidden Markov model (HMM), Gaussian mixture model (GMM) [27], support vector machine (SVM) [28], and artificial neural network (ANN). However, it is still challenging to find the salient feature set from LLDs to recognize distinct emotions because of the aforementioned challenging factors. As deep learning has become the best way to find the distinguished feature, many studies focus on SER using deep neural network (DNN) from acoustic features. Convolutional neural network (CNN) [29] and recurrent neural network (RNN) [30], which are the two important DNN models, are widely used to recognize the emotion in speech. CNN can extract high-level local feature representations using the receptive field of the neuron, and have been used for acoustic modeling and feature extraction in SER systems. RNN, including long short-term memory (LSTM) [31], is designed to handle long-range temporal dependencies in the speech signal sequence. Convolutional and recurrent neural network (CRNN) is a mixed architecture formed by combining the feature learning ability of CNN with the sequence modeling ability of RNN.

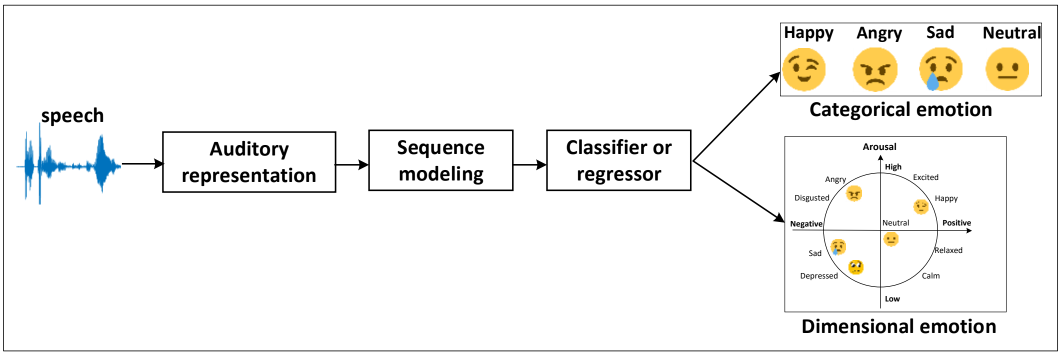


Figure 1.1: The general framework of speech emotion recognition

The goal of this study is to explore auditory representations of computational auditory models and deep learning methods to improve the performance of emotion recognition. There are mainly two kinds of methods for emotion recognition based on the different emotion description, categorical and dimensional emotion recognition. Categorical emotion describes an emotional state as discrete labels such as “happy,” “angry,” etc. Compared with categorical emotions, dimensional emotion can describe more mixed emotions and captures the gradual emotion transitions in spontaneous or natural speech [32]. There are advantages and disadvantages in categorical and dimensional emotion descriptions. To investigate the effectiveness of human auditory characteristics on emotion recognition, this study investigates both description methods. Figure 1.1 illustrates the general framework of speech emotion recognition. Auditory representations of the speech signal are first extracted from the auditory perceptual model. Then, sequence modeling of auditory representation is used to effectively capture temporal information. Finally, classification or regression models are used to identify categorical or dimensional emotions.

1.2 Challenges

The human auditory system can easily perceive the emotional states of speech even in a noisy environment, so this study is to verify whether it is possible to improve the performance of emotion recognition based on the auditory-based features using deep learning methods. However, due to the complexity of the human auditory system, the mechanism of auditory signal processing is not completely clear. We still do not know what kind of auditory representations of speech should be more distinguishable for different emotions. Moreover, we also do not know how to model the auditory representation sequence to simulate the auditory system for emotion recognition effectively. While emotion recognition based on the auditory representation of speech is an area that has increased its presence in the speech community, there are still important challenges that need to be addressed to achieve natural communication between humans and robots.

(1) Extraction for the distinguishable auditory representation of speech

In the computational auditory model, the speech signal is processed by auditory filtering and modulation filtering to generate temporal modulation cues. These cues contain four-dimensional spectral-temporal representations, including acoustic frequency, modulation frequency, amplitude, and temporal information. Little, however, is known about exactly what distinguishable auditory representations of speech are most important to identify the emotional states. The current methods extract MSF from temporal modulation cues by calculating the static features of each modulation channel. Although MSF can achieve better recognition performance than acoustic features, this static feature can not reflect the real emotional state in speech due to the loss of temporal cues. Previous research found that the auditory system responds to joint spectral-temporal patterns in the speech signal rather than temporal-only or spectral-only patterns [33]. Therefore, how to extract the distinguishable auditory representations from temporal modulation cues is a challenging task.

(2) Auditory representation sequence modeling for categorical emotion recognition

In the extraction of auditory representation, we propose 3D convolution to learn joint spectral-temporal representations from temporal modulation cues, and use LSTM to capture the temporal dependence of speech sequence. The speech sequence is segmented into non-overlapping subsequences. This feature cannot reflect the change of emotion, and it does not consider how to extract salient emotion regions. Generally, the regions with salient emotions in an utterance are very short, and most of the rest may be non-emotional or silent. Some studies addressed the silence regions using voice activity detection (VAD) [34] or by null label alignment [35], and then use an attention model to capture salient emotion regions. However, the existing attention models only focus on the salient regions of emotion but ignore the continuity of cognition. In the auditory system, selective auditory attention captures salient emotion regions by continuous scanning and encoding of the speech signals [36]. Therefore, how to effectively model the auditory representation sequence by mimicking the auditory attention to capture salient emotion regions is also an important issue of SER.

(3) Auditory representation extraction and sequence modeling for dimension emotion

recognition

In the categorical emotion recognition method, the speech signal is mapped to the high-dimensional data space through auditory and modulation filtering, and then the joint spectral-temporal feature is extracted from this representation by 3D convolution. High-dimensional data increases the complexity of the emotion recognition model, especially for the lack of large-scale speech emotion database, which may make the training model poor generalization. In dimension emotion recognition, the speech signal is usually a long time series, and the annotated dimensional value is very short. Therefore, the 3D convolution used in categorical emotion recognition is not suitable for dimension emotion recognition. In addition, emotional information in speech usually changes dynamically with time. The dynamic information of emotion in speech sequence is very important for emotion recognition, especially for dimensional emotion recognition, because the target dimensional values are continuous and have a short-time gap between two adjacent predictions [37]. However, the acoustic features, especially for the suprasegmental features, are not good at capturing the temporal dynamic for dimensional emotion recognition. The human auditory system can easily track the temporal dynamics of emotion by perceiving the intensity and fundamental frequency of speech. Therefore, how to effectively extract auditory representations and model auditory representation sequence to track the temporal dynamics of emotion for dimensional emotion recognition is also an important issue of SER.

1.3 Proposed approach

The dissertation presents novel SER methods to address the challenges mentioned in Section 1.2. This section summarizes the proposed solutions.

(1) A three-dimensional convolutional neural network architecture is proposed to obtain discriminative spectral-temporal auditory representations from the temporal modulation cues

This study attempts to extract auditory representation from human auditory models to improve the recognition performance of emotion. However, due to the complexity of the human auditory system, the mechanism of auditory signal processing is not completely clear. We still do not know which auditory model can better simulate the human auditory system. Therefore, this study first investigates the cochlear auditory filterbank, then introduce modulation filterbank to generate temporal modulation cues. Multi-dimensional spectral-temporal auditory representations can be obtained from temporal modulation cues, which contain acoustic frequency components, modulation frequency components, and temporal information. This study then proposes a 3D convolutional neural network architecture to extract the discriminative spectral-temporal auditory representations from the temporal modulation cues.

(2) An attention-based sliding recurrent neural network is proposed to continuously obtain segment-level features and capture the salient regions of emotion representation

As selective auditory attention in the auditory system can capture salient emotion regions by continuous scanning and encoding of the speech signals. We investigate the relation of the auditory features and human attention mechanism and propose an attention-based sliding recurrent neural network (ASRNN) model to seize the salient emotion regions from joint spectral-temporal representation. Among them, a sliding window is used to continuously obtain segment-level features, so that the features between segments are partially overlapped, and each segment contains context-related features. Then, a temporal attention model is used to capture the salient regions of emotion representation in each utterance. In this method, ASRNN effectively models auditory representation sequence by mimicking the auditory attention to capture salient emotion regions for categorical emotion recognition.

(3) A multi-resolution modulation-filtered cochleagram feature is proposed to capture the temporal and contextual modulation cues for dimensional emotion recognition

Since emotion in speech often changes with the time, the temporal dynamics are very important factors in emotion recognition. Temporal modulation cues obtained directly from the time-domain model of auditory perception can reflect its temporal dynamics compare to acoustic features usually processed in the frequency domain. A recent neuroscientific study suggests that the cortex derives multi-resolution representations through the temporal modulation analysis. Therefore, this study proposes a novel auditory feature to extract high-level auditory representation from temporal modulation cues, and designs a parallel LSTM network architecture to track the temporal dynamics of auditory representation sequence. The proposed novel feature, multi-resolution modulation-filtered cochleagram (MMCG), is constructed by combining four modulation-filtered cochleagrams at different resolutions to capture temporal and contextual information. Each kind of modulation-filtered cochleagrams is extracted from temporal modulation cues of the amplitude envelope. Considering that each modulation-filtered cochleagram in MMCG contains different temporal and contextual modulation cues, a parallel LSTM is designed to model multi-temporal dependencies of MMCG and track the temporal dynamics of auditory representation sequence.

1.4 Contributions

This study explores different methods of feature extraction based on human auditory characteristics and combines the current popular deep learning methods to identify categorical and dimensional emotional representation. This dissertation presents the following contributions to the area of SER.

Temporal modulation cues play an important role in speech perception and contain multi-dimensional spectral-temporal information. Therefore, this study proposes a 3D CNN architecture to extract the discriminative auditory representations from temporal modulation cues for categorical emotion recognition. This deep model obtains both the local features and periodicity information of emotional speech by a joint spectral-temporal feature learning. It is confirmed that temporal modulation cues contain joint spectral-temporal representations and high-level discriminative auditory representation could be extracted from temporal modulation cues by joint spectral-temporal feature learning of 3D convolution.

Capturing salient emotion regions using the attention model is a perspective way for emotion recognition. Some recent studies proposed attention models to adjust weights of LLD-based features. Unlike these studies, this study proposes an ASRNN to continuously scan the temporal sequence and focus on the emotional region. In ASRNN, the continuous segment-level internal representations are extracted by a sliding window, and then a temporal attention model is focused on the salient emotion regions for utterance-level emotional states. Moreover, the results of the listening tests indicate that there is a strong correlation between human auditory attention and the attention model. Therefore, the ASRNN architecture can effectively capture the salient emotional regions, which is similar to the human selective auditory attention.

Inspired by the multi-resolution modulation signal processing of the auditory system, the MMCG feature is further proposed to capture the temporal and contextual modulation cues. This feature is constructed by combining four modulation-filtered cochleagrams at different resolutions to capture various spectral and temporal features. It is confirmed that the MMCG feature could effectively capture the temporal and contextual cues at different resolutions for dimensional emotion recognition. The results also show that the parallel LSTM can track the temporal dynamics of emotion from each modulation-filtered cochleagram at different scales.

1.5 Dissertation organization

The rest of this dissertation is organized as follows. Figure1.2 shows the organization of this dissertation.

Chapter 2 reviews the related knowledge of SER, including the representation of emotion, emotion database used in this study, traditional acoustic feature, auditory feature, and deep learning methods. The typical evaluation metrics of classification and regression models are also introduced.

Chapter 3 analyzes the existing problems of SER features and extracts discriminative auditory features for categorical emotion recognition. First, a multi-channel parallel CRNN is proposed for two-stage emotion recognition based on Gammatone auditory filterbank. Then, a 3D CRNN is proposed for end-to-end emotion recognition based on temporal modulation cues. Finally, the experimental results of the two methods are analyzed and compared.

Chapter 4 proposes an ASRNN to focus on the salient emotion regions by extracting segment-level features in a sliding window manner and utterance-level features with a temporal attention model. Then, a subjective evaluation is conducted to investigate the correlation between the temporal attention model and human auditory attention in perceiving emotional speech.

Chapter 5 proposes a novel MMCG feature to capture the temporal and contextual modulation cues and designs a parallel LSTM network architecture to extract more temporal dynamics from modulation-filtered cochleagram for dimension emotion recognition. Finally, the performance of dimension emotion recognition is analyzed and compared with all evaluated features.

Chapter 6 concludes the proposed methods for speech emotion recognition, and discusses the future works in the end.

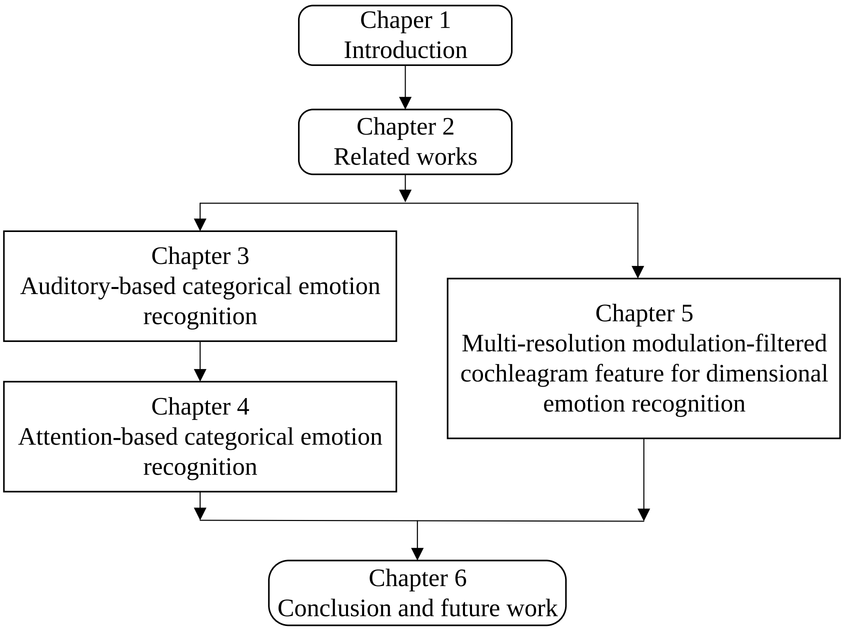

Chaper 1
Introduction
Chapter 2
Related works
Chapter 3
Auditory-based categorical emotion recognition
Chapter 4
Attention-based categorical emotion recognition
Chapter 5
Multi-resolution modulation-filtered cochleagram feature for dimensional emotion recognition
Chapter 6
Conclusion and future work

Figure 1.2: Organization of this dissertation

Chapter 2
Related works

2.1 Representation of emotion

Speech emotion recognition (SER) aims to identify the emotional states of human beings from speech automatically. The process of SER is to extract emotion features from the speech signal, then train an emotion recognition model and identify the emotional states using a specific emotion description method. Emotion plays an essential role in the understanding of the speaker's intention. However, compared with other psychological phenomena such as cognitive, there is no universally agreed theoretical definition because of the subjectivity of emotion recognition.

At present, there are many description models about emotion definition, which are mainly divided into two kinds of emotion description: categorical emotion description and dimensional emotion description. During the 1970s, psychologist Paul Ekman identified six basic or prototypical emotions that he suggested were universally experienced in all human cultures. The emotions he identified were happiness, anger, sadness, surprise, fear, and disgust [38]. In addition to identifying categorical emotion types, people use dimensional emotion to describe more abundant emotion types. Dimension emotional description uses the continuous numerical value to describe an emotional state, so it is also called continuous emotional description. It regards the emotional state as a point in multi-dimensional emotional space, and each dimension corresponds to different psychological attributes of emotion, such as arousal, valence, expectation, dominance, liking, etc. The dimensional emotion is more close to the analysis of continuous and complex emotion information naturally expressed by human beings in daily communication activities. Valence and arousal (V-A) are the universal primitives in emotion dimensional space [39], as shown in Fig. 2.1. Valence is related to subjective appraisal and experience with positive or negative emotions. Arousal is associated with an intensity level, unusually low or high degree of arousal. Through different valence and arousal, we can distinguish different emotions. For example, neutral remains in the middle of the V-A space. Happy and angry both have a high activation level. However, happy has a positive valence value, and angry has a negative valence value. Sad has a negative valence and low arousal.

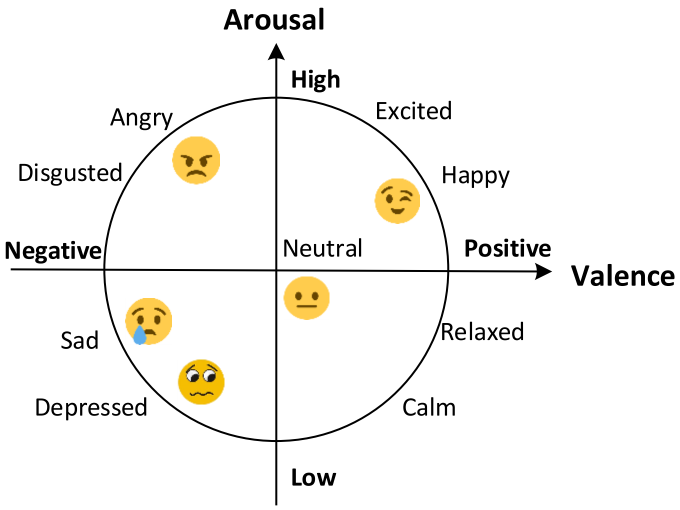


Figure 2.1: The 2D valence-arousal emotion space, with the approximate positions of some categorical descriptors shown in the plane

However, both categorical and dimensional emotion description methods have their limitations. Because emotions are complex and subjective, fewer discrete categories may not reflect the subtle differences and complexity of emotional states. The dimensional emotion space can reflect more subtle and fuzzy emotions without boundary, and it does not need to define a large number of emotion states classification in advance. However, the dimensional evaluation value may lose significance due to the lack of consistent evaluation criteria. In order to investigate the effectiveness of human auditory characteristics on emotion recognition, this study utilizes both categorical and dimensional emotion descriptions.

2.2 Emotional speech corpus

This dissertation utilizes both categorical and dimensional emotion corpus to train, develop, and test the proposed SER frameworks.

2.2.1 Categorical emotion corpus

In this dissertation, three categorical emotion databases are used in total, including CISIA database [40], the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [41], and the MSP-IMPROV database [42].

1) CISIA database: A acted Mandarin emotional speech database made by the Chinese Academy of Sciences. CISIA database comprises a total of 9600 recordings from four actors (2 females and 2 males). This database was recorded in laboratory environments with fixed lexical content and acted emotions. Recordings for every speaker were made

300 same sentences and 100 different sentences. Each speaker utters 400 sentences with six emotions, which are happy, fear, angry, sad, surprise, and neutral emotion. CISIA database is used for two-stage categorical emotion recognition.

2) IEMOCAP database: This database is a well-known dataset for speech emotion recognition comprising of scripted and improvised multimodal interactions of dyadic sessions. It consists of around 12 hours of speech from 10 human subjects and is labeled by three annotators for emotions such as happy, sad, angry, excited, and neutral, along with dimensional labels such as valence and arousal. All recordings have the structure of a dialogue between a man and a woman either scripted or improvised on the given topic. This study includes excitement utterances with happiness ones and takes 5,531 utterances (1636 happy, 1084 sad, 1103 angry, 1708 neutral) for all sessions. The mean length of all the turns is 4.55 s (max.: 34.14 s, min.: 0.58 s), emotional state distribution is shown in Fig 2.2. IEMOCAP database is used for end-to-end categorical emotion recognition.

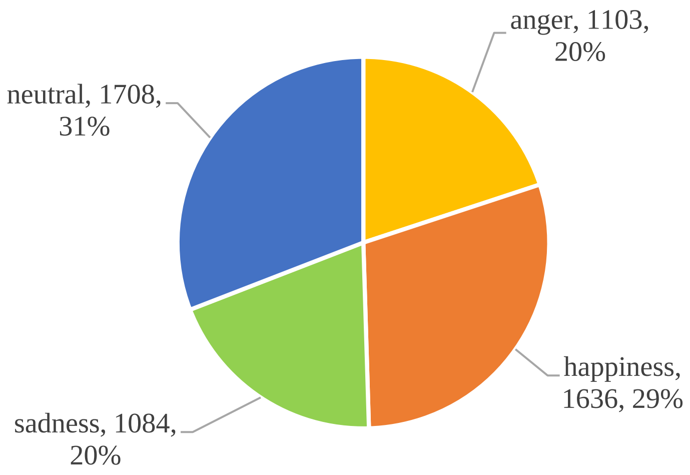


Figure 2.2: Emotional state distribution in IEMOCAP

3) MSP-IMPROV database: This database is an audio-visual dyadic emotion corpus. It consists of six sessions in the same manner (12 unique speakers). Each session includes all the speaking turns of the improvisation and the natural interaction based on the 20 target sentences in the improvised scene. The emotional expressions of the speakers were elicited through carefully designed scenarios that include improvisations and target sentences with specific lexical content. The final database contains a total of 7798 utterances (2644 happy, 885 sad, 792 angry, 3477 neutral). The mean length of all the turns is 4.09 s (max.: 31.09 s, min.: 0.41 s), emotional state distribution is shown in Fig 2.3.

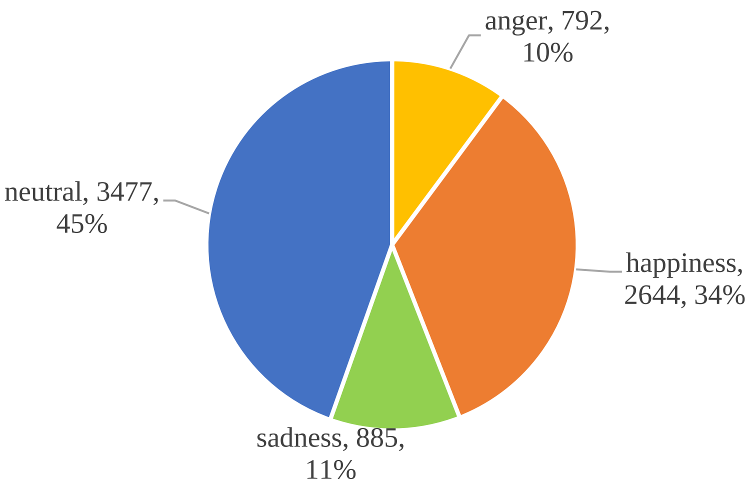


Figure 2.3: Emotional state distribution in MSP-IMPROV

In this study, the IEMOCAP and MSP-IMPROV databases are used in the experiment of attention-based categorical emotion recognition. Both databases are composed of multimodal interactions of dyadic sessions and labeled by three annotators for emotions such as happy, sad, angry, excited, and neutral, along with dimensional labels such as valence and arousal. Only four emotional categories are used in both databases: happy, sad, angry, and neutral.

2.2.2 Dimensional emotion corpus

Two databases are used in this study, namely, RECOLA [43] and SEWA [44] databases. These two databases consist of spontaneous data, and a selected subset of these two databases are used as per Audio/Vision Emotion Challenge (AVEC) 2016 [45] and AVEC 2017 [46].

1) RECOLA database: This database is a multi-modal corpus of remote collaborative and affective interaction. There are 27 French speakers in the database, which are divided into three partitions (9 train, 9 development, and 9 test) by balancing gender, age and mother tongue of the subjects. The recordings are annotated time-continuously in terms of the emotional dimensions including arousal, valence and dominance. The affective behavior of the participants was evaluated by six different annotators and averaged over all annotators by considering the inter-annotator agreement to provide a gold standard.

2) SEWA database: The database is also a multimodal database for remote collaboration and emotional interaction. This subset of the database is used in 2017, 2019, and 2020 AVEC challenges on emotion recognition. The SEWA database recruited 408 speakers and divided them into six groups according to different cultural backgrounds

(UK, Germany, Hungary, Greece, Serbia and China). There are also significant differences in age and gender in each group, with at least three pairs in each age group being able to speak in their mother tongue. This study uses the 2017 AVEC database. This database contains 64 German subjects and is divided into three partitions (34 train, 14 development, and 16 test). The recordings are annotated time-continuously in terms of the emotional dimensions, including arousal, valence and liking. To validate efficacy of the proposed approach for dimensional emotion recognition, a subset of SEWA database is also used.

In this study, we use gold-standard labels and investigate arousal and valence prediction for both databases. Similar to the studies [26,47], we use the same training set and development set to train and validate the recurrent model with different acoustic-based features and auditory-based features. Specifically, 18 recordings in the RECOLA database and 48 recordings in the SEWA database are adopted.

Although both databases were obtained from dyadic conversations, differences between RECOLA and SEWA are as follows:

1) Each recording includes only the audio from the target speaker in RECOLA, whereas each recording includes the mix of the target speaker and interlocutor in SEWA.

2) The duration of each recording is lasting for 5 minutes in RECOLA, while it is variable from 47 seconds to 3 minutes in SEWA.

3) The sampling rate of the emotion annotation is 25Hz in RECOLA while it is 10Hz in SEWA. It means that the values of each primitive emotion are continuously labeled on 40-ms consecutive frames in RECOLA and 100-ms consecutive frames in SEWA.

2.3 Acoustic feature extraction

In deep learning-based speech emotion recognition, acoustic feature extraction usually consists of two steps. Firstly, LLDs are extracted from speech, and then the statistical function of LLDs calculated on a block of continuous frames are used to get High-level statistics functions (HSFs). In this study, acoustic features are used as baseline features for performance comparison with auditory-based features.

2.3.1 Low-level descriptors

LLDs refers to some traditional hand-tuned acoustic features, which are generally calculated on a frame of speech, and are used to represent the features of a frame of speech. The hand-tuned acoustic features most widely used are prosodic features, voice quality features and spectral features.

Voice quality feature: refer specifically to the properties of speech affected by the stuff inside your larynx. These features usually include formant frequency and bandwidth, jitter, shimmer and glottal parameter.

Prosody feature: refers to the change of pitch intensity and speaking rate in speech besides the voice quality feature, also known as a suprasegmental feature, including duration, fundamental frequency (F0), energy and other features. The F0 refers to the vibration frequency of the pitch, which determines the pitch of the voice and can reflect the emotional state to a large extent. For example, when the pitch is high, it may correspond to the emotional state of happiness or anger; when the pitch is low, it may correspond to the emotional state of sadness. Pitch is often used to express the perception of F0 in subjective psychology. The reciprocal of its F0 is called a fundamental period, which is also a common acoustic feature.

Spectral feature: signal properties in the frequency domain, thus providing useful additions to voice quality and prosody features, including linear predictor coefficient (LPC) and Mel-frequency cepstrum coefficient (MFCC). MFCC is the most commonly used spectral feature. It is a feature based on the Mel scale, which is closer to the response of the auditory system than the linear interval frequency band. MFCC extraction method is: firstly, the speech signal is blocked into short-term overlapping frames, then each speech frame $s(t)$ is multiplied by an analysis window $w(t)$ and the short-term Fourier transform (STFT) is computed and subsequently compute the filter banks. As the filterbank coefficients calculated in the previous step are highly correlated, discrete cosine transform (DCT) is calculated to decorrelate the filter bank coefficients and yield a compressed representation of the filter banks. Finally, only the first 12 DCT coefficients are reserved.

Table 2.1 lists the ComParE acoustic feature set with 65 low-level descriptors.

2.3.2 High-level statistics functions

Although the frame-level features can be used directly for machine learning, a more common approach in SER is to compute features at segment-level or utterance-level, by applying a number of descriptive functions (typically statistical) to the contours of frame-level features, and often to their derivatives as well to extract local dynamic cues.

HSFs are obtained by calculating the statistic functions on LLDs, such as mean, maximum, etc. At present, the researches usually use local or global feature statistical functions to get segment-level or utterance-level HSFs, which can extract the same dimension features from different speech segments. The Munich open Speech and Music Interpretation by Large Space Extraction (openSMILE) as a prevailing emotion recognition toolkit is employed in this dissertation to extract LLDs and HSFs [48].

Table 2.1: The 65 low-level descriptors provided in the ComParE acoustic feature set

4 energy related LLD	**group**
Sum of auditory spectrum	prosodic
Sum of RASTA-filtered auditory spectrum	prosodic
RMS Energy, Zero-Crossing Rate	prosodic
55 spectral LLD	**group**
RASTA-filt. aud. spect. bds. 1–26 (0-8 kHz)	spectral
MFCC 1–14 cepstral	cepstral
Spectral energy 250–650 Hz, 1 k–4 kHz	spectral
Spectral Roll-Off Pt. 0.25, 0.5, 0.75, 0.9	spectral
Spectral Flux, Centroid, Entropy, Slope	spectral
Psychoacoustic Sharpness, Harmonicity	spectral
6 voicing related LLD	**group**
F0 (SHS & Viterbi smoothing)	prosodic
Prob. of voicing	voice quality
log. HNR, Jitter (local & DDP), Shimmer (local)	voice quality

The following describes the main HSFs used in this study:

IS09_emotion [49]: Emotion feature set of the INTERSPEECH 2009 Emotion Challenge. For 16 LLDs and their first-order delta features, 12 statistical functions are applied to obtain 384-dimensional features, including statistics of short-time energy, MFCC, short-time zero-crossing rate, time domain and frequency domain information.

IS10_paraling [50]: Emotion feature set of the INTERSPEECH 2010 Emotion Challenge. It includes 1582 dimensional features, 34 LLDs and their delta features use 1428 dimensional features generated by 21 functions, and 19 functions use 4 pitch based LLDs and their delta features.

IS13_ComParE [51]: Emotion feature set of the INTERSPEECH Emotion Challenge since 2013. It includes 6373-dimensional features, including 4 energy features, 55 spectrum features, 6 acoustic features and delta features. 130 LLDs are obtained.

Emobase: This feature set includes the following LLDs, intensity, loudness, 12 MFCC, pitch (F0), probability of voice, F0 envelope, 8 LSF (line spectral frequencies), zero-crossing rate, and then calculate statistical features (arithmetic mean, linear fitting, delta features, standard deviation, etc.) for these features.

Emobase2010 [50]: According to the use of documents of openSMILE, this feature set is basically the same as IS10_paraling. The only difference is that the “maxpos” and “minpos” features are standardized in the INTERSPEECH 2010 paralinguistic challenge set. This configuration is standardized as segment length.

Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [52]: The statistical functionals include mean, standard deviation, percentiles, and slope of F0 and loudness contours and only mean and standard deviation for the other LLDs (MFCC, spectral descriptors, etc.). In addition to that, some rhythm-related features are computed, namely, rate of loudness peaks, the mean length and the standard deviation of continuously voiced and unvoiced regions, and the number of continuous voiced regions per second.

HSFs features based on utterance-level statistics have the advantages of low dimension and become the mainstream feature extraction method, but its disadvantage is that after calculating statistical function, some local information reflecting the dynamic changes of emotion will be lost. With the development of deep learning and high-performance computing technology, more and more researches consider the use of segment level HSFs. LLDs and HSFs are common emotion recognition methods. In addition, some researches directly use short-time Fourier transform (STFT) bins [53], Mel filterbank [54,55] or spectrogram [56] for emotion recognition. Only MFCC imitates human’s auditory physiological features to a certain extent. It first maps the linear spectrum to Mel nonlinear spectrum based on auditory perception and then converts it to cepstrum. However, due to the use of triangular filterbanks for frequency-domain filtering, the energy leakage between adjacent frequency bands is very serious, which is not conducive

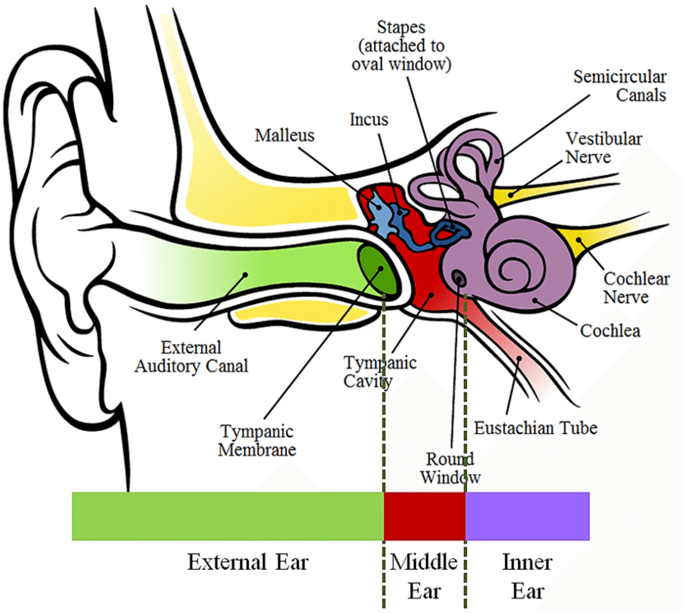


Figure 2.4: Human auditory system

to the extraction of formants and other characteristics, and its frequency band division is based on the uniform distribution of the central frequency Mel scale, which is not fully in line with the concept of the critical bandwidth in the auditory characteristics.

2.4 Auditory model

2.4.1 Human auditory system

In the auditory system, the mechanical stimulus is transformed into nerve impulses, and these impulses are transferred to the auditory cortex of the brain along the auditory nerve. The transformation enables the brain to extract spoken words and nonverbal elements such as emotions. According to the processing of speech and audio signals, the auditory system can be roughly divided into the peripheral auditory system and central auditory system. The peripheral auditory system consists of the outer ear, the middle ear and the inner ear, as shown in Fig 2.4.

The outer ear is composed of the pinna and external auditory canal. Its main function is to collect sound, amplify it and judge the direction of a sound source.

The middle ear is mainly composed of the tympanic membrane and three auditory ossicles (malleus, incus and stapes). Its primary physiological function is to amplify the gain of input speech and transfer it efficiently into the cochlea from the external auditory canal.

The inner ear is mainly composed of vestibule and cochlea. Its primary physiological function is the sound sensing function of the cochlea. The cochlea is shaped like a snail’s shell, which is composed of three ducts that run in parallel: scala tympani, scala media, scala vestibuli. The organ of Corti located in the scala media is the main component of the cochlea, which is responsible for transforming the mechanical vibration transmitted to cochlea into the nerve impulse of the auditory nerve fiber. The vibration of the basilar membrane of the cochlea stimulates the hair cells located above it, and causes the afferent nerve fibers at the bottom of the hair cells to produce action potentials, which leads to the release of chemical transmitters at the nerve endings, and the nerve impulses to the central auditory system.

The central auditory system consists of the cochlear nucleus, superior olivary complex, inferior colliculus, medial geniculate body and auditory cortex. The afferent nerve signals arrive at the auditory cortex along with the central auditory system and generate auditory perception and cognition in the auditory cortex.

2.4.2 Computational auditory modeling

In the light of the human auditory system, many computational auditory models have been developed which describe the signal processing occurs in the ears. Recently, computational auditory models are being used in feature extraction for SER tasks. Different auditory models are used to simulate different stages of auditory signal processing.

1) Auditory filterbank

The auditory filter simulates the time-frequency signal decomposition of the cochlear basilar membrane. The auditory filterbank is used to simulate the time-frequency signal decomposition of the cochlear basilar membrane. Two kinds of cochlear models are commonly used as a simulation of the cochlea in the processing of speech and audio. One is Lyon's cochlear model, and the other is the auditory filterbank model based on equivalent rectangular bandwidth (ERB_N) [57]. Auditory filterbank is approximate to simulate the frequency separation of sounds within the cochlea from the basilar membrane. Auditory filterbank model well the basilar membrane motion (BMM) of the auditory system.

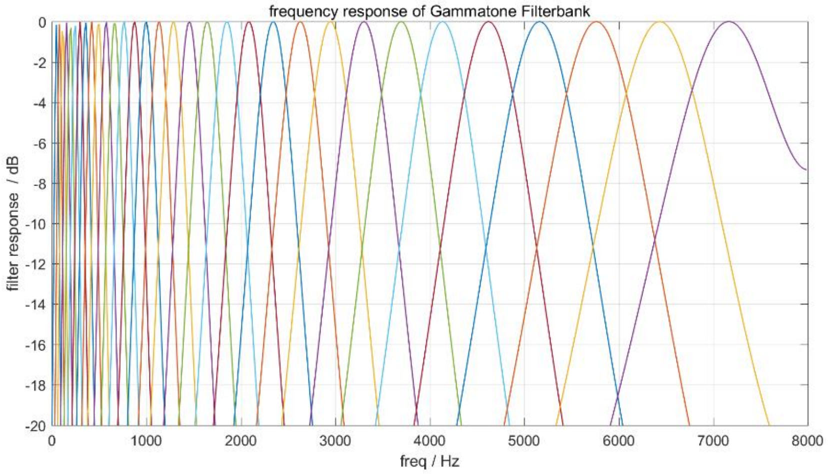


Figure 2.5: Frequency response of Gammatone filter

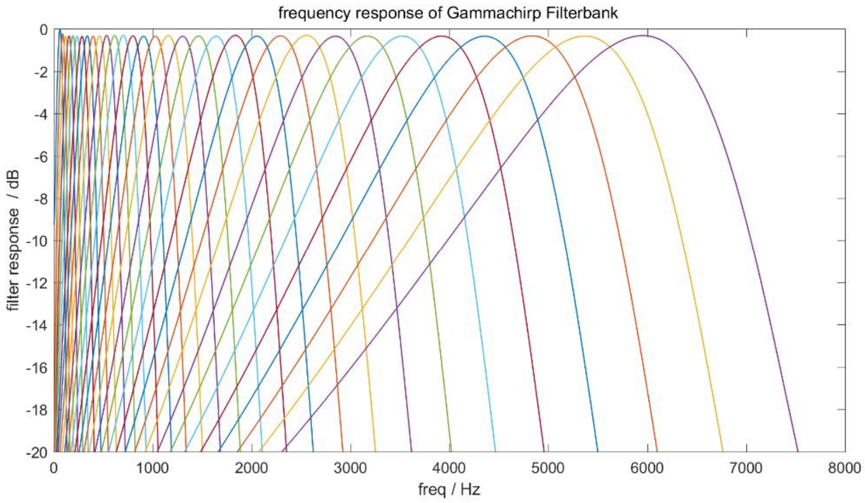


Figure 2.6: Frequency response of Gammachirp filterbank

Gammatone [58] or Gammachirp [59] filterbanks are the commonly used auditory filterbank. The impulse response of a Gammatone filter is the product of a Gamma distribution and a sinusoidal tone. The bandwidth of each filter is described by an ERB_N, which is a psychoacoustic measure of the width of the auditory filter at each point along the cochlea. Fig. 2.5 illustrates the frequency responses of the Gammatone filterbank. Compared to Gammatone filter, Gammachirp filter is an asymmetric and Nonlinear filter which is similar to auditory filter shapes. The frequency responses of the Gammatone filters, as seen in Fig. 2.6, are asymmetric and exhibit a sharp drop-off on the high frequency side of the center frequency compared to Gammatone filter. This corresponds well to auditory filter shapes derived from masking data.

Both Gammatone and Gammachirp filter is used to simulate the basilar membrane, each of which has its own advantages and disadvantages. The calculation efficiency of Gammatone is higher than that of Gammachirp, but Gammachirp is better than Gammatone in simulating the asymmetric and level-dependent auditory filterbank. Experiments show that they have little influence on the performance of emotion recognition. In this study, Gammatone filter is used in two-stage emotion recognition and dimension emotion recognition, while the Gammachirp filter is used in end-to-end mode.

2) Temporal Envelope Extraction

On basilar membrane, the frequency selectivity of position changes logarithmically. The IHCs are embedded in the floor of the basilar membrane and will be excited when the basilar membrane moves upwards. IHCs detect the movement of the basilar membrane and transduce it into neural signals. Each mechanical-to-neural signal transduction contains a temporal envelope, which is very important for speech perception.

The temporal amplitude envelope simulates the signal transduction of the IHCs. The temporal envelope from each band is usually extracted through either half-wave or full-wave rectification and low-pass filter. Recently, the Hilbert transform was used as another way to extract a temporal envelope. The half-wave or full-wave rectification produces distorted frequency components in the modulation domain, whereas the Hilbert transform provides a clear separation between the signal's temporal envelope and fine structure [60]. Hence, the Hilbert transform is used for temporal envelope extraction in this study.

3) Modulation Filterbank

There are both physiological and psychology evidence suggested the existence of modulation filterbank in the auditory system. From a physiological point of view, the processing of amplitude modulation frequencies is performed in the higher processing stages of the auditory system [31]. This temporal periodicity code is assumed to be translated into a frequency selective rate-based representation between the cochlear nucleus (CN) and the inferior colliculus (IC). Furthermore, in the IC, a periodotopic arrangement of neurons is suggested that are tuned to certain modulation frequencies. These neurons were found to be arranged almost orthogonally to the tonotopic arrangement of neurons that are tuned to certain acoustic frequencies. Physiological studies have shown that temporal modulation is the processing of high-resolution temporal information by tuning the IC to a specific modulation frequency [11]. Recent psychoacoustic experiments show that temporal modulation is very important in speech perception and understanding. A modulation filterbank is introduced to analyze the envelope fluctuations of the stimuli in each peripheral auditory filter. Temporal modulation cues of high frequency-domain resolution can be obtained by the modulation filter.

2.5 Deep learning

CNNs and RNNs are two important deep learning algorithms, which are used to recognize the emotion in this study. CNNs are used to extract high-level local feature representations and RNNs are used to handle long-range temporal dependencies in time series.

2.5.1 Convolutional neural network

The CNN is designed especially for visual recognition tasks and consists of many pairs of alternating convolutional and subsampling layers [30]. Inspired by neuroscience, Yann LeCun et al. proposed CNN architecture called LeNet-5 for the task of recognizing handwritten character recognition in the 1990s [10]. CNN architecture is mainly composed of an input layer, convolutional layer, pooling layer, fully connected layer and output layer. The convolutional layer and the pooling layer are connected in an alternating manner, that is, a convolutional layer is connected to a pooling layer, and then a convolutional layer is connected, and so on. Figure 2.7 shows the basic structure of CNN network. The CNN network introduces three core ideas of the local receptive field, weight sharing, and spatial or temporal downsampling, so that CNN can better reflect the local features of the image and maintain the invariance of the displacement, scale or deformation of the features to a certain degree.

(1) Local receptive fields

The role of the receptive field is to allow the convolutional layer to extract local features of the image and maintain the spatial continuity of the image [61]. Similar to the local perception mechanism of the cat visual cortex, the neurons in adjacent layers use a local connection mode to extract basic visual features (such as directional edges, endpoints, corners, etc.) and then combine the basic features at higher layers to form global features.

(2) Shared Weights

The weight parameters of each neuron in the hidden layer of the same feature in the local connection can be shared with other neurons, which greatly reduces the training parameters. Neurons with receptive fields located in different areas of the image have the same weight. Compared with the traditional fully connected neural network, the convolutional network greatly reduces the number of network parameters through weight sharing, thus making it feasible to train large-scale networks.

(3) Spatial or temporal sub-sampling

Downsampling is also called pooling, and its principle is the scale invariance of features. The pooling function reduces the scale of output features and increases the receptive field of subsequent convolutional layers, which helps to extract high-level features while reducing the computational complexity of the network. It uses a fixed-size pooling window and a certain stride to slide on the feature and calculates the maximum or average value of the feature in the window according to the different pooling functions.

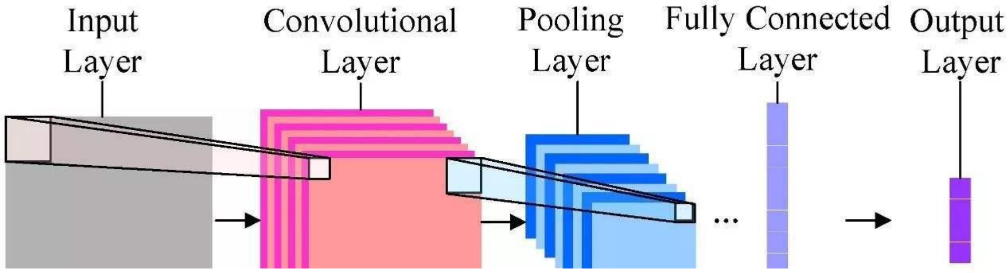


Figure 2.7: Schematic diagram of the basic structure of the CNN network

According to the calculation method of pooling function, there are Max Pooling and Average Pooling.

The function of the activation function in the convolutional neural network is to increase the nonlinearity of the neural network. The commonly used activation functions are:

(1) Sigmoid function

Also called S-shaped function, it is strictly monotonically increasing and differentiable. The function expression is as follows:

$$y = f(x) = \frac{1}{1 + e^{-x}} \tag{2.1}$$

The value range of the Sigmoid function is (0,1), and the function value changes faster where the independent variable is close to 0. The Sigmoid function is the most commonly used activation function in neural networks.

(2) Hyperbolic tangent function (tanh)

$$y = f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2.2}$$

This function can be derived from the sigmoid function. The range of the tanh function is [-1, 1], and the relationship is:

$$tanh\, x = 2sigmoid(2x) - 1 \tag{2.3}$$

(3) Rectified linear unit (ReLU)

This function is considered to have a certain biological principle, and is often better than other activation functions in practice, and is the most widely used in DNN. The function expression is as follows:

$$y = max(0, x) \tag{2.4}$$

There are three main changes in ReLU than the Sigmoid function: 1) unilateral suppression, 2) wide excitement boundary, and 3) sparse activation.

In order to prevent overfitting during training, dropout optimization methods are usually used. Secondly, the convolution operation can be performed in different dimensions. For example, 1D convolution is mostly used for feature extraction of 1D speech or other signal sequences, and 2D convolution is mostly used for computer vision and speech spectrograms. Moreover, 3D convolution is mostly used for feature extraction of 3D structure. In short, CNN has been widely used in image vision, speech recognition, signal processing and other fields, and achieved excellent results.

2.5.2 Recurrent neural network

RNN is a type of neural network where the output from the previous step are fed as input to the current step, and is powerful in modeling the sequential data [31]. Unlike CNN, the RNN introduces the concept of memory to process arbitrary sequences of inputs. Shown in Fig 2.8 is the recurrent neural network structure. On the left, it contains a self-loop connection. Among them, x is the input sequence, h is the hidden vector sequence, o is the output vector sequence. W, U and V are the weights matrices of the hidden layer input layer and output layer, respectively. On the right, it shows the structure obtained by unfolding it in time. h_t represents the memory of the sample at time t, $h_t = f(W * h_{t-1} + U * x_{t-1})$. h_{t-1} and h_{t+1} represent the memories of $t-1$ and $t+1$ time steps, respectively.

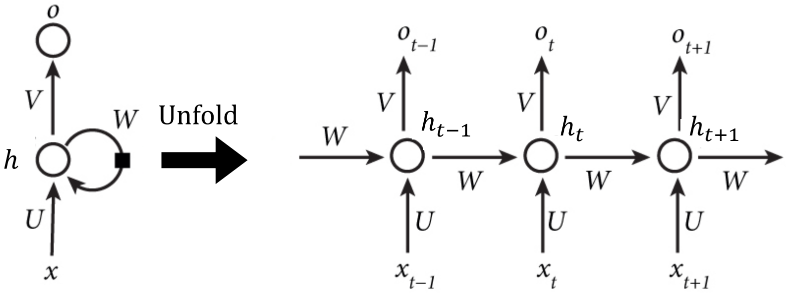


Figure 2.8: Recurrent neural network structure

During the reverse propagation, the RNN will encounter the problem of gradient disappearance. The most popular way to train an RNN is by backpropagation through time. However, the problem of the vanishing gradients often causes the parameters to capture short-term dependencies while the information from earlier time steps decays, and the RNN becomes worse at modeling long-term dependencies. The emotional context is very important in SER, so we don’t want to lose that information. LSTM is a kind of recurrent neural network specially designed to solve the long-term dependence problem of general RNN [30].

2.5.3 Long short-term memory network

LSTM architecture is the state-of-art model for sequence analysis since it can exploit long-term dependencies in the sequences by using memory cells to store information. Given an input feature sequence $\mathrm{x} = \{x_1, \ldots, x_T\}$, LSTM computes the hidden vector sequence $\mathrm{h} = \{h_1, \ldots, h_T\}$, and output vector sequence $\mathrm{y} = \{y_1, \ldots, y_T\}$ by iterating the following equations from t=1 to T:

$$(h_t, c_t) = \mathrm{H}(x_t, h_{t-1}, c_{t-1}), \tag{2.5}$$

$$y_t = w_y * h_t + b_y, \tag{2.6}$$

Where the H term is the LSTM layer function, c is the cell activation vector with the same size as the hidden vector h. The w terms denote weight matrices and the b terms denote the bias vectors.

Figure 2.9 shows the gate structure in the LSTM. It includes forget gate, input gate and output gate. These three gate structures effectively avoid the phenomenon of gradient disappearance. Among them, C represents the cell state, σ represents the Sigmoid function. A horizontal line at the top of the figure is the memory flow, representing the memory information of the current time step as the input of the next time step. The horizontal line runs through the top of this figure. A horizontal line at the bottom of the figure is the data flow, representing the input data at the current time step and the hidden layer data at the previous time step.

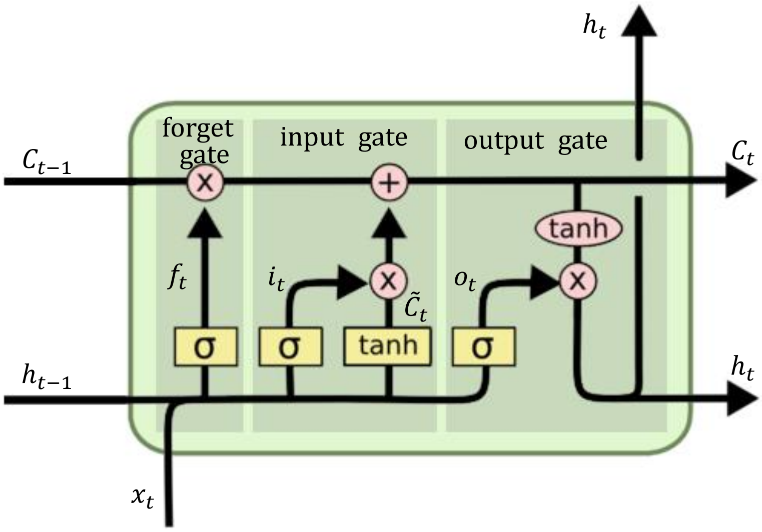


Figure 2.9: The forget gate, input gate and output gate in Long Short-Term Memory Network

BLSTM increases the flow of information in reverse time on the basis of LSTM, so it can use “future” information, which is usually better than the one-way model in effect. The basic idea of BLSTM is to combine two LSTMs in opposite directions, and these two hidden layers are connected to an output layer.

2.6 Evaluation metrics

2.6.1 Classification model evaluation metrics

The classification method based on deep learning usually adopts a softmax function label to be converted into one-hot vector. The softmax function is usually added to the last layer of various neural networks. The category corresponding to its maximum output value is used as the identification classification of the sample. The softmax function normalizes the k-dimensional vector in the node to another k-dimensional vector, so that the range of each element in the vector is $[0,1]$, and the sum of all elements is equal to 1. The definition of the Softmax function is:

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \tag{2.7}$$

Where z_i represents the ith element in the original vector.

Accuracy is the most common and basic evaluation standard in classification models. For the average recognition accuracy of all categories, weighted accuracy (WA) and unweighted accuracy (UA) can be used to evaluate separately. WA is the ratio between the accuracy prediction of each category and the total number of each category, and finally, get the average value of the accuracy of each category. UA is the ratio between the number of correctly predicted samples in all categories of the test set and the total number of predicted samples input to the classifier. UA is an ideal evaluation index for data with an uneven distribution of sample categories.

2.6.2 Regression model evaluation metrics

At present, there are many evaluation methods for regression models, including Mean Absolute Error (MAE), Mean Square Error (MSE), Pearson's Correlation Coefficient (PCC) and Consistency Correlation (concordance correlation coefficient, CCC), etc. The CCC is used as a differentiable objective function that unites both PCC and MSE and can be thought of as a PCC that enforces the correct scale and offset of the outputs [62]. Hence, the CCC between the prediction values of emotion dimensions and the gold standard values is used to determining the weight of each feature. CCC is also the official evaluation index recommended by the AVEC Challenge in recent years. CCC (ρ_c) is a measure of how well the prediction values of emotion dimensions (Y) compares to a "gold standard" measurement (X).

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \text{ ,} \tag{2.8}$$

where ρ is the Pearson correlation coefficient (PCC) between the two time series prediction and gold-standard, σ_x^2 and σ_y^2 is the variance of each time series, μ_x and μ_y are the mean value of each. ρ is the PCC coefficient between the two variables, and its calculation formula is as follows:

$$\rho = \frac{\frac{1}{n}\sum_i^n (x_i - \mu_x)(y_i - \mu_y)}{\sigma_x\sigma_y} \text{ .} \tag{2.9}$$

Therefore, the prediction that is well correlated with the gold standard but shifted in value is penalized in proportion to the deviation. This means that the CCC measure combines the PCC with the square difference between the mean of the two compared time series. To measure the weight of each feature, the CCC measure between the prediction values of emotion dimensions and the gold stranded values is used. The value range of CCC is [-1, +1], where +1 indicates that the two sequences are completely positively correlated, -1 is completely negatively correlated, and 0 is completely uncorrelated.

Chapter 3
Auditory-based categorical emotion recognition

3.1 Introduction

Due to the importance of the auditory system in speech perception, research has focused on designing emotion recognition systems by mimicking the human auditory system. In the auditory system, the sound signal is first decomposed by the cochlea, then extracted and compressed by the IHC, the afferent nerve and the central auditory system. Finally, the emotional information of speech is perceived by the auditory cortex. According to the physiological and psychological characteristics of the human auditory system, researchers designed computational auditory models to simulate the various stages of the auditory system, including computational models of cochlear mechanics, IHC, auditory nerve and brainstem signal processing. Different models can be combined to extract various auditory features. The cochlea is the main part of the auditory peripheral, which decomposes the acoustic signal along the length of the basilar membrane into multi-channel acoustic frequency components. Using an auditory filter to simulate the perception process of the cochlea has important applications in understanding the auditory mechanism, speech perception and recognition. Gammatone filter is a kind of commonly used auditory filter simulating cochlea.

SER based on computational auditory models using deep learning methods is a new way to identify the emotional state. Based on CNN and RNN, this study explores the neural network models of feature extraction and time series modeling. Since CNN keeps the spectral-temporal translation invariance for speech signal processing, it is often used to extract high-level features for SER. Mao et al. [56] achieved good performance of SER by trying to learn salient feature maps from the spectrogram of speech using an autoencoder followed by CNN. Lim et al. [53] used deep CNN to extract salient features by transforming the speech signal to 2D representations using a short-time Fourier transform (STFT). Keren et al. [54] and Neumann et al. [55] presented CNN in combination with LSTM to improve the recognition rate based on log Mel filter-banks. SER can be treated as a classification problem based on speech sequences. Some studies deal with it as a sequence classification problem employing RNN or LSTM model.

Chernykh et al. [63] used deep recurrent neural networks to train on a sequence of acoustic features calculated over small speech intervals. Lee et al. [35] extract a high-level representation of emotional states with regard to its temporal dynamics.

Considering that the length of an utterance input to CNN should be the same, it is usually divided into fixed-length segments. Han et al. [64] firstly extracted the segment-level emotion state distributions utilizing the features (F0 and MFCC) based on the DNN model and used an extreme learning machine (ELM) to identify utterance-level emotions. In this chapter, we propose a two-stage emotion recognition method based on Gammatone auditory filterbank using multichannel parallel convolutional recurrent neural networks (MPCRNN). The segmented raw waveform is input to MPCRNN after time-frequency decomposition of Gammatone filterbank to obtain the segment-level emotion probability distribution. Then the utterance-level statistic features are calculated from a segment-level probability distribution and then fed into SVM classifier to predict the emotional states of utterances.

In the two-stage method, the amplitude-frequency response curve of the Gammatone filter is symmetrical about the center frequency and independent about sound levels, which cannot reflect the level-dependent asymmetry of the auditory filter. To solve this problem, Irino proposed a Gammachirp filter to better simulate the basilar membrane filter [59]. Physiological acoustics research shows that the modulation and filtering of the time-domain envelope signal generated by the auditory filter plays an important role in speech perception and understanding. Modulation filtering can obtain high-resolution time modulation cues, which include multi-dimensional information such as acoustic frequency, modulation frequency, amplitude and time [25]. This kind of time modulation cue can be represented by the 3D space of speech signal. Recently some studies proposed 3D convolution models to better capture the spectral-temporal relationship of the feature representations for emotion recognition. Chen et al. [65] proposed attention-based CRNN from a 3D feature representation by computing the log Mel-spectrogram with deltas and delta-deltas for emotion recognition. Kim et al. proposed deep 3D CNN for spectral-temporal feature learning by dividing the speech signal into several sub-segments and these sub-segments contain 2D feature maps with 256 points log-spectrogram for every 20 ms [66]. In this study, the temporal modulation cues from the auditory front-ends contain 3D spectral-temporal representation. The back-ends of the SER system are responsible for extracting high-level features from the 3D representation. CNN has superior feature extraction power inspired from biological neural networks and can extract high-level local feature representations using the spectral-temporal receptive field of the neuron. Therefore, we propose an emotion recognition method in an end-to-end manner using three-dimensional convolutional recurrent neural networks (3D CRNN) based on temporal modulation cues. Temporal modulation cues contain four-dimensional spectral-temporal integration representations directly as the input of 3D CRNN. The convolutional layer is used to extract high-level multiscale spectral-temporal representations, and the recurrent layer is used to extract long-term dependency for emotion recognition.

The rest of the chapter is organized as follows. In Section 3.2, a two-stage discrete emotion recognition method based on Gammatone filterbank is proposed, This section includes the design of MPCRNN architecture based on auditory filterbank, segment-level and utterance-level feature extraction, as well as the analysis of experimental results, and finally analyzes the shortcomings of this method; In Section 3.3, an end-to-end discrete emotion method based on the auditory model is proposed. This section includes the extraction of 3D spectral-temporal features, 3D CRNN, and the experimental results are analyzed and discussed; Finally, Section 3.5 summarizes this chapter.

3.2 Emotion recognition based on auditory filterbank

In this section, we explicitly emphasize a deep learning algorithm based on auditory filterbank to learn discriminative features for SER from the raw waveform. We firstly introduce the details of the two-stage MPCRNN model for SER and then present the methods for segment-level features and utterance-level emotion.

3.2.1 Multi-channel parallel convolution recurrent neural network

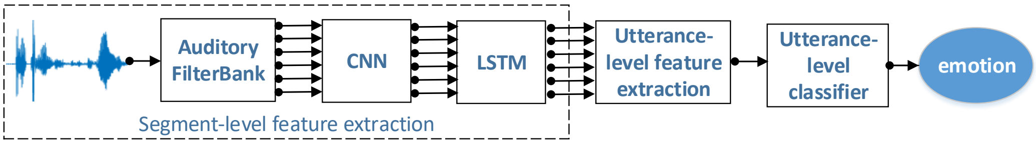

Auditory FilterBank
CNN
LSTM
Utterance-level feature extraction
Utterance-level classifier
emotion
Segment-level feature extraction

Figure 3.1: MPCRNN architecture for SER

Figure 3.1 shows the MPCRNN architecture for SER. The first stage is to extract the segment-level robust and compact features from raw audio. The speech signal is firstly segmented into finite-length chunks. To mimic the function of basilar membrane, multichannel auditory features are extracted based on Gammatone auditory filterbank in each segment. The multichannel auditory features are subsequently processed to obtain a compact representation of the most salient acoustic characteristics for each channel signal in parallel. Hence, we employ parallel 1D convolution for each channel in CNN operation, and then feed each channel data into LSTM as a sequential task to get the relations of each channel. We finally get the emotion probability distribution for each segment using the MPCRNN model.

The second stage is to extract the utterance-level statistical features from the different segments that belonged to the same utterance and feed into an SVM classifier to determine the emotional state of the whole utterance.

3.2.2 Segment-level feature extraction

For extracting the segment-level features, we firstly segment and filter out the raw waveform followed Gammatone auditory filterbank, and train subsequently an MPCRNN to predict the probability distribution of each emotional state.

1) Segmentation and filter for the raw waveform

For segment-level feature extraction, we firstly segment each wav file into 415ms-duration segments. For comparing the traditional method, we get 40 frames for each segment, which includes 25ms windows and 10ms shift.

The energy of each segment y is the sum of the square about each sampling value y_i, as shown in Eq. (3.1). Moreover, according to the energy of segments, all segments are arranged from lowest to highest.

$$\text{energy(y)} = \sum_{i=1}^{n} {y_i}^2 \tag{3.1}$$

There are many segments with low energy, so that we cannot perceive any emotion in these segments.

Therefore, we set a threshold value and filter out the segments whose energies are less than the threshold value in accordance with the subjective listening experiments.

2) Gammatone auditory filterbank

Gammatone auditory filterbank models well the basilar membrane motion of the auditory system. The impulse response of a Gammatone filter is the product of a Gamma distribution and a sinusoidal tone. The bandwidth of each filter is described by an equivalent rectangular bandwidth (ERB), which is a psychoacoustic measure of the width of the auditory filter at each point along the cochlea.

$$g_{\mathrm{t}}(n,t) = \mathrm{A}t^{a_1-1}\exp(-2\pi w_f \mathrm{ERB}(f_{\mathrm{n}})t)\cos(2\pi f_{\mathrm{n}}\, t) \tag{3.2}$$

As shown in Eq. (3.2), where A, w_f and a_1 are parameters, and $\mathrm{A}t^{a_1-1}\exp(-2\pi w_f \mathrm{ERB}(f_{\mathrm{n}})t)$ is the amplitude term represented by the Gamma distribution, f_{n} is the center frequency of the filter, and $\mathrm{ERB}(f_{\mathrm{n}})$ is an equivalent rectangular bandwidth in $f_{\mathrm{n}}(t)$.

In the auditory front-end, the emotional speech signal $s(t)$ is first filtered by a bank of cochlea filters. An output of the n-th channel signal is given by

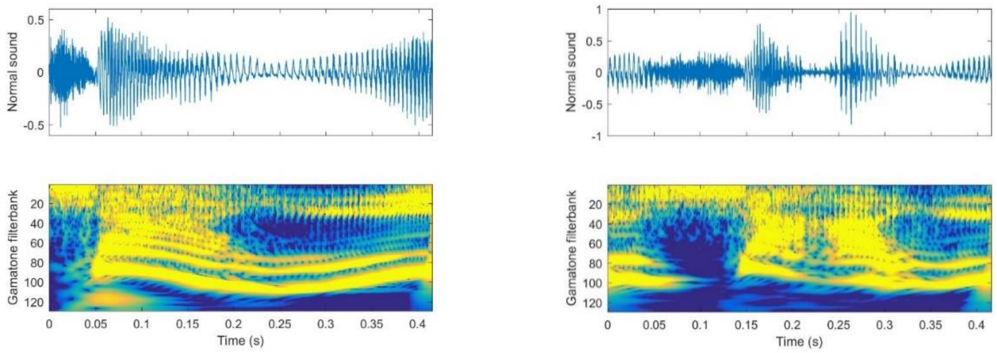
$$s_g(n,t) = g_t(n,t) * s(t),\ 1 \le n \le N, \tag{3.3}$$

where $g_{\mathrm{t}}(n,t)$ is an impulse response of the n-th channel, t is the sample number in the time domain, N is the number of channels in the cochlea filterbank, and $*$ denotes the convolution.

Figure 3.2 shows graphical representations of Gammatone auditory filterbank for different emotions, which are coming from the second segment of wangzheangry201.wav and wangzhehappy201.wav with the same sentence. As shown from the upper part of the figure, we find that the speaking rate is faster and the energy is higher with feelings of angry compared to that of happy. Additionally, we find that graphical representations of Gammatone auditory filterbank for different emotions are different with 128 channels filterbank whose center frequencies equals to 600Hz from the lower part of the figure. In other words, different emotions are reacted with different frequency channels in the human auditory system.

3) Emotion state probability distribution of each segment using MPCRNN

After adopting the Gammatone auditory filterbank from raw audio, we preprocess each Gammatone channel with zero mean and unit variance and then feed into MPCRNN. The normalization ensures that each channel can catch its own characteristic using the same super parameters in MPCRNN. As shown in Fig. 3.3, we employ parallel 1D convolution for each channel in CNN operation, and then feed each channel data into LSTM as a sequential task.



Happy

Angry

Figure 3.2: The graphical representations of Gammatone auditory filterbank for different emotions

For the CNN part, we use a two-layer CNN model to extract different features. The S-Conv layer extracts fine-scale spectral information with a short window from the high sampling rate signal, while the L-Conv layer extracts more long-term characteristics of the speech with a long window. The max-pooling operation is employed in each layer.

For the LSTM part, we consider the multichannel convolutional data as a sequence datum and feed the data into a two-layer LSTM model. Additionally, a fully connected layer is followed by LSTM, which maps the hidden node number (128) into six different emotions: happy, fear, angry, sad, surprise, and neutral. The softmax function is then employed to get the probability distribution of each emotion for each segment. At last, the sequence of the probability distribution over the emotion states is generated from the segment-level MPCRNN.

Given the sequence of the probability distribution over the emotion states generated from the segment-level multichannel parallel convolutional networks, we can form the emotion recognition problem as a sequence classification problem.

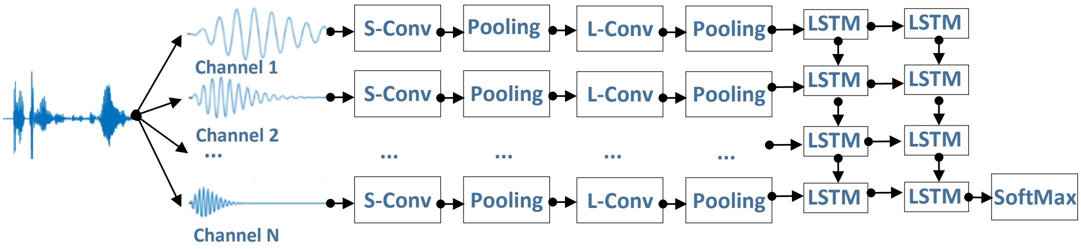


Figure 3.3: Segment-level features extraction using MPCRNN based on Gammatone auditory filterbank from raw waveform

3.2.3 Utterance-level feature extraction

The probability of each segment changes across the whole utterance. Different emotions dominate different regions in the utterance. The true emotion for this utterance is the prominent segment computed from statistics of the segment-level probabilities.

In the two-stage SER method, our experiments are based on the hypothesis that the emotional states of all segments belonged to a certain utterance are the same with the emotional state of this utterance in the training phase. Hence, in the training phase, we assign the same label to all the segments in one utterance. Furthermore, since not all segments in an utterance contain emotional information and it is reasonable to assume that the segments with the highest energy contain most prominent emotional information, we only pick out segments with the highest energy in an utterance as the training samples.

The features in the utterance-level classification are computed from statistics of the segment-level probabilities—the maximal, minimal and mean of the segment-level probability of the kth emotion over the utterance, respectively. The segment number of each utterance is different from the range from one to eleven.

As shown in Fig. 3.4, eighteen utterance-level statistical features are computed with three statistical features for each emotion state and six different emotions totally. The utterance-level statistical features are fed into a classifier for emotion recognition of the utterance. MPCRNN provides good segment-level results, which can be easily classified with a simple classifier. Therefore, we use an SVM classifier with basic statistical features to determine emotions at the utterance-level. In the testing phase, we get the segment-level probabilities distribution using softmax function, and utterance-level emotions are predicted by means of the statistic of the segment-level probabilities.

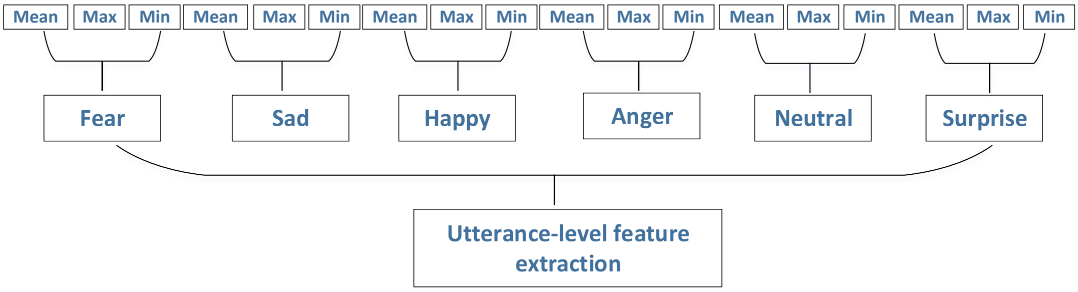


Figure 3.4: Utterance-level features extraction

3.2.4 Experiment results and analysis

1) Experiment setup

We develop MPCRNN as a fast and optimized algorithm for SER based on Gammatone auditory filterbank. We carried out experiments on CISIA emotional speech database. The input signal is sampled at 16 kHz and convert into frames using a 25-ms window sliding at 10-ms each time. So the total length of a segment is 10 ms × 40 + (25 − 10) ms = 415 ms. In fact, emotional information is usually encoded in one or more speech segments whose length varies on factors such as speakers and emotions. According to some studies[67,68], a speech segment longer than 250 ms has been shown to contain sufficient emotional information.

The threshold value of the energy to filter out the segment is 50. We get 15915 segments as the inputs of MPCRNN from 61938 segments in total. Hence, about 25.7% of segments with the highest energy in an utterance are used in the training and the test phase finally.

To get the data from Gammatone auditory filterbank, frequency distributed on ERB_N scales is between 60 Hz and 6 kHz, and the central frequency f_0 equals to 600 Hz. Meanwhile, we apply the four order Gammatone with N equals to four.

2) Hyperparameters for MPCRNN

For training MPCRNN, the S-Conv layer with a 2.5 ms window and 40 kernels in order to extract fine-scale spectral information. The L-Conv layer with a 250 ms window and 40 kernels in order to extract more long-term characteristics of the speech. The pool size equals to 2 in the first layer and 10 in the second layer.

We employ parallel convolutional networks with 32 Gammatone channels, and then use two LSTM layers with 128 cells each. For the sequence data with 32 Gammatone channels, we use many-to-one methods to extract the sequence features.

Additionally, for all random weight initializations, we choose L2-regularizer initialization. We employ cross-entropy as the objective loss function. We then use Adam gradient descent with the learning rate 1e-5. Moreover, we employ ReLU as the activation function, which brings the non-linearity into networks.

To avoid overfitting in training our networks, we employ dropout as a first measure. Dropout has been specifically proposed for cases where labeled data is scarce. It works by randomly omitting a certain percentage of nodes in the network at the training phase while using the full network at the test phase.

Since deep networks need to be trained on a huge number of training databases to achieve satisfactory performance, if the original database contains limited training data, it is better to do data augmentation to boost the performance. Data augmentation is employed by shifting the original speech audio 300ms and 600ms as a new start point for segmentation as a second measure to avoid overfitting.

3) Experiment results

We train the model in a speaker-independent manner, i.e., we use utterances from three speakers to construct the training databases and use the other speakers for the test. The experiments are performed using Nvidia GTX1080 GPU.

In order to analyze the performance of MPCRNN based on Gammatone auditory filterbank, we also obtain firstly probability distribution for emotional recognition using CNN and LSTM respectively based on Gammatone auditory filterbank. After that, we use an SVM classifier with segment-level statistical features to determine utterance-level emotions.

In addition, we compare our approach with other emotion recognition approach. We extract 289 statistics features from 12 MFCC Coefficients for each utterance based on IS09_emotion configure file. We employ these features with SMO (Sequential minimal optimization) classifier, which can get better accuracy than other machine learning methods in these experiments.

There are 2400 utterances for each speaker in the CISIA database, but some utterances are filtered out as the low energy. Finally, 1520 utterances are remained as the train and test set (Class distribution: angry: 346; fear: 237; happy: 218; neutral: 117; sad: 237; surprise: 365). Results obtained for each method are shown in Fig. 3.5.

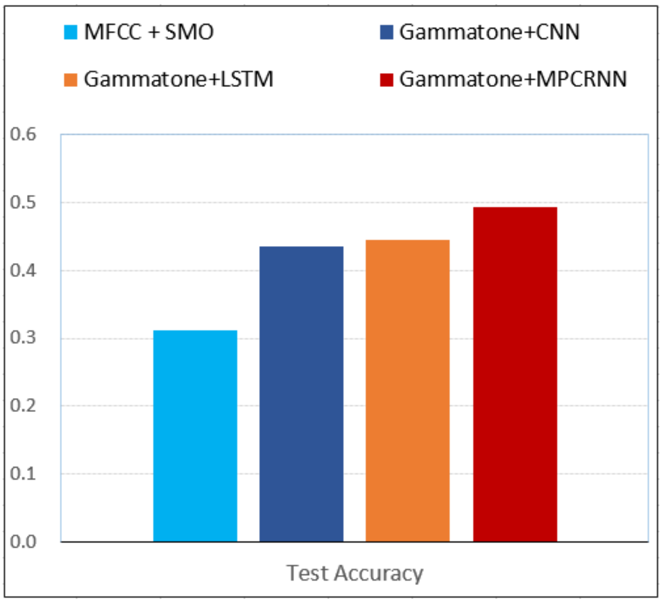


Figure 3.5: Experiment results on CISIA database

In all of the experiments, our study performs better than other methods, with the accuracy equals to 0.494. We found that MPCRNN outperforms LSTM and CNN by around 10% relatively. The accuracy of MFCC and SMO classifiers equals 0.32 in a speaker-independent manner. The proposed approach gives absolute 17.4% better accuracy over the MFCC+SMO approach. Figure 3.6 shows the confusion matrix on CASIA. The recognition rate of surprise is higher than other emotions. A lot of confusion is concentrated between anger and surprise. We think this is because there is no distinguishing between anger and surprise in valence and arousal space. There is some confusion between neutral and surprise. This is because the neutral has the least samples to extract the salient features.

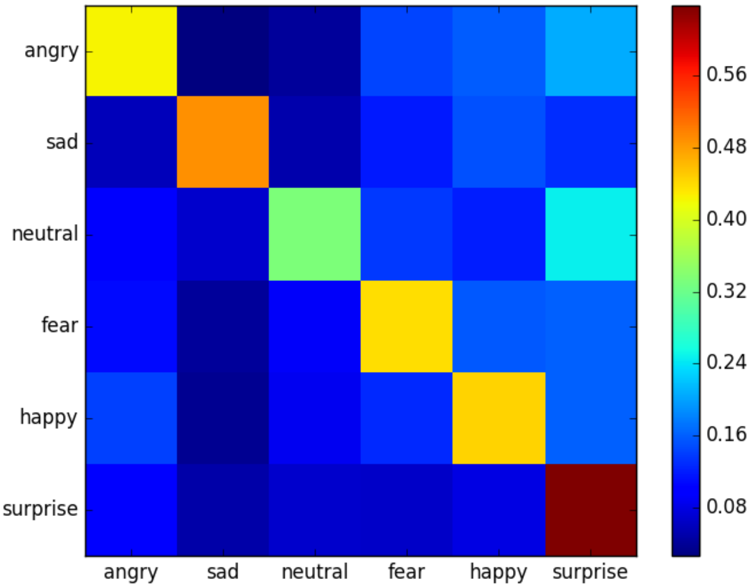


Figure 3.6: Confusion matrix on CISIA database

3.2.5 Summary

In this section, we studied the recognition of emotional speech by utilizing Gammatone auditory filterbank to train a deep model that combines multichannel parallel convolutional recurrent neural networks. We estimated emotion states for each speech segment in an utterance, constructed an utterance level feature from segment-level estimations, and then employed an SVM classifier to recognize the emotions for the utterance. Our experimental results indicate that this approach substantially boosts the performance of emotion recognition from speech signals and it is very promising to use neural networks to learn emotional information based on Gammatone auditory filterbank.

However, part of the corpus cannot involve the two-stage model training, and temporal envelope modulation, which plays an important role in speech perception and understanding, is not investigated. For these reasons, we should extract the joint spectral-temporal features from temporal modulation cues to accurately describe emotion.

3.3 Emotion recognition based on modulation filterbank

3.3.1 Modulation perception model

1) Auditory signal processing

The spectral-temporal representations are extracted using the signal processing steps depicted in Fig. 3.7. In the auditory front-end, the emotional speech signal $s(t)$ is first filtered by a bank of Gammachirp auditory filters. The output of the nth channel signal is given by

$$s_g(n,t) = \ g_c(n,t) * s(t),\ 1 \le n \le N, \tag{3.4}$$

where $g_c(n,t)$ is the impulse response of the *n*-th channel, t is the sample number in the time domain, N is the number of channels in the auditory filterbank, and $*$ denotes the convolution.

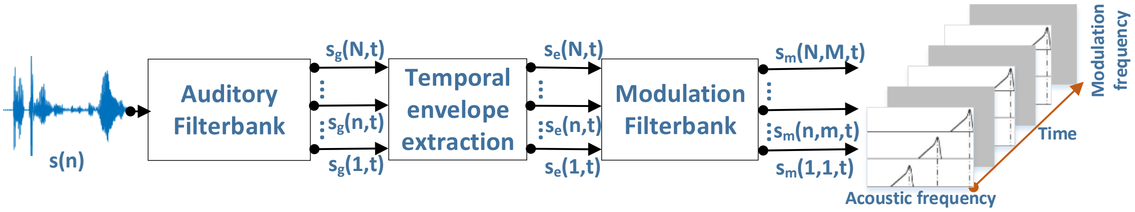


Figure 3.7: Signal processing steps to extract spectral-temporal representation

The center frequencies of these filters are proportional to their bandwidths, which in turn are characterized by the equivalent rectangular bandwidth ($\mathrm{ERB_N}$) [57]:

$$ERB_N(f_\mathrm{n}) = \frac{f_\mathrm{n}}{Q_{ear}} + B_{min}, \tag{3.5}$$

where f_n is the center frequency of the nth filter, Q_ear is an asymptotic filter quality at large frequencies, B_min is minimum bandwidth at low frequencies. Filter quality is a measure of its center frequency divided by the bandwidth. The most widely accepted is provided by [69] in which Q_ear and B_min are 9.26449 and 24.7, respectively. This impulse response of Gammachirp filter is the product of the Gamma distribution and sinusoidal tone.

$$g_c(n,t) = \mathrm{A}t^{a_1-1}\exp\big(-2\pi w_f ERB_N(f_\mathrm{n})t\big)\cos(2\pi f_\mathrm{n}\, t + c_1\ln(t) + \varphi), \tag{3.6}$$

where $\mathrm{A}t^{a_1-1}\exp(-2\pi w_f ERB_N(f_\mathrm{n})t)$ is the amplitude term represented by the Gamma distribution, A, a_1 and w_f are the amplitude, filter order, and bandwidth of the filter, respectively. The $c_1\ln(t)$ term is the monotonic frequency modulation term, φ is the original phase, and $ERB_N(f_\mathrm{n})$ is a bandwidth of the auditory filter in f_n. The chirping properties of the Gammachirp filter are largely determined by those of its “passive” asymmetric filter at all levels and have been shown to fit those of auditory nerve fibers well [59].

The temporal amplitude envelope is extracted using the Hilbert transform to calculate the instantaneous amplitude $s_\mathrm{e}(n,t)$ of the *n*-th channel signal. The $s_\mathrm{e}(n,t)$ is computed from $s_g(n,t)$ as the magnitude of the complex analytic signal $\widehat{s_g}(n,t) = s_g(n,t) + \mathrm{j}\mathcal{H}\{s_g(n,t)\}$, where $\mathcal{H}\{\cdot\}$ denotes the Hilbert transform. Hence,

$$s_{\mathrm{e}}(n,t) = \left|\widehat{s_g}(n,t)\right| = \sqrt{s_g^2(n,t) + \mathcal{H}^2\{s_g(n,t)\}}. \tag{3.7}$$

Furthermore, the *m*-th modulation filter in the nth channel signal is used to obtain the spectral-temporal modulation signal $s_{\mathrm{m}}(n,m,t)$.

$$s_{\mathrm{m}}(n,m,t) = m_{\mathrm{f}}(m,t) * s_{\mathrm{e}}(n,t), \qquad 1 \le m \le M, \tag{3.8}$$

where $m_{\mathrm{f}}(m,t)$ is the impulse response of the modulation filterbank and M is the number of channels in the modulation filterbank.

This type of signal generates a frequency-domain-specific time-domain signal for each sub-channel and many sub-channels comprise the 3D spectral-temporal representation. Due to the high time-resolution of the spectral-temporal representations, a reduction in the number of samples for the time domain has to be carried out. The reduction in the time-resolution is simply carried out by downsampling spectral-temporal representations with an 800-Hz rate. This operation reduces the sequence length by a factor of 20.

2) Spectral-temporal representations

A modulation filterbank is used to extract the spectral-temporal modulation representations over the joint acoustic-modulation frequency plane. By incorporating the cochlear filterbank and the modulation filterbank, a richer 4D spectral-temporal representation is formed and used to analyze spectral and temporal relations. Figure 3.8 shows the modulation representation for the four emotions with a time-averaged pattern, where each one shown is the average over all the time frames for an emotion. “AFC” and “MFC” denote the acoustic and modulation frequency channels, respectively.

Such representations show that the energy of human vocal sound is mostly concentrated at 10 to 15 acoustic frequency channel for anger and happiness and at 5 to 12 acoustic frequency channel for neutral emotion and sadness. The energy is mostly concentrated at the lower modulation frequency channel with a peak at 4 Hz for neutral emotion. The peak shifts to a higher modulation frequency for anger and happiness, suggesting a faster speaking rate for these emotions. Happiness, however, shows a more abundant energy distribution in higher acoustic channels compared to anger’s energy distribution. In contrast to anger and happiness, neutral emotion, and sadness exhibit lower modulation frequency more prominently, suggesting lower speaking rates. The neutral emotion, however, also exhibits a prominent energy distribution in higher acoustic channels between 20 and 25. Sadness exhibits a discriminative energy distribution in the lower acoustic frequency channels over all modulation frequency channels.

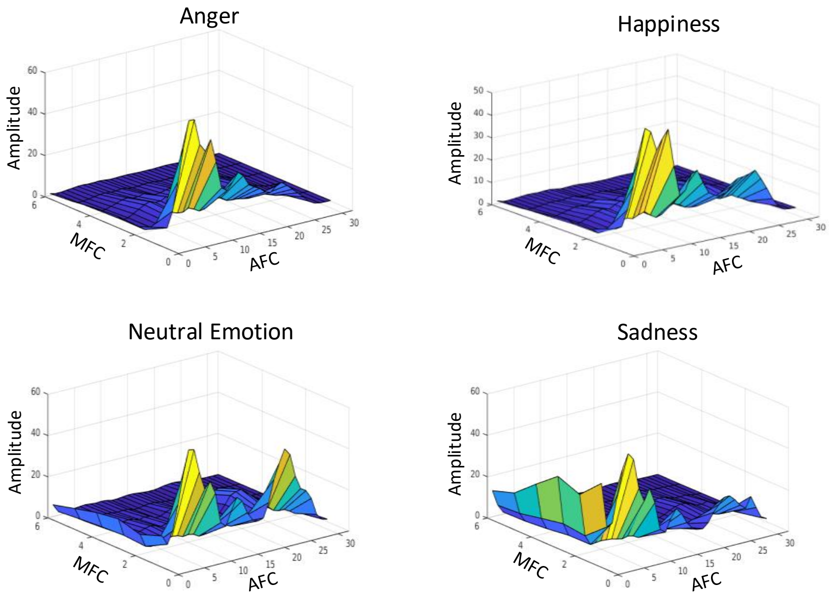


Figure 3.8: Time-averaged modulation representation of sounds

This shows that different emotions have discriminative spectral-temporal modulation representations, which are suitable to extract high-level spectral-temporal representations for convolutional networks.

3.3.2 Three-dimensional convolution recurrent neural network

1) 3D CRNN model

Inspired from biological neural networks, shadow or deep artificial neural networks were designed to extract features. CNNs can extract high-level multiscale spectral-temporal representations using different receptive fields. RNNs can handle long-range temporal dependencies. For processing audio signals, CNNs/RNNs are used to achieve the function of the primary auditory cortex.

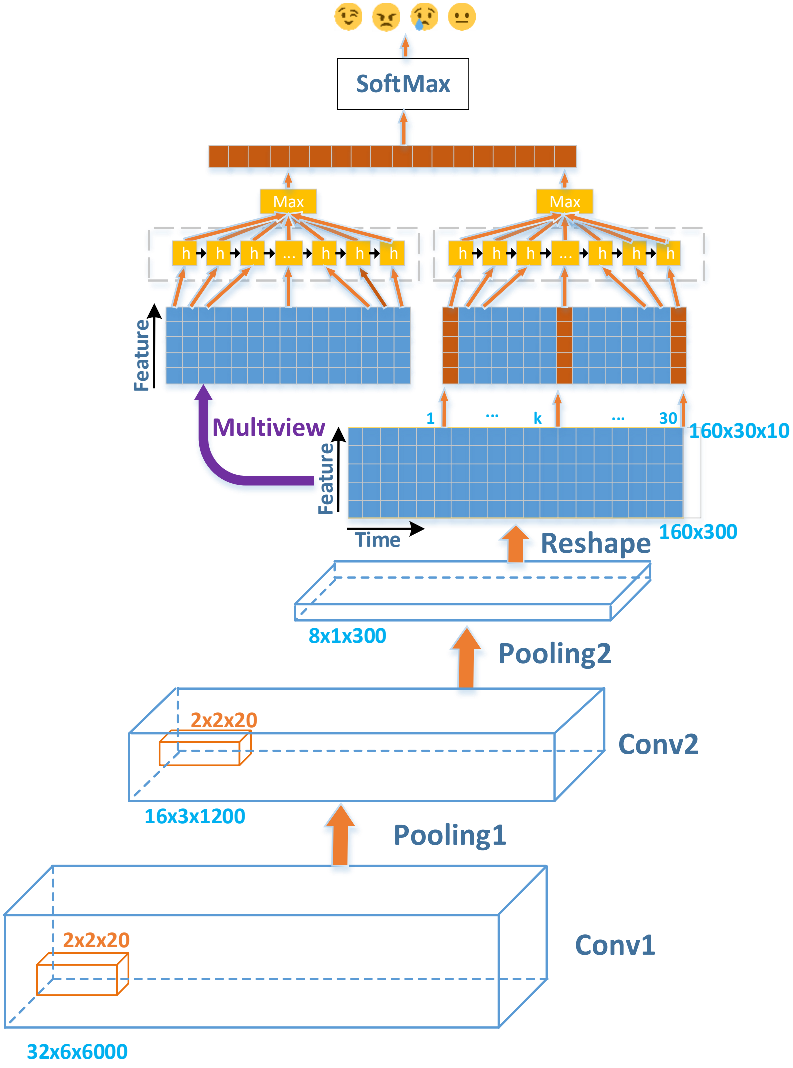


Figure 3.9: Overview of 3D CRNN model

We put forward a 3D CRNN model combining a CNN and RNN for emotion recognition from speech. Figure 3.9 shows an overview of the proposed methods. First, we feed the spectral-temporal representations into the 3D CNN to learn high-level multiscale spectral-temporal representations straightforwardly for a sequence of varied length. Nevertheless, LSTM/RNN is more suitable to learn temporal information.

Eventually, fully connected features are generally used as the LSTM input, but keeping the spatial correlation information in LSTM processes enables more informative spectral-temporal representations to be learned.

Table 3.1: 3D CRNN architecture

Layer	Input size	Output size	Kernel	Stride
Conv1	32x6x6000	32x6x1200	2x2x20	1x1x5
Pool1	32x6x1200	16x3x1200	2x2x1	2x2x1
Conv2	16x3x1200	16x3x600	2x2x20	1x1x2
Pool2	16x3x600	8x1x300	2x3x2	2x3x2
RNN1	10x30x160	30x128	-	-
RNN2	30x128	128	-	-
MV	128	3x128	-	-
FC	3x128	4	-	-

2) 3D convolutional layer

3D CRNN architecture is described in Table 3.1. The first convolutional layer (Conv1) is used to extract 3D features that are composed of acoustic frequency, modulation frequency, and short-time windows. These features are another time sequence, which is the input of the second convolutional layer (Conv2) that models spectral-temporal representations. The data format of the input and output data is reported as "DxHxW", where D, H, and W are the data in the acoustic frequency channels (depth), modulation frequency channels (height), and time sequence (width), respectively. Specifically, the input size is 32x6x6000. Additionally, the shape of the kernels is [2, 2, 20] in the conv1 and conv2 layers following the max-pooling operation. Finally, we get the output of pool2 with the shape of 8x1x300 and then reshape it to 2D shapes. The batch size and convolution filter size are equal to 20. Batch normalization is used before each convolutional layer. Experiments in this study also demonstrate that there will be a substantial speedup in training when using batch normalization.

3) Recurrent layer

We also use two recurrent layers to obtain different scale dependencies using the first recurrent layer (RNN1) for relatively short-term dependencies and the second recurrent layer (RNN2) for utterance-level dependencies.

Figure 3.10 shows the first and second recurrent layer. For RNN1, the input of the layer is 300x160, representing the time sequence length and feature size, respectively. The time sequence is divided into 30 windows, and each window includes 10 time frames. The time sequence is fed frame by frame into the first recurrent layer. Then, the hidden states of the recurrent layer along the different frames of the window are used to compute the extracted features[54].

The output of this layer for each window is the cell state vector of the last time frame in each window. For each window, each layer extracts 128 features. Finally, we create a new sequence with a length of 30x128 to put into RNN2. For RNN2, the whole sequence is fed into the LSTM model, and max-pooling is used to generate 128 feature sequences. After applying max-pooling, the resulting sequence contains temporal features of the sequence and can be fed into the fully connected (FC) layer for classifying.

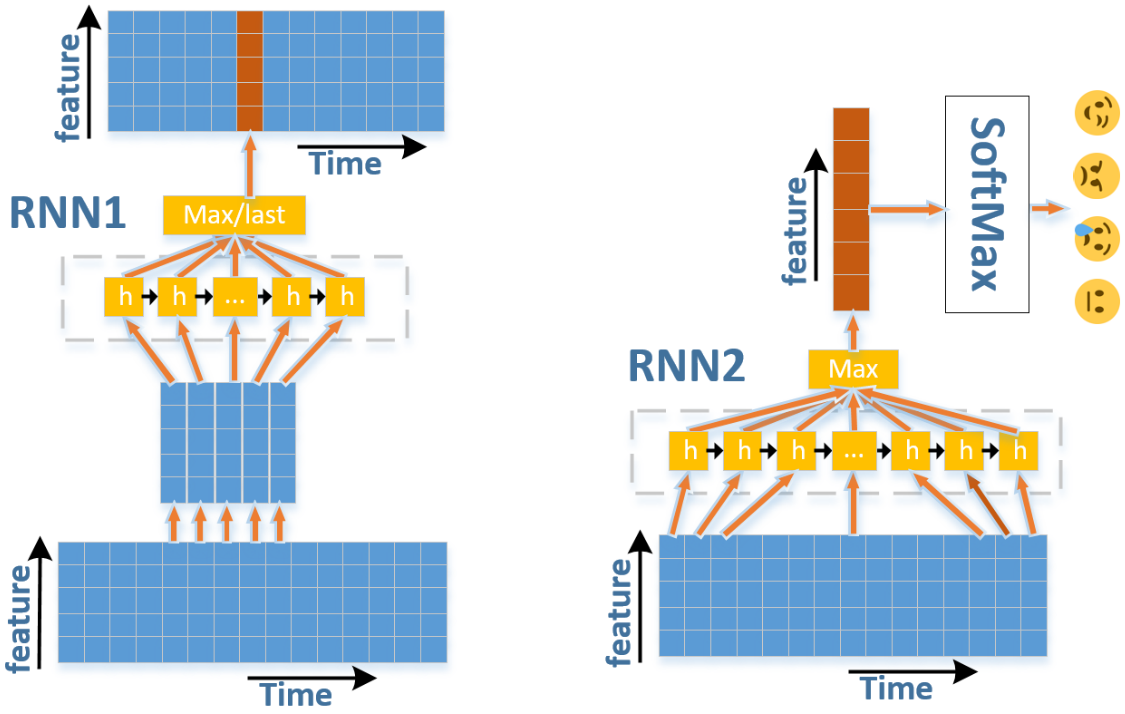


Figure 3.10: First (left) and second (right) recurrent layers

4) Multi-view features

For RNN1, 30 time windows are obtained, and each time window includes 10 time frames. Moreover, we utilize multi-view (MV) features to obtain more information. In this study, we just use unidirectional LSTM because of the varied length for each utterance in the database. For obtaining more dependency information, we shift the time sequence twice, and each shift is equal to 3 and 6, respectively. Finally, we feed this shifted time sequence into RNN2 as a new sequence.

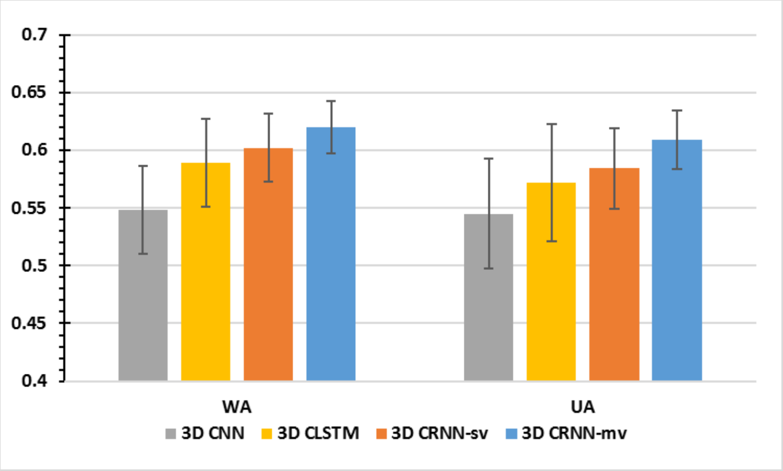


Figure 3.11: Caption of the figure Comparison of recognition accuracy of different models on IEMOCAP database

3.3.3 Experiment setup

1) Setup for modulation spectral features

For our experiments, we used the interactive emotional dyadic motion capture (IEMOCAP) database. Since the input length for a CNN has to be equal for all samples, we set the maximal length to 7.5 s (mean duration plus standard deviation). Longer turns were cut at 7.5 s, and shorter ones were padded with zeros.

We first applied a pre-emphasis filter to the signal to amplify the high frequencies to compensate for the energy loss in the outer-middle ear and then used normalization to remove the difference of the speakers by mapping the values of signals to mean 0 and the standard derivation to 1.

Furthermore, we introduced the compressive Gammachirp filterbank to accommodate the compressive characteristics. To get the data from the Gammachirp filterbank, the frequency distributed on the ERB_N scales was between 100 Hz and 8 kHz. The modulation filterbank was also used to control the envelopes of octave bands from 2 to 64 Hz, consisting of one low-pass filter and five band-pass filters. The detailed setup is shown in Table 3.2.

2) Hyperparameters for 3D CRNN

For all random weight initializations, we chose L2 regularization. The parameters were learned in an end-to-end manner, meaning that all parameters of the model were optimized simultaneously using the Adam optimization method with a learning rate of 1e-4 to minimize the chances of having a cross-entropy objective. Moreover, we used a ReLU as the activation function, which brought the non-linearity into the networks. To avoid overfitting when training our networks, we used a dropout rate of 0.5 after the second recurrent layer.

Table 3.2: Setup for modulation spectral features

Name	Value
Sampling frequency	16000 Hz
Modulation filterbank sampling frequency	800 Hz
Gammachirp channels	32
Modulation sub-channel	6
Sound pressure level	60 dB

3.3.4 Experiment results and analysis

1) Comparison Experiments

There were three comparison experiments named 3D CNN, 3D CLSTM, and 3D CRNN-sv. All these models had the same layers from conv1 to pool2 with the shape of 300x160.

3D CNN: Adding two extra 2D convolutional layers and pooling layer (with 2x2 kernel and 2x2 stride) onto the top of pool2, and then was followed by a fully connected layer.

3D CLSTM: Similar to the 3D CRNN model except without the RNN1 and MV layer. For RNN2, the whole sequence with the shape of 300x160 was fed into the LSTM model, and max-pooling was used to generate 128 feature sequences.

3D CRNN-sv: This was a single-view way for the 3D CRNN model. Similar to the 3D CRNN model except without the MV layer. The output size of FC was 128.

2) Experiments results

To train the models in a speaker-independent manner, we used leave-one-session-out cross-validation. We used utterances from eight speakers to construct the training databases and used the other two speakers for the test.

We used two measures to evaluate the performance: WA and UA. WA is the classification accuracy of the entire test data set, and UA is the average of the classification accuracy for each emotion. The results obtained for each method are shown in Fig. 5. They show that the 3D CRNN with multi-view results in better recognition accuracy with 61.98% and 60.93% in WA and UA measures. This shows that more multiscale information was obtained from the multi-view model. The results also show that the 3D CNN had poorer accuracy than that of the other models because of the absence of a recurrent layer. This also demonstrates the importance of the sequential dependencies information for emotion recognition from speech.

Table 3 shows that the proposed method outperformed the other methods. Han et al. [64] firstly extracted the segment-level emotion state distributions utilizing the features (F0 and MFCC) based on the DNN model and used an ELM to identify utterance-level emotions. Chernykh et al. [63] proposed a CTC approach based on RNN to recognize the utterance-level emotions utilizing MFCC and spectrum properties like flux and roll-off features. The method of Ghosh et al. [70] learns utterance specific representations by a combination of stacked autoencoders and bidirectional LSTM trained on 128 bin FFT spectrograms. Overall, the proposed approaches significantly outperform the previous best accuracy result with 5.88% (from 56.1% to 61.98%) and 6.93% (from 54% to 60.93%) absolute accuracy improvement in WA and UA measures, respectively.

Table 3.3: Comparison of the proposed method and other methods on IEMOCAP database

Method	Features	Models	WA	UA
Han et al.[64]	MFCC and F0	DNN-ELM	54.3%	48.2%
Ghosh et al. [70]	FFT spectrograms	BLSTM-autoencoder	48.1%	49.09%
Chernykh et al. [63]	MFCC and spectrum	RNN with CTC	54%	54%
Neumann et al. [71]	13 MFCCs	Attentive CNN	56.1%	-
Our work	Auditory features	3D CRNN	61.98%	60.93%

3.4 General discussion

In the two-stage method, we set an energy threshold and filter out the segments whose energy is less than the threshold. Eventually, about 25.7% of segments are obtained to train the model. We also try to reduce the threshold to increase the training data but found that these low-energy segments have little effect on emotion recognition. After using the energy-based filtering method, some utterances have 11 segments, but some utterances are completely filtered out because of the low energy. Each speaker in the CISIA database has 2400 utterances. At last, only about 60% of these utterances can recognize emotion when they are used for testing. This shows that the two-stage method based on energy has obvious defects. In the end-to-end emotion recognition experiment, each utterance is processed by a soft segmentation method, all the data will not be filtered out, and then use the max-pooling method to automatically grasp the significant parts of the speech.

Secondly, the two-stage method only considers the use of the auditory filter, and does not consider the spectral-temporal modulation cues, which are more important for speech perception. In the end-to-end mode, a modulation filterbank is introduced to generate high-resolution spectral-temporal modulation cues provided by the time domain envelope and its modulation frequency components. These cues contain multi-dimensional information. Therefore, an end-to-end 3D CRNN is designed to extract the high-level emotion feature sequence from the spectral-temporal representation and construct the temporal-dependence of the sequence. The chapter studied auditory-inspired end-to-end recognition of emotional speech using a 3D CRNN model based on temporal modulation cues. Convolutional networks can reconstruct multiscale spectral-temporal representations, and recurrent networks can obtain the long-term dependencies for emotion recognition. The experimental results demonstrate that our method is an effective way to design an emotion recognition system by mimicking the human auditory system.

3.5 Summary

In this chapter, we first investigated the two-stage emotion recognition from multichannel acoustic frequency components of Gammatone filterbank. Since part of the corpus do not involve the two-stage model training, and temporal envelope modulation is not considered, we proposed the end-to-end emotion recognition using a 3D CRNN model based on temporal modulation cues. Convolutional networks are used to learn the joint spectral-temporal representations from temporal modulation cues, and recurrent networks are used to obtain long-term dependencies for emotion recognition. The experimental results demonstrate that proposed methods are effective to identify the emotional states by mimicking the human auditory system.

However, to reduce the training cost, the speech sequence is segmented into non-overlapping subsequences through soft segmentation in the end-to-end method. These discontinuous segment-level features cannot fully reflect the dynamic changes of emotions. Therefore, how to effectively simulate the auditory system to capture salient emotion regions is also an important issue of SER. Additionally, the modulation frequency components in the end-to-end method only include the local information about variations of intensity and duration. The periodicity information is also effective for emotion recognition, so whether this information can be extracted from temporal modulation cues.

Chapter 4

Attention-based categorical emotion recognition with auditory front-ends

4.1 Introduction

Recent CNNs show powerful abilities of feature learning and have been used for acoustic modeling and feature extraction for SER. Inspired by auditory signal processing in chapter 2, we proposed an end-to-end SER system using 3D CNNs to learn a joint spectral-temporal feature from temporal modulation cues containing acoustic frequency components, modulation frequency components, and temporal features. The modulation frequency components consist of six filters spaced on a logarithm scale from 2 to 64 Hz. Such modulation frequency components include the local information about variations of intensity and duration. However, it did not take into account obtaining the periodicity information about F0 from the modulation frequency band. The frequency band between about 50 and 500Hz is related to the periodicity information about F0, which has been shown to be important for speech perception [72]. To obtain both the local features and periodicity information, in this study, we improve the 3D convolution model by increasing the modulation filters and reducing the convolutional kernel size.

To capture the variations of local features and periodicity information from the feature sequence, we need to extract utterance-level features for classifying emotional speeches through time series modeling. Long short-term memory recurrent neural networks (LSTM-RNNs) have powerful abilities of time series modeling to handle temporal dynamic information. LSTM can effectively capture the long-range time dependencies for sequence classification. However, it cannot avoid the slow training speed caused by backpropagation-through-time (BPTT) in long sequences. To reduce the training cost, in the end-to-end method of chapter 3, the time sequence is divided into non-overlapping subsequences in the extraction of segment-level features. These discontinuous segment-level features cannot fully reflect the dynamic changes of real emotions. From a cognitive point of view, people can obtain important information by scanning the temporal sequence continuously and transmit it for higher-level processing. In addition, people have superior abilities in paying attention to the emotional regions, meanwhile ignoring the emotionless regions. Most of the studies did not take into account the human mechanism how to focus on the emotional segments while ignoring the emotionless segments. An utterance consists of a number of voiced and unvoiced segments. The voiced segments can express emotion more than the unvoiced ones. It is unknown what kind of auditory features attract humans to pay more attention to the salient regions of emotion representation. Therefore, we investigate the relation of the auditory features and human attention mechanism and propose a sliding recurrent method to realize the attention mechanism. In the temporal attention method, the continuous segment-level internal representations are extracted by a sliding window, and are used to capture the salient regions of emotion representation.

To fully utilize the human auditory mechanism and attention mechanism, in this chapter, we begin with the investigation of temporal modulation cues from auditory front-ends and then find out a method to capture the salient regions of emotion representation. Based on the achievements, we propose a joint deep learning model that combines 3D convolutions and attention-based sliding recurrent neural networks (ASRNNs) as the back-ends of the SER system. To show the benefit of the proposed model, we evaluate it on the IEMOCAP [41] and MSP-IMPROV [42] databases by comparing various models with the proposed model. Our results show that the proposed model can achieve better results compared with the traditional model on both databases. We also conduct a subjective evaluation to investigate the relevance between the attention patterns of the temporal attention model and human attention in perceiving emotional speech.

The rest of the study is organized as follows. In Section 4.2, we introduce the auditory front-ends to produce temporal modulation cues. Section 4.3 details the 3D convolutions to learn a joint spectral-temporal feature representation from those cues and ASRNNs to focus on the salient regions of emotion representation. In Section 4.4, we also investigate the impacts of experiments on different situations. We discuss the implications of this study in Section 4.5. Finally, we draw conclusions in Section 4.6.

4.2 Joint spectral-temporal representations

4.2.1 Overview of the emotion recognition method

An overview of the proposed SER method is illustrated in Fig. 4.1. The auditory front-ends of this system are used to functionally simulate the signal processing in the auditory system, as depicted in the left part of Fig. 4.1.

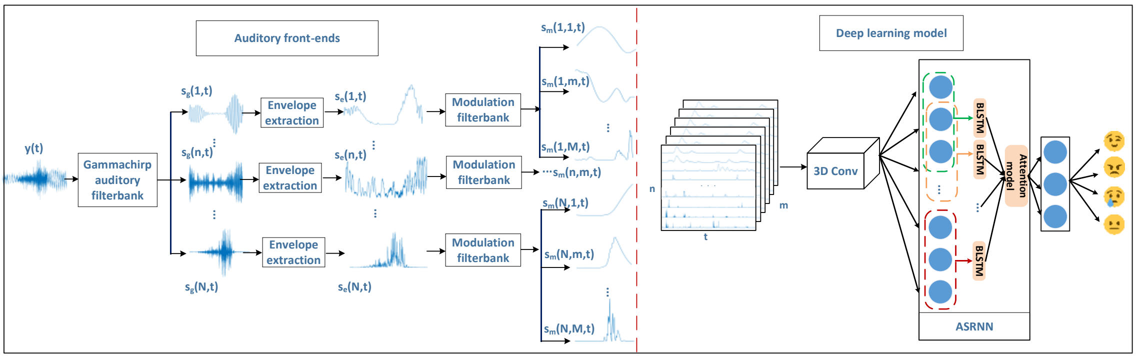


Figure 4.1: Speech emotion recognition with auditory front-ends

The auditory front-ends are composed of three parts: auditory filterbank, temporal envelope extraction and modulation filterbank. The auditory filterbank is responsible for decomposing speech signals into acoustic frequency components as a function of the acoustic frequency analyzer in the cochlea. In this study, we use the Gammachirp filterbank [59] as the auditory filterbank because this filter is adequate for reproducing psychophysically estimated human auditory filters over a wide range of center frequencies and levels [73,74]. Furthermore, temporal envelope extraction from the acoustic frequency components is used to effectively simulate the mechanical-to-neural signal transduction in the IHCs. Modern psychophysical models of temporal modulation processing suggest that the temporal envelope is processed by joint spectral-temporal modulations [11]. The spectral-temporal modulation contains the 3D modulated spectrum with dynamic peaks, which relates directly to speech perception [33]. Hence, the modulation filterbank is introduced to generate 3D joint spectral-temporal representations from the temporal envelope.

The back-ends of this system are depicted in the right part of Fig. 1. 3D convolutions are firstly used to extract joint frame-level features, including not only variations information of intensity and duration but also the periodicity information. Further, ASRNNs are used to focus on the salient emotion regions by extracting segment-level features in a sliding window manner and utterance-level features with a temporal attention model.

4.2.2 Modulation-spectral representations

In this study, the acoustic F0 and the modulation frequency will not overlap, because the modulation filter acts on the envelope signal of each subband from Gammachirp instead of the original voice signal. The modulation frequency band between about 50 and 500Hz is related to the periodicity information about F0. The periodicity information has been shown to be important for speech perception. To obtain both the local features and periodicity information, in this study, we improve the 3D convolution model by increasing the modulation filters and reducing the convolutional kernel size.

The modulation frequency components consist of six filters spaced on a logarithm scale from 2 to 64 Hz. Such modulation frequency components include the local information about variations of intensity and duration. However, it did not take into account of obtaining the periodicity information about F0 from the modulation frequency band. The modulation frequency band between about 50 and 500Hz is related to the periodicity information about F0, which has been shown to be important for speech perception [72]. To obtain both the local features and periodicity information, in this study, we improve the 3D convolution model by increasing the modulation filters and reducing the convolutional kernel size.

Figure 4.2 shows the different emotion examples of the modulation spectral representation with 32 acoustic channels and nine modulation channels from the IEMOCAP database. Each utterance comes from the same speaker, named Ses01F_impro04_F000 (Neutral emotion), Ses01F_impro05_F009 (Angry), Ses01F_impro03_F001 (Happiness), and Ses01F_impro02_F005 (Sadness), respectively. The y-axis and x-axis of these representations are acoustic and modulation channels, respectively. Both channels are spaced on a logarithm-scale frequency. Modulated signals with standard deviation are projected into the modulation and acoustic frequency space. Panels (a) to (d) in Fig. 4.2 shows the modulation spectral representations of anger, happiness, neutral emotion and sadness, respectively. As slow modulation frequency, particularly below 16 Hz (modulation channel equals to 4), can extract local information about variations of intensity, duration, attack, decay, and segmental cues of speech [16]. From these panels, we can find that the different emotion has different low-frequency modulation information, suggesting they could for speech perception be discriminated from each other. In chapter 3, we used six modulation filters to extract low-frequency information (below 64 Hz) for emotion recognition.

Although fast modulation frequency is less important than slow modulation frequency, it still contains the periodicity information to reflect emotional changes. Figure 4.2 also shows that the periodicity information is retained between the seventh and ninth modulation channels. In addition, for the same fast modulation frequency, it shows that the acoustic frequency of anger and happiness is higher than that of sadness and neutral emotion. For this reason, we use nine modulation filters with an upper limit of modulation frequency (512 Hz) instead of six filters to obtain periodicity information for emotion recognition.

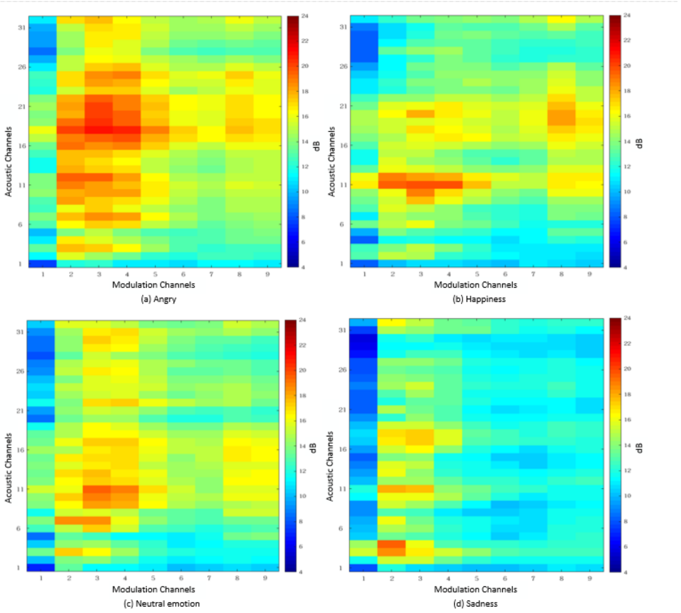


Figure 4.2: Different emotion examples of the modulation spectral representation with 32 acoustic channels and nine modulation channels from the IEMOCAP database

4.2.3 Three-dimensional convolution

The architecture of 3D CNNs is described in Table 4.1. The first convolutional layer (Conv1) is used to extract 3D features that are composed of acoustic frequency, modulation frequency, and time sequences. These features are fed into the next two convolutional layers (Conv2 and Conv3) to model high-level feature representations for time series. The data format of the input and output data is designed as "DxHxW," where D, H, and W are the data in the acoustic channels (depth), modulation channels (height), and time sequence (width), respectively. In this study, the input size is set as 32x9x6000 and the size of the kernels is 2x2x4. To reduce computational complexity, the stride for Conv1 is set to 1x1x2, and that for the other convolutional layers is set to 1x1x1. Each convolutional layer includes batch normalization and ReLU operations. Batch normalization is used to accelerate the training of deep networks [75]. The first pooling layer (Pool1) before conv2 has a kernel size of 2x2x1 and stride of 2x2x1 with the max-pooling operation. The second pooling layer (Pool2) has a kernel size of 2x2x2 and a stride of 2x2x2. This means that spectral-temporal pooling is executed on Pool2. The third pooling layer (Pool3) has a kernel size of 2x1x2 and stride 2x1x2. This means that the acoustic frequency channel and temporal pooling is executed while the modulation frequency channel remains on Pool3. The max-pooling operations in each pooling layer is used to extract robust features against background noise, especially for the waveform signals. These three pooling layers reduce the output size of the time sequence by a factor of 20 on the temporal length. This means that the 3D convolution only learns the frame-level features in 22.5ms for each point. The feature maps of the three convolution layers are 20, 32, and 64, respectively. Finally, we obtain the output of Pool3 with the shape of 750x4x2x64 after transposing the axis of the tensor then reshape it to 2D shapes of 750x512.

Table 4.1: 3D convolutional neural networks architecture

Layer	Input size	Output size	Kernel	Stride
Conv1	32x9x6000	32x9x3000	2x2x4	1x1x2
Pool1	32x9x3000	16x4x3000	2x2x1	2x2x1
Conv2	16x4x3000	16x4x3000	2x2x4	1x1x1
Pool2	16x4x3000	8x2x1500	2x2x2	2x2x2
Conv3	8x2x1500	8x2x1500	2x2x4	1x1x1
Pool3	8x2x1500	4x2x750	2x1x2	2x1x2
Reshape	4x2x750	750x512		

4.3 Attention-based sliding recurrent neural network

Part of the attention system of the brain is involved in the control of thoughts, emotions, and behavior. In the auditory system, selective auditory attention tracks the temporal dynamics of emotion by continuous scanning and encoding of the speech signals [36]. Inspired by the selective auditory attention in the auditory system, we propose an ASRNN model to seize the emotional parts from temporal dynamics information in speech. Among them, a sliding window is used to extract the continuous segment-level emotional features containing temporal dynamics information. Then, a temporal attention model is used to capture the important information related to emotion in each utterance.

1) Sliding recurrent neural networks

The sliding recurrent neural networks (SRNN) are used to continuously extract the intermediate segment-level representations for the short-term sequence depicted in Fig. 4.3. The input of the SRNNs is TxD, where T represents the total length of the time sequence and D represents the feature vector size. x_k is the input to the LSTM block of kth sliding input sequence with Z time frames.

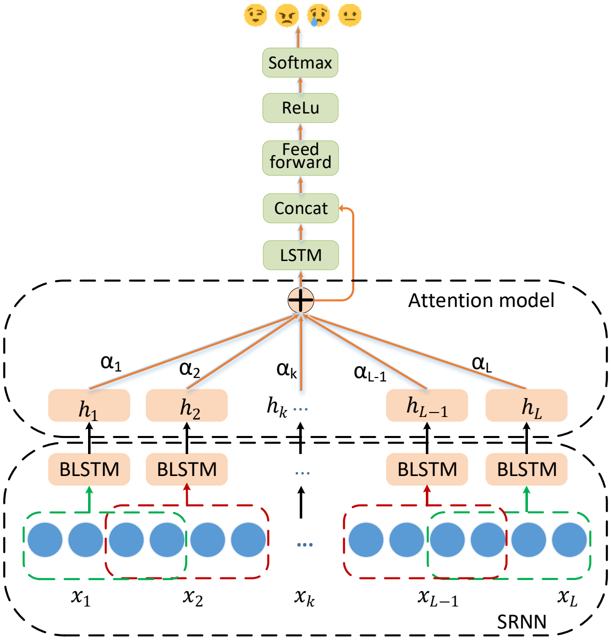


Figure 4.3: Attention-based sliding recurrent networks

$$x_k = \{x_{(k,1)}, \dots, x_{(k,Z)}\}, \qquad x_{(k,t)} \in \mathbb{R}^D, \qquad 1 \le t \le Z \tag{4.1}$$

Each x_k is fed frame-by-frame into the LSTM units. The formulation of LSTM with peephole connections can be described by the following equations:

$$i_{(k,t)} = \sigma(W_{ix}x_{(k,t)} + W_{ih}h_{(k,t-1)} + W_{ic}c_{(k,t-1)} + b_i) \tag{4.2}$$

$$f_{(k,t)} = \sigma(W_{fx}x_{(k,t)} + W_{fh}h_{(k,t-1)} + W_{fc}c_{(k,t-1)} + b_f) \tag{4.3}$$

$$\widetilde{c_{(k,t)}} = tanh(W_{cx}x_{(k,t)} + W_{ch}h_{(k,t-1)} + b_c) \tag{4.4}$$

$$c_{(k,t)} = f_{(k,t)} \odot c_{(k,t-1)} + i_{(k,t)} \odot \widetilde{c_{(k,t)}} \tag{4.5}$$

$$o_{(k,t)} = \sigma(W_{ox}x_{(k,t)} + W_{oh}h_{(k,t-1)} + W_{oc}c_{(k,t)} + b_o) \tag{4.6}$$

$$h_{(k,t)} = o_{(k,t)} \odot tanh\left(c_{(k,t)}\right), \tag{4.7}$$

where $i_{(k,t)}$, $f_{(k,t)}$, $o_{(k,t)}$, $c_{(k,t)}$, and $h_{(k,t)}$ are the input gate, forget gate, output gate, cell state, and output of the LSTM block, respectively, at the current time step t. The weight matrices W_{i*}, W_{f*}, and W_{o*} transform x_k and hidden state $h_{(k,t-1)}$, respectively, to cell update $\widetilde{c_{(k,t)}}$ and three gates $i_{(k,t)}$, $f_{(k,t)}$, and $o_{(k,t)}$. Finally, b_i, b_f, b_o are the additive biases of the input gate, forget gate, and output gate, respectively. The set of activation functions consists of the logistic sigmoid function σ(·), element-wise multiplication ⊙, and hyperbolic tangent function tanh(·).

Specifically, we use a bidirectional LSTM (BLSTM) network in this study, where the sequence of received signals is once fed in the forward direction into one LSTM cell, and once fed in backward into another LSTM cell. The forward LSTM reads the time sequence in its original order and generates a hidden state $fh_{(k,t)} = \{fh_{(k,1)}, \dots, fh_{(k,Z)}\}$ at each time step. Similarly, the backward LSTM reads the time sequence in its reverse order and generates a sequence of hidden states $bh_{(k,t)} = \{bh_{(k,Z)}, \dots, bh_{(k,1)}\}$. The last state of the forward and backward LSTM cells carry information of the entire source sequence. We concatenate the last state of the forward and backward LSTM cells to produce the h_k of k sequence.

$$h_k = [fh_{(k,Z)}, bh_{(k,1)}] \tag{4.8}$$

Each hidden state h_k contains information of each sliding window sequence. The hidden states of the recurrent layer along the different frames of the window are used to compute the extracted features. The output of this layer for each sliding window is the cell state vector of the last time frame in each sliding window. After processing in each sliding window, we shift S time frames to compute the next sliding window with the valid padding. The number of sliding window L is calculated as

$$\mathrm{L} = \ulcorner (\mathrm{T} - \mathrm{Z})/S \urcorner\ . \tag{4.9}$$

The BLSTM has 512 hidden units for both directions in each sliding window. Finally, we create a new sequence with the shape of L x1024 to put into the attention model. The same parameters of the LSTM cell are used in each sliding sequence, and then a new context sequence h is produced.

$$h = \{h_1, \dots, h_L\}, \qquad h_k \in \mathbb{R}^{2D}, \qquad 1 \leq k \leq L \tag{4.10}$$

2) Temporal attention model

Because there are many speech frames that are unrelated to the expressed emotion, such as silence, the attention mechanism is mainly used to focus only on the significant emotional part of the speech signal. Recently, some studies proposed attention models to adjust weights for each of the speech frames depending on their importance based on LLDs using an RNN. The silence regions can be addressed using voice activity detection (VAD) [34] or by null label alignment [35]. Wang et al. [34] proposed an attention model of learning utterance-level representations to improve classification after using a VAD to filter out silence frames and mini-batch training in each utterance. Lee et al. [35] extracted high-level representation of emotional states with regard to its temporal dynamics using the BLSTM approach, in which they assume that different frames should have different labels and the label sequence should be alternating between the utterance-level label and a newly introduced NULL state. Neumann et al. [71] proposed an attentive convolutional neural network (ACNN) to test the emotional discrimination of different feature sets. In addition, a self-attention based deep model [76,77] demonstrated the effectiveness of improving the performances for SER. Unlike these studies, we apply a temporal attention model to the sliding window sequence instead of applying one based on LLDs.

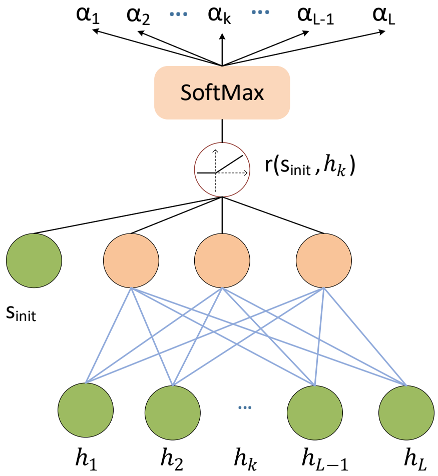


Figure 4.4: Attention weights

Sequence h is fed into feedforward neural networks then concatenated with s_{init}, as depicted in Fig.4. Subsequently, a ReLU is used to produce non-linear transformations $\mathcal{R}(s_{\text{init}}, h_k)$.

$$\mathcal{R}(s_{init}, h_k) = U_k ReLU(s_{init} + W_k h_k + b_k), \tag{4.11}$$

where W_k, U_k are the trainable parameter matrices, b_k is the bias vector, and s_{init} is the initial hidden state of the sliding recurrent sequence. We use the non-linear function of the ReLU due to its good convergence performance. For each h_k, the α_k can be computed as follows:

$$\alpha_k = \frac{exp\left(\mathcal{R}(s_{init,} h_k)\right)}{\sum_{l=1}^{L} exp\left(\mathcal{R}(s_{init,} h_l)\right)} \tag{4.12}$$

We then obtain the attention weights α_k of each sliding sequence from the attention model. The output of the attention layer, $attention\_sum$, is the weighted sum of h.

$$attention\_sum = \sum_{k=1}^{L} \alpha_k h_k. \tag{4.13}$$

The weighted sum of sequence h is fed into a unidirectional LSTM cell to obtain a hidden vector h_s. The features concatenated by h and h_s are fed into feedforward neural networks. Subsequently, we use a ReLU as the activation function, which brings the nonlinearity into the networks. Finally, we use the softmax to produce the emotion state distribution. To avoid overfitting when training our networks, we use a dropout rate of 0.5 before feed-forward layers during training.

Table 4.2: Accuracy comparison of static features on IEMOCAP and MSP-IMPROV databases (%)

Static features	UA	
	IEMOCAP	MSP-IMPROV
IS09	53.4	41.2
emobase2010	54.9	40.9
IS13 ComParE	54.5	40.6
MFCC	51.5	40.5
MSF	52.5	43.2

4.4 Experiment results and analysis

We conduct speaker-independent experiments using the IEMOCAP and MSP-IMPROV datasets. The class distribution is unbalanced in both databases, especially for MSP-IMPROV database, the number of utterances belonging to happy/neutral class more than three times that of angry/sad. UA is a better measurement if the class distribution is not balanced. Hence, we use UA as the performance metric of the proposed framework to avoid being biased to the larger classes.

4.4.1 Results of baseline features

Firstly, we investigate the conventional emotion recognition system with static features that are computed using fixed statistical functions to the hand-crafted LLDs. We extract MFCC, emobase2010, IS09_paraling [49], and IS13_ComParE [51] features use openSMILE toolkit. All features are first normalized by specific z-normalization. Secondly, to investigate the effectiveness of static modulation features on emotion recognition, we also extract the MSFs by calculating the spectral centroid, spread, skewness, and kurtosis from the modulation spectral representation. For each feature set, we train a linear SVM model to recognize the speech emotion using LibSVM [78] and Weka toolkits [79]. All results are presented by leave-one-session-out cross-validation. Table 4.2 shows the accuracy comparison of static features on IEMOCAP and MSP-IMPROV databases. The best result is 54.9 percent for IEMOCAP using the original static features with 1,582 dimensions whereas the best result is 43.2 percent for MSP-IMPROV using the static modulation features with 160 dimensions. The results also show that MFCC features achieve the worst results, which may be due to the minimum number of MFCC features (only 39 dimensions features). Similar to the results from [25], the MSFs perform better than MFCC for emotion recognition on both databases. Emotion information from speech changes dynamically over time, but the static features do not contain temporal dynamics information which plays a key role in the emotion recognition process.

4.4.2 Experiment setup

In the front-end signal processing, we first resample the speech signal with a sampling frequency of 16000 Hz and apply a pre-emphasis filter to compensate for the effect of a sound source. We subsequently use normalization to remove the difference of the speakers by mapping the signal values to mean 0 and the standard derivation to 1 in each utterance. The sound-pressure level is set to 60 dB, which approximates to a normal voice. Furthermore, we introduce the compressive Gammachirp filterbank with 32 filters to provide the compressive characteristics. The frequency of Gammachirp filterbank distributed on the ERB_N scales is between 0.1 and 8 kHz. The modulation filterbank is also used to control the envelopes of octave bands from 2 to 512 Hz, consisting of nine filters (one low-pass filter and eight band-pass filters). The low-pass filter is a 2nd order Butterworth infinite impulse response (IIR) filter with a cut-off frequency of 2 Hz. The cut-off frequencies of the band-pass filters are equally spaced on a logarithm scale from 2 to 512 Hz.

In the back-ends of the SER system, a joint deep learning model combined 3D convolution and ASRNN is used. To train the model with a speaker-independent property, we use leave-one-session-out cross-validation. In each experiment, four sessions are used for training the deep model and one session is divided into two sub-sessions depending on the gender in both databases. For all random weight initializations, we choose L2 regularization. The parameters are learned in an end-to-end manner, meaning that all parameters of the model are optimized simultaneously using the Adam optimization method with a learning rate of 1e-4 to minimize cross-entropy loss. The batch size is 10, and the maximum epoch is 30 with early-stopping. The process stops if the UA does not improve for 8 consecutive epochs.

4.4.3 Effect of different window and shift

SRNNs are used to obtain continuous internal representations while maintaining good computational efficiency. The continuous internal representations can be extracted using a sliding window. At the same time, computational efficiency can be improved by segmenting a feature sequence into multi sub-sequence. However, choosing different lengths of window and shift will affect the recognition accuracy and computational efficiency of the emotional recognition system.

To reach higher recognition accuracy and computational efficiency, we investigate the effect of the sliding window and shift lengths using IEMOCAP database. First, the entire feature sequence is divided into multi-subsequences in a sliding manner. The length of each subsequence is much shorter than the original sequence, and the model can be trained rapidly using BPTT. Then, we run the proposed system five times and obtain the average accuracy in the case of the different sliding window and shift lengths. We consider the different sliding window lengths of 10, 20, 30, 40, 50, and 100, which mean the duration of the sequence from 200 to 2000 ms. We also consider the shift lengths of 5, 10, and 20, which means that it will produce 150, 75, and 38 sliding subsequences in the same padding manner for the duration of the convolutional sequence with 750x512. When the sliding window length is 100 with a shift length of 10, the training time of the ASRNN architecture is close to that of the entire sequence fed into the recurrent networks. Hence, we do not consider a longer sliding window that will take a longer time to train the model. One session in the database is chosen for testing and others for training. We find that the computational efficiency will be improved with the shortening of window length and the lengthening of shift. But in this case, the recognition accuracy will decrease due to the inability to extract more emotional features. In addition, because only the feature of the last time frame in each sliding window is retained, when the window length is too long, not only the computational efficiency will be reduced, but also the recognition accuracy will be reduced. The results obtained for each method are shown in Fig. 4.5. Recognition accuracy is closer when the shift length is 5 or 10, but it became worse when the shift length is 20. This figure also shows that the ASRNN architecture resulted in better accuracy when the sliding window length is 20 or 40. Therefore, we only consider sliding window lengths of 20 and 40 and shift lengths of 5 and 10.

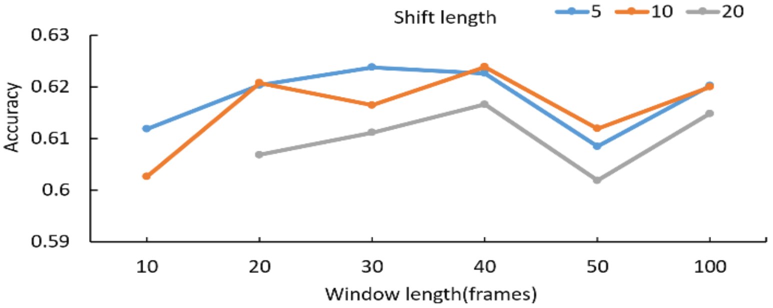


Figure 4.5: The impact of sliding window and shift length on recognition accuracy

Table 4.3 shows the recognition results using different lengths and shift of the sliding window with the ASRNNs architecture for both databases. One can see that the ASRNNs architecture with the sliding window length of 20 and shift length of 10 performed better than the others, whose recognition accuracy is 62.6% for IEMOCAP and 55.7% for MSP-IMPROV. The results are much better than those obtained using the traditional parameters shown in Table 4.2. According to the results, the window length of 20 frames (about 400ms) is suitable for expressing segment-level emotions, while the shift length of 10 is better for classification than that with the shift length of 5. Comparing with the best results of traditional recognition system in Table 4.2, the proposed system achieved +7.7 and +12.5% absolute accuracy improvements on IEMOCAP and MSP-IMPROV, respectively. These results indicate that the proposed system with temporal dynamics information is better to recognize emotional states than the conventional system with static features.

Table 4.3: Accuracy comparison with different sliding-windows and shift lengths in ASRNN architecture on IEMOCAP and MSP-IMPROV databases (%)

Sliding window length	Shift length	UA	
		IEMOCAP	MSP-IMPROV
20	5	62.3	54.9
20	10	62.6	55.7
40	5	61.0	54.2
40	10	62.1	55.3

Tables 4.4 and 4.5 show the confusion matrix of the best results for the IEMOCAP and MSP-IMPROV databases, respectively. In general, the class distributions of the confusion matrix for the different sessions are basically similar. One can see that happiness is easily confused with neutral emotion and vice versa. Anger is more easily misclassified as happiness than happiness being misclassified as anger. Unlike the study [55], the proposed system reduces the confusion between anger and happiness categories to a major extent, especially in MSP-IMPROV. Sadness is easily confused with neutral emotion in IEMOCAP, while it is easily confused with happiness in MSP-IMPROV. The confusion in the proposed method mainly happens between the neutral one and the others. This implies that emotion recognition based on auditory front-ends is basically consistent with people's recognition of emotion. In terms of the databases, the overall performance on IEMOCAP is better than MSP-IMPROV. The reason for this seems to be that the MSP-IMPROV database is highly imbalanced.

Table 4.4: Confusion matrix (%) of ASRNN with an average accuracy of 52.6% on the IEMOCAP database

		Output			
	Emotion	Neutral	Happiness	Anger	Sadness
Input	Neutral	58.5	17.0	8.0	16.5
	Happiness	20.6	55.6	12.7	11.1
	Anger	12.9	18.1	64.4	4.6
	Sadness	15.6	9.4	3.0	72.0

Table 4.5: Confusion matrix (%) of ASRNN with an average accuracy of 55.7% on the MSP-IMPROV database

			Output		
	Emotion	Neutral	Happiness	Anger	Sadness
Input	Neutral	45.0	34.8	8.6	11.6
	Happiness	17.5	67.5	10.3	4.7
	Anger	13.2	25.1	59.7	2.0
	Sadness	22.5	23.3	3.7	50.5

4.4.4 Results of modulation channel, sliding widow and attention model

In order to evaluate the effects of modulation channel number, sliding window and attention model on the SER system, we design a number of comparative experiments in different situations.

First, we evaluate the effects of the nine modulation filterbank in obtaining local features and periodicity information by comparing it to the one with six modulation filters (ASRNN-6MFB). ASRNN-6MFB is set as the same layers as the ASRNN, but different inputs shape of 32x6x6000 result in different kernels and stride. Compare to ASRNN, the difference is that the kernel and stride are 2x1x2 instead of 2x2x2 in Pool2. In addition, the convolutional maps are 40 instead of 64 to keep similar features in each frame. Finally, the output shape is 4x3x750 in pool3. Then this layer is reshaped to 2D shapes of 750x480.

Second, an attention-based recurrent neural network (ARNN) is designed to evaluate whether the sliding window can obtain more temporal dynamics information or not. ARNN is a special case of an ASRNN. That is, when the sliding window length of an ASRNN is equal to the length of the entire convolution sequence and the shift length is equal to 0, it becomes an ARNN. Hence, the attention model is used on the entire time sequence.

Table 4.6: Accuracy comparison (%) between RNN architectures on IEMOCAP and MSP-IMPROV databases

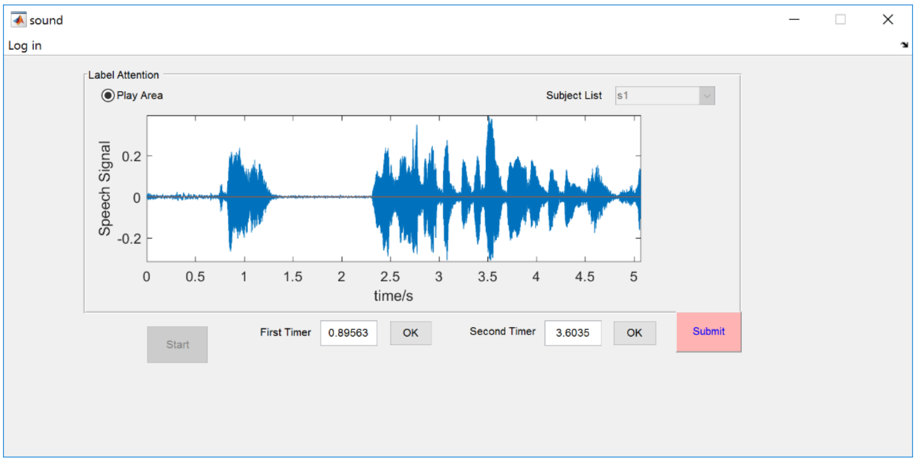
RNN architecture	UA	
	IEMOCAP	MSP-IMPROV
SRNN-Max-pooling	61.5	54.2
SRNN-Mean-pooling	61.7	53.9
ARNN	61.3	55.2
ASRNN-6MFB	61.7	54.8
ASRNN	62.6	55.7

Third, SRNNs with max and mean pooling are designed to evaluate whether the attention model can seize the emotional regions. An SRNN has the same sliding window and shift lengths as the ASRNN. There are two types of pooling used in an SRNN: maximum and average, denoted as SRNN-Max-pooling and SRNN-Mean-pooling, respectively. These models mentioned above use the same convolutional networks with the input shape of 32x9x6000.

Table 4.6 shows the comparison of results on different types of SRNNs with attention and non-attention models and one ARNN. Compared with ASRNN-6MFB, the ASRNN achieves the same improvements of +0.9% on both databases. This means that the proposed system with nine channels may extract more information from speech than ASRNN-6MFB. Compared with ARNN, ASRNN achieves +1.3% and +0.5% absolute improvements on the IEMOCAP and MSP-IMPROV databases, respectively. This means that the segment-based attention model is better than the frame-based attention model. Compared with SRNN-Max-pooling and SRNN-Mean-pooling, the ASRNN achieves +0.9% and +1.5% absolute improvements on the IEMOCAP and MSP-IMPROV databases, respectively. This means that the attention model is better than max- and mean-pooling.

4.5 Listening test for temporal attention

Recently, Kell et al. [80] demonstrated that a deep neural network made human-like error patterns. If our attention model reflects the human mechanism, its result should be similar to human behaviors when they recognize speech emotion. For this reason, listening testing is designed to evaluate the similarity of the behaviors between the proposed attention model and humans. Thirty sentences from IEMOCAP database are used for the listening tests. Each sentence with a duration between 4.5 to 7.5 s is presented to at least 25 listeners (14 female and 11 male with ages ranging from 20 to 28) in random orders. Figure 4.6 lists the subjective evaluation test interface of the temporal attention behavior of the human auditory system, which requires the listener to focus on each utterance and select the two positions that can best express the most salient emotion. After selecting two attention positions, clicking the submit button will randomly play the next utterance. The listeners are asked to concentrate on listening to each utterance and choose the two locations that best show the emotions of the utterance.


Subject List s1

First Timer 0.89563

Second Timer 3.6035

Figure 4.6: The subjective evaluation test interface of temporal attention behavior in human auditory system

Figure 4.7 illustrates an example of comparisons between the attention model and human temporal attention. The top panel shows the waveform of an emotional sentence, and the upper-middle panel shows the spectrogram of the sentence. The lower middle panel shows the attention weights (α_i) that are calculated based on auditory front-ends and deep frameworks. The bottom panel shows a histogram that is the point numbers of attention position given by subjective judgments, and a dashed line that is the moving-average on two neighbor data points. One can see that the curve of the attention weights is similar to that of subjective judgment. Pearson's correlation coefficient is used to quantitatively measure the similarity between the attention model and human temporal attention. The correlation coefficient is $P = 0.552$ $(\rho < 0.001)$ between the attention weights and histogram in this particular utterance. If we calculate the correlation between the moving average values and the attention weights, the correlation coefficient becomes $P = 0.715$ $(\rho < 0.001)$. This indicates that there is a strong correlation between human temporal attention and the attention model. This implies that the proposed attention model can reflect human selective attention to a large extent.

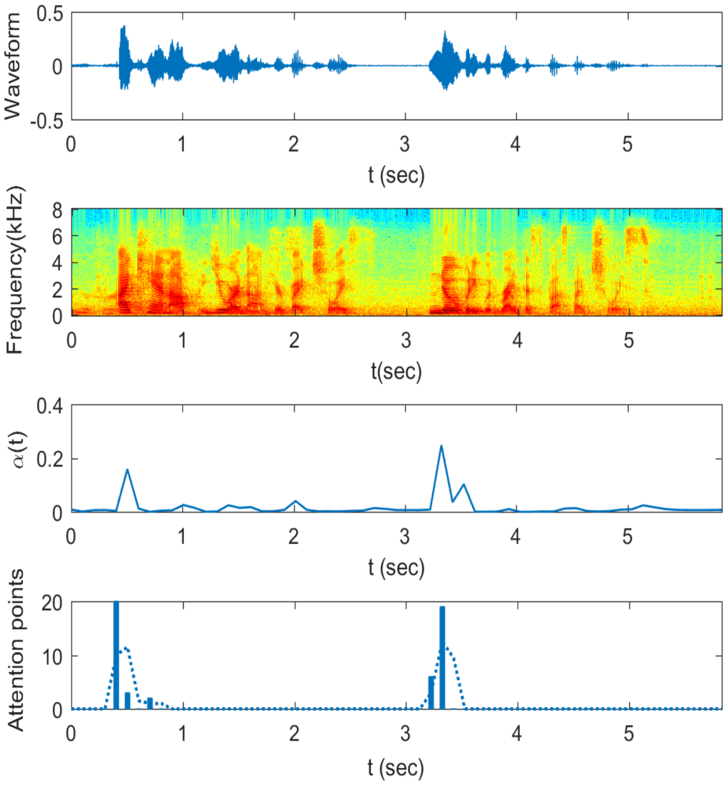


Figure 4.7: Analysis and comparison of attention model and human selective attention for test example. Top panel: raw waveform (Ses01F_impro04_F033.wav from IEMOCAP database); upper middle panel: spectrogram; lower middle panel: attention weight $(\boldsymbol{\alpha}_i)$ over sliding window time sequence; bottom panel: histogram shows attention numbers for subjective judgments, and dashed line shows moving-average with 2 data points.

4.6 General discussion

Taking into account that the human auditory system has a strong ability to perceive the intensity and fundamental frequency of speech, furthermore, it can track the temporal dynamics of emotion from the perceived information and focus on the salient emotion regions. Therefore, we propose an SER system by combining the auditory mechanism and attention mechanism of the auditory system.

The auditory front-ends of the SER system are used to produce temporal modulation cues, which contain local features and periodicity information of the emotional speech. During the process of temporal modulation cues extraction, an additional correlation in neighboring channels will be introduced because of the partially overlapped frequency. Traditional methods use discrete cosine transform to de-correlate the temporal modulation features in the acoustic and modulation frequency domains. Since CNN can successfully de-correlate the features in neighboring channels, we directly use 3D CNN to learn a joint spectral-temporal feature from temporal modulation cues. Furthermore, temporal dynamic information is obtained by continuously scanning the temporal sequence and then is transmitted to the higher-level processing center. To focus on the emotional regions while ignore the emotionless regions, an attention model is used to extract utterance-level features.

Table 4.7: Accuracy comparison of the proposed system and other systems on IEMOCAP and MSP-IMPROV databases (%)

Literature	Features	Backend	UA	
			IEMOCAP	MSP-IMPROV
Ref [81]	Raw speech	CRNN	60.23	52.43
Ref [55]	Log Mel-filterbank	Attentive CNN	59.54	45.76
Ref [82]	Mel-filterbank	CNN	61.8	53.8
Ref [83]	LLDs	Deep belief network	62.4	-
Ref [70]	FFT bins	BLSTM	52.8	-
Ref [84]	LLDs	Attention-based BLSTM	60.1	-
Proposed in chapter 3	Temporal modulation	3D CRNN	60.93	-
Proposed	**Temporal modulation**	**ASRNN**	**62.6**	**55.7**

To show the benefit of the proposed model, we compare our results with the studies [55, 81,82], the authors used the raw speech as input to parallel convolution layer and showed that on both databases as presented in Table 4.7. In [81], the authors used Mel filterbank features as the input to CNN and showed that CNN with these features could produce competitive results to the popular feature sets. In [55], the authors used Log-Mel filterbank features as the input to autoencoder and used attentive CNN for representation learning. In [82], the authors used the raw speech as input to parallel convolution layer and showed that

CNN-LSTM could capture multi-temporal dependencies. Compared to these studies, we are achieving a better result of 62.6% and 55.7% respectively on both databases using 3D convolutions and ASRNNs from temporal modulation cues. This indicates that the auditory front-ends can provide spectral-temporal representations, and deep frameworks can effectively extract emotional information from such representation for emotion recognition.

In addition, four representative studies with reported results on IEMOCAP are selected as comparisons. In [83], the authors used static features of LLDs for representation learning and deep belief network for emotion recognition. In [70], the authors used FFT bins with autoencoder for representation learning and used RNN to identify the emotional states. In [84], the authors used attention-based BLSTM models on LLDs for emotion recognition. Additionally, compared with our previous study in chapter 3, we are able to obtain faster training speed with SRNNs, and this system can better identify happiness and anger. This may be benefited by the 9-channel modulation filterbanks that contain fundamental frequency information, which is important for emotions. In contrast, our study exceeded the accuracy compared to the leading studies.

Other studies used attention models to identify emotions on IEMOCAP databases, but the experimental conditions are different. For example, [25,29,30] did not merge happy and excited into one class, while [71] just reported weighted accuracy. Unlike these frame-based attention models, we use a sliding window-based attention model to focus on the salient regions of emotion representation. The results of the experiments showed that this model could effectively obtain emotional information. The subjective evaluation shows that the attention patterns of the attention model are basically consistent with human behaviors in recognizing emotions.

4.7 Summary

We proposed an SER method using 3D convolutions and attention-based sliding recurrent neural network based on auditory front-ends. As the human auditory system is powerful in spectral-temporal signal analysis and processing, and the auditory model, which mimics the function of the human auditory system, is used as a front-end to extract spectral-temporal features in the SER system. Additionally, compared with modulation spectral features, these 3D features contain temporal dynamics characteristics and can avoid the modulation correlation problem.

Considering that local features and periodicity information can better express emotions, we used 3D convolutions to extract frame-level features from nine modulation filters. We then used recurrent networks to obtain temporal dynamics information in each utterance. We also used an attention model to focus on the emotionally salient parts of a speech signal.

Therefore, we propose a joint deep learning model that combines 3D convolutions and attention-based sliding recurrent neural networks. Our experiments demonstrated that the proposed system could obtain spectral-temporal representations and exhibit better recognition accuracy compared to that of state-of-the-art SER systems on both databases.

In summary, an auditory model as a front-end can extract rich spectral-temporal information, and the proposed method can effectively extract high-level features for emotion recognition. This system is possibly applied to other audio-event perception and recognition. For future work, we further plan to investigate the feature extraction from temporal modulation cues and emotion recognition model for dimensional speech emotional databases.

Chapter 5
Multi-resolution modulation-filtered cochleagram features for dimensional emotion recognition

5.1 Introduction

To recognize the dimensional emotion continuously from speech, the first step is to extract sequential acoustic features that can represent discriminative characteristics of each short-term segment. The sequential acoustic features from the speech can be extracted directly from sequential LLDs, and can also be extracted from the statistical features of LLDs calculated on a block of continuous frames. For dimensional emotion recognition, temporal dynamic information is very useful because the target dimensional values are continuous and have a short time gap between two adjacent predictions [37]. For example, it is usually difficult to distinguish between happy and angry, but there are obvious differences between them by mapping them into the V-A space. Both emotions have high arousal, but their valence is completely different: happiness has positive valence, and anger has negative valence. Unlike the arousal, which can usually be characterized by the amplitude envelope (energy) of the signal, its valence needs to be characterized by temporal dynamics of the amplitude envelope [87]. However, as the LLDs-based and functional-based acoustic features are not good at capturing the temporal dynamics for this task, especially for the suprasegmental information of emotional speech, valence prediction performances are commonly lower.

Moreover, the MSFs cannot reflect the real emotion in speech well since they are static features and do not contain detailed temporal cues. Temporal modulation cues contain multi-dimensional modulation spectral representations (MSR) of speech after using signal processing of auditory front ends. In Chapter 3 and Chapter 4, we proposed different CNNs to extract high-level emotional features from MSR for categorical emotion recognition. The MSR is a kind of 3D spectral-temporal representation, which is a mapping of speech signals to a high-dimensional data space through auditory and modulation filtering. High-dimensional data increases the complexity of the emotion recognition model, especially for the lack of large-scale speech emotion database, which may make the training model poor generalization. Avila et al. [26] proposed a feature pooling scheme to improve robustness for dimensional emotion recognition using combined MSF and MSR. Firstly, a 23-channel Gammatone filterbank was used to carry out auditory filtering and extract its amplitude envelope, then discrete Fourier transform (DFT) was used to obtain modulation spectrum, and then 8-channel modulation filter was used to generate MSR. Feature pooling is achieved by sliding window analysis of the fused MSF and MSR features and dimension reduction with Principal Component Analysis (PCA). However, this method uses DFT to convert the envelope signal into the frequency domain before temporal modulation, thus increasing the computational complexity.

Recent studies in cognitive neuroscience show that the cortical encoding of natural sounds entails the formation of multiple representations of sound spectrograms with different degrees of spectral and temporal resolution [14]. Chen et al. [88] proposed a multi-resolution cochleagram (MRCG) feature for speech separation, which is extracted from four cochleagrams of different resolutions to capture both local information and spectral-temporal context. Experimental results showed that the multi-resolution feature obtains the best results for speech separation among all evaluated features. The cortex derives these multi-resolution representations through frequency-specific neural processing channels and the combined analysis of the spectral and temporal modulations [89]. Inspired by this knowledge, we investigate multi-resolution temporal modulation cues extracted from frequency-specific auditory-filtering signals for emotion recognition. Then, we propose a novel auditory-based feature called multi-resolution modulation-filtered cochleagram (MMCG), which encodes modulation spectral representation of temporal envelope by the multi-resolution ways. Inspired from the feature extraction of MRCG, we also combine four modulation-filtered cochleagrams at different resolutions to construct the MMCG features for capturing the local and global temporal modulation cues as well as spectral-temporal modulation cues at different scales.

Next, a regression model should be considered to capture the temporal dynamics of emotion from MMCG feature sequences in dimensional emotion recognition. LSTM networks are widely used to model time sequences in learning more effective emotional representations from speech [37,90]. Comparing to Support Vector Regression (SVR), LSTM networks achieve a higher prediction accuracy due to their ability to model long-term time dependencies [91]. As each kind of modulation-filtered cochleagram contains different temporal modulation or contextual information, a parallel LSTM network architecture is designed to capture more temporal dynamics from different resolution modulation-filtered cochleagram.

The remainder of this chapter is organized as follows. Section 5.2 introduces the acoustic-based and auditory-based baseline features. Section 5.3 proposes the MMCG feature from temporal modulation cues. Section 5.4 presents the time series modeling using plain and parallel LSTM network architecture to capture temporal dynamics from the multi-resolution feature. Section 5.5 describes our experiment and compare it with the state-of-the-art results. Finally, we discuss the dynamic fitting ability of the MMCG feature and the effectiveness of each resolution modulation cochleagram feature in Section 5.6 and conclude this chapter in Section 5.7.

5.2 Baseline features

To evaluate the suitability of the proposed features for emotion recognition, we utilize various widely used acoustic features and auditory-based features as baselines. There are two methods to extract acoustic features either by the 20–40 ms frame length LLDs features with 10 ms shift (LLDs-based strategy), or by the statistical features of LLDs calculated on a block of continuous frames (functional-based strategy).

5.2.1 Acoustic-based feature

Since the value of each primitive is not labeled on one frame but rather on consecutive frames, such as four frames in RECOLA and ten frames in SEWA, we use frame stacking to extract the LLD-based features and match the granularity of the annotation in each primitive. Frame stacking consists of concatenating a block of continuous frames. For example, a context of 3 frames means that frames at times t - 1, t and t + 1 are concatenated to create one feature vector at time t. Frame stacking allows a recurrent model to use contextual information when learning a prediction function. The functional-based features are calculated on the LLDs-based features by using functions such as mean and standard deviation.

In this chapter, we use MFCC and eGeMAPS feature sets as the acoustic features, each using the two abovementioned strategies. We also use auditory-based features based on different stages of the auditory system, including early-stage MRCG features and late-stage MSF features. The following is a brief description of these two kinds of feature sets.

LLDs-based MFCC: This acoustic feature set contains 39 MFCCs (12 MFCCs + logarithmic energy, 13 delta and 13 double delta features), with a window size of 25 ms and a shift of 10 ms. In RECOLA, we stack four frames to form a 40-ms feature vector and obtain a total of 156-dimensional MFCC LLD features. Similarly, we stack 10 frames in SEWA to form a 100-ms feature vector and obtain 390-dimensional MFCC LLD features.

Functional-based MFCC: We then applied statistical functionals (mean and standard deviation) to extract 78-dimensional functional-based MFCC features by computing from the LLDs over segments of 4 seconds with a shift of 40 ms and 100 ms for both databases respectively. The same way is used in [92–94].

LLDs-based eGeMAPS: The eGeMAPS contains spectral, cepstral, prosodic and voice quality information of the voice record. Such features have been used in the RECOLA baseline with other modalities. Similar to extract MFCC features, we extract 23-dimensional acoustic LLDs acoustic features with the same window and shift. Finally, we get 92 and 230 eGeMAPS LLDs features in both databases.

Functional-based eGeMAPS: We then applied statistical functionals to extract 88-dimensional functional-based eGeMAPS features by computing from the LLDs over segments of 4 seconds with a shift of 40 ms and 100 ms for both databases respectively. Altogether, we get 88 statistical features for both databases.

5.2.2 Auditory-based feature

Modulation spectral feature (MSF): This feature set contains seven statistical features extracted from modulation spectral representations with a 200-ms window and a 40-ms or 100-ms shift for both databases respectively, including the mean of energy, centroid, flatness, spectral spread, spectral skewness, spectral kurtosis, and spectral tilt. This feature set is calculated in 32 acoustic channels and 9 modulation channels, respectively. Finally, we obtain 63 acoustic-frequency-domain features and 224 modulation-frequency-domain features, altogether 287 features.

Multi-resolution cochleagram feature (MRCG): We use the method proposed by [88] to extract MRCG features. This feature set contains four cochleagram features generated at different levels of resolution. The cochleagram is generated by applying the Gammatone filter to audio signals [95]. The high-resolution level encodes local information while the remaining three lower resolution levels capture spectral-temporal information. Different from the study, we use a Gammatone filterbank of 32 instead of 64 channels. Finally, 128 MRCG features are extracted from each time-frequency unit.

5.3 Multi-resolution modulation-filtered cochleagram features

5.3.1 Feature description

In this section, we propose the MMCG feature to encode temporal modulation cues of a temporal envelope to produce multi-resolution spectral-temporal features.

1) Temporal modulation cues from auditory front-ends

In this chapter, a Gammatone filterbank is applied to obtain multiple high-resolution sub-band signals in the frequency domain [96][58]. The temporal amplitude envelope $s_{\mathrm{e}}(n,t)$ is extracted using Hilbert transform to calculate the instantaneous amplitude of the *n*-th channel signal. Furthermore, the *m*-th modulation filter in the *n*-th channel envelope signal is used to obtain the spectral-temporal modulation signal $s_{\mathrm{m}}(n,m,t)$.

$$s_{\mathrm{m}}(n,m,t) = \ m_{\mathrm{f}}(m,t) * s_{\mathrm{e}}(n,t), \qquad 1 \le m \le M, \tag{5.1}$$

where $m_{\mathrm{f}}(m,t)$ is an impulse response of the modulation filterbank and M is the channel number of modulation filterbanks. In this chapter, M is set to nine, which spans from 2 to 512 Hz on a logarithm scale. Such modulation frequency components contain rich spectral-temporal information to describe the variations of intensity, duration and period of speech [72].

Then, each sub-band modulation signal is divided into a number of different-duration modulation units, where the shift length is the same as the dimensional emotion database. It is defined as:

$$s_{mu}(n,m,i) = \ \mathrm{w}(t_w) \cdot s_m(n,m,(i-1) \cdot Len_s + t_w), \tag{5.2}$$

where $\mathrm{w}(t_w)$ is a window function, t_w is the sample number in each time window. The Hamming window is chosen in this chapter. $s_{mu}(n,m,i)$ is the sub-band modulation unit of *n*-th acoustic channel and *m*-th modulation channel at the *i*-th modulation unit, $1 \le i \le \frac{Len_t}{Len_s}$, where Len_t is the total length of speech signal $s(t)$ and Len_s is a window shift. Finally, a total of $n * m$ channel signals are generated. $s_{mu}(n,m,i)$ is expressed as $s_{mu}(c,i)$, where c equals to $n * m$.

The modulation-filtered cochleagram $MCG(c,i)$ is calculated by convolving each modulation unit:

$$MCG(c,i) = \sum_{i=0}^{L-1} s_{mu}(c,i) * s_{mu}(c,i), \tag{5.3}$$

where L equals to $\frac{Len_t}{Len_s}$.

1) Multi-resolution modulation-filtered cochleagram features

Each modulation unit in the modulation-filtered cochleagram contains temporal modulation cues. Recent psychoacoustic investigations indicate that humans are able to detect and discriminate multi-scale modulations that occur in one dimension alone (temporal or spectral) as well as combined spectral-temporal modulations [89]. To obtain multi-scale information, we extract multi-resolution temporal modulation cues from modulation units. The extraction of the proposed MMCG feature is given in Fig. 5.1.

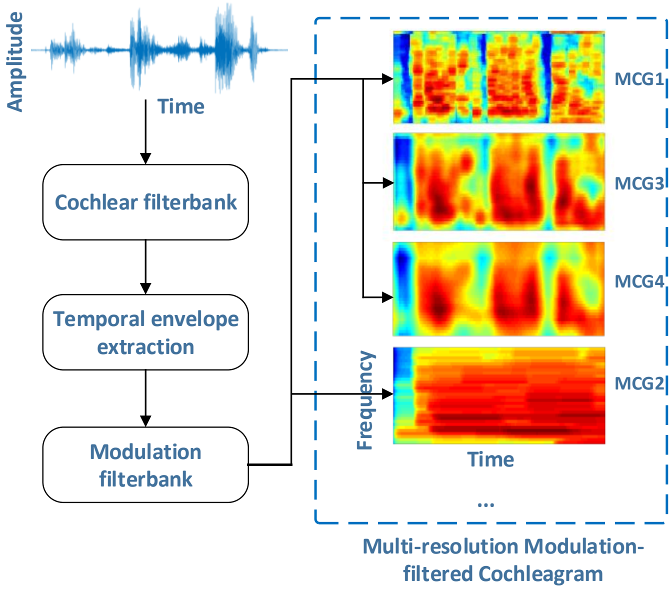


Figure 5.1: Extraction of MMCG feature

The first modulation-filtered cochleagram (MCG1) generates a high-resolution cochleagram feature from the modulation units, and each modulation unit performs discrete convolution with itself. As we all know, human auditory nonlinearity expands small sounds and compresses large sounds. As auditory frequency selectivity in a log frequency scale for amplitude modulation best matches the auditory perception of modulation frequency. A log function is used to match the auditory perception in the MCG1. The window length of MCG1 is set to 200 ms, and the mathematical expression of MCG1 is:

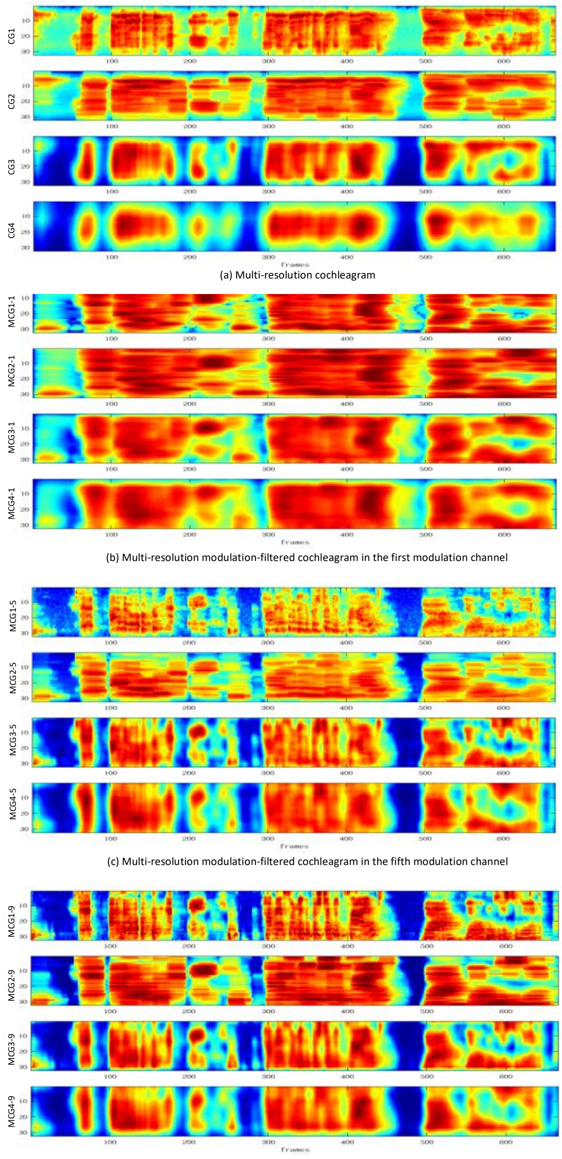
$$MCG1(c,i) = log10(\sum_{i=0}^{L-1} s_{mu}(c,i) * s_{mu}(c,i)), \qquad (5.4)$$

Similarly, the second modulation-filtered cochleagram (MCG2) can be obtained. Unlike MCG1, the window length changes to 2000 ms, which is a low-resolution cochleagram feature.

The third modulation-filtered cochleagram (MCG3) is derived by averaging MCG1 across a square window of 5 frequency channels and 5 time steps centered at a given modulation unit. If the window goes beyond the given cochleagram, the outside units take the value of zero (i.e., zero padding). It can be expressed as:

$$MCG3(c,i) = (\sum_{\mathrm{n}=c-2}^{c+2} \sum_{j=i-2}^{i+2} MCG1(c,i))/(5*5), \qquad (5.5)$$

The fourth modulation-filtered cochleagram (MCG4) is calculated in a similar way to MCG3, except that a square window of 11 frequency channels and 11 time steps is used.



(a) Multi-resolution cochleagram

(b) Multi-resolution modulation-filtered cochleagram in the first modulation channel

(c) Multi-resolution modulation-filtered cochleagram in the fifth modulation channel

(d) Multi-resolution modulation-filtered cochleagram in the ninth modulation channel

Figure 5.2: The multi-resolution cochleagram and modulation-filtered cochleagram. (a) The multi-resolution cochleagram contains four different scale features (CG1-CG4). (b)-(d) The 1st, 5th, 9th modulation-channel multi-resolution modulation-filtered cochleagram, where containing MCG1-MCG4 for each figure. The x-axis represents the number of time step and the y-axis represents the channel number.

It can be shown as:

$$MCG4(c,i) = (\sum_{k=c-5}^{c+5} \sum_{j=i-5}^{i+5} MCG1(c,i))/(11 * 11), \qquad (5.6)$$

MCG1, MCG2, MCG3, and MCG4 are connected to obtain the MMCG feature, which has 288 × 4 dimensions for each time step. The MMCG feature is denoted as:

$$MMCG(c,i) = [MCG1(c,i); MCG2(c,i); MCG3(c,i); MCG4(c,i)]. \qquad (5.7)$$

5.3.2 Feature analysis

Figure 5.2 shows an MRCG and three different MMCG with 40-ms window length and shift. The MRCG feature contains four different scale cochleagrams (CG1-CG4) as shown in Fig. 5.2(a), whereas the MMCG feature contains the 1st, 5th, 9th modulation channel as shown in Fig.5.2 (b)-(d), respectively. The MMCG features are obtained by the modulation of a temporal envelope on different frequency-specific channels. Each modulation channel contains multi-resolution features. One can see that the MMCG features have higher time and frequency-domain resolution structure of speech than the MRCG. Therefore, it is expected that the MMCG is more capable of capturing the temporal dynamics of speech emotion cues.

5.4 Time series modeling

In this section, we introduce emotion recognition models for dimensional emotion recognition. A plain LSTM network architecture is first used as the baseline method, and then a parallel LSTM network architecture is designed to extract spectral and temporal features from the MMCG feature. The details of the abovementioned network architectures are elaborated in the following part.

5.4.1 Plain long short-term memory network

An LSTM architecture is the state-of-art model for sequence analysis since it can exploit long-term dependencies in the sequences by using memory cells to store information. As the LSTM network is widely used to model time sequences in learning more effective emotional representations from speech [97]. Trigeorgis et al. proposed to use 1D convolution to directly learn high-level emotion feature representation directly on the raw waveform, and then use LSTM to learn its time-dependent features from this representation sequence to predict dimensional emotion [98]. Wöllmer et al. presented a fully automatic audiovisual recognition approach based on LSTM modeling of word-level audio and visual features [99]. For dimensional emotion recognition, temporal information is very useful because the target dimensional values are continuous and have a short time gap between two adjacent predictions. In this chapter, we use LSTM networks as a regression model to predict continuous variations of the dimensional variables and explore different types of architectures to capture the temporal dynamics in speech.

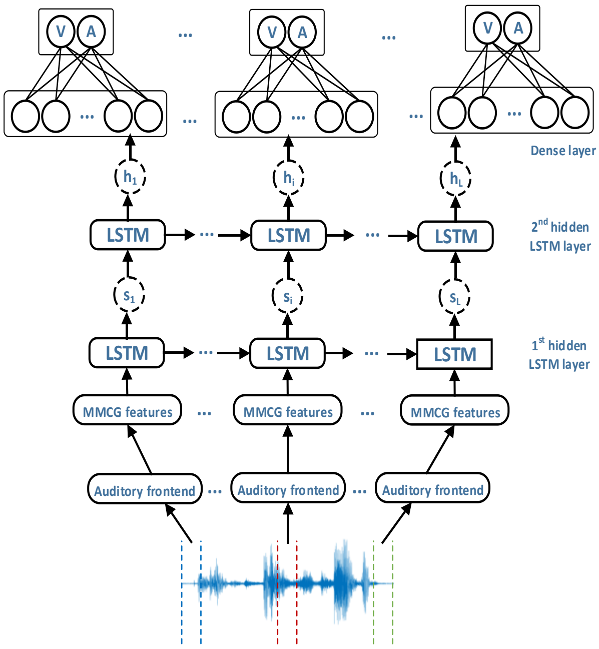


Figure 5.3: A plain LSTM network architecture for dimensional emotion recognition

First, a plain LSTM network architecture is used as the benchmark regression model. Figure 5.3 depicts the architecture of the plain LSTM network for dimensional emotion recognition. The plain LSTM network architecture contains one input layer with 1152-dimensional MMCG features, two hidden layers, and then following a dense layer, a regression layer. In our experiments, the model consists of 128 and 64 nodes for the first and second hidden layers. Then the dense layer is used to connect the hidden layer, followed by a ReLU activation function. Finally, we use the regression layer to predict the value of arousal and valence emotion. To avoid overfitting when training our networks, we use a dropout rate of 0.75 before the regression layer during training.

5.4.2 Parallel long short-term memory network

MMCG contains different temporal and contextual modulation cues. Each modulation-filtered cochleagram feature has its own temporal and contextual dependency, which can be obtained by different LSTM units. So, in our solution, several LSTM units are used in parallel to handle different-resolution temporal modulation cues. Because the plain LSTM can't extract the dependency of different scales, we propose a parallel LSTM network as a regression model to capture temporal and contextual information for dimensional emotion recognition.

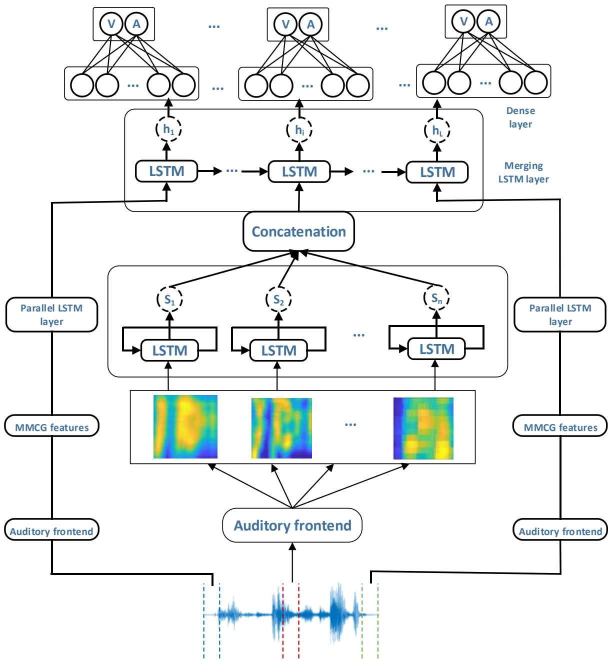


Figure 5.4: A parallel LSTM network architecture for dimensional emotion recognition

Figure 5.4 depicts the architecture of the parallel LSTM network. It consists of an input layer, a parallel LSTM layer, the merging LSTM layer, the dense layer, and the regression layer. In the input layer, the four 288-dimensional features (MCG1-MCG4) are fed into different LSTM units for parallel processing. Every LSTM unit has the same hyper-parameters, such as neurons number, to handle each feature equally. The LSTM unit deals with the segment data step by step and iterates through the loop. The outputs s_i at the time step i of each LSTM unit are collected and combined as the preliminary features with a size of 4* h1, where h1 is the number of the hidden neurons of each LSTM unit. After that, a merging LSTM layer is added. It is fed with the preliminary features and iterates through the order of the outputs from the previous layer. The output h_i of the merging LSTM layer is selected as the frame-level emotional feature, which contains the time dependency. It is a vector with the size of h2, where h2 represents the number of the hidden neurons of the merging LSTM unit. Then the dense layer is used to connect the hidden layer, followed by a ReLU activation function. To avoid overfitting when training our networks, we use a dropout rate of 0.75 before the regression layer during training.

In our experiments, the model consists of 128 and 64 nodes for the first and second hidden layers. Finally, the regression layer to the predicted valence and arousal primitives.

5.4.3 Loss function

Because CCC combines PCC and MSE, it is more reliable in evaluating performance, so CCC is used as an evaluation indicator in dimensional emotion recognition tasks. As the CCC loss consistently improves results in the emotion recognition task compared to mean squared error loss and mean absolute error loss [45]. We utilize a CCC-based loss function (L_c) as the objective function of recurrent model. L_c is defined as:

$$L_c = \frac{2-\rho_c^a-\rho_c^v}{2}, \tag{5.8}$$

where ρ_c^a and ρ_c^v are the CCC of the arousal and valence, respectively.

Table 5.1: Performance comparison (in term of ρ_c) under different features using plain LSTM networks on RECOLA database

Features	Arousal	Valence
LLD-based MFCC	.679	.320
functional-based MFCC	.651	.331
LLD-based eGeMAPS	.662	.312
functional-based eGeMAPS	.701	.329
functional-based MSF	.709	**.368**
MRCG	**.734**	.351

5.4.4 Multitask learning

In the V-A space, we first investigate whether there is a correlation between valence and arousal. If the two are strongly correlated, multitask learning can be used to train the regression model at the same time. Otherwise, the respective models should be trained independently. We use PCC to measure the correlations between arousal and valence. The PCC coefficient is $\rho = 0.518$ on the train and development set in RECOLA, while $\rho = 0.652$ in SEWA. This indicates that there is highly correlated between arousal and valence. Therefore, we employ a multi-task learning method to predict the arousal and valence simultaneously on two kinds of LSTM networks. The regression models of each LSTM network are trained with two outputs and two CCC losses at the same time.

5.5 Experiment results and analysis

In this section, we employ the plain LSTM network architecture to compare performances between the baseline features and the proposed MMCG. Then we conduct extensive experiments to improve prediction performance using the parallel LSTM network architecture.

5.5.1 Experiment setup

The RECOLA database is used to find the best emotion recognition model in valence and arousal space, and then a subset of the SEWA database is used to validate our proposed method. We implement our methods with the TensorFlow deep learning framework. We train the regression model throughout all experiments with Adam optimizer with a fixed learning rate of 1e-4. Additionally, for all random weight initializations, we choose L2-regularizer initialization. For RECOLA, the mini-batch size utilized is 10 with a sequence length of 750 frames (30 s) when training, and the model is tested on the entire records without segmentation. After we preprocess the raw signal to have zero mean and unit variance, we segment it to 30-second sequences and use them as input. For SEWA, due to the variable length in this database, we train the deep model using zero-padding and test it directly using the original data. The mini-batch size utilized is 10 with a sequence length of 880 frames (almost 90 s) when training, and the model is tested on the entire records without segmentation.

5.5.2 Results of baseline features

In this set of experiments, we aim to investigate the baseline features (MFCC, eGeMAPS, MSF and MRCG) for emotion dimension prediction. Among them, the acoustic features (MFCC, eGeMAPS) include LLDs-based and functional-based features. The LLDs-based features are extracted from openSMILE toolkit [48]. Feature normalization is then applied to each feature dimension. First, we apply the plain LSTM network to the baseline features to find the dominant feature for predicting each emotion dimension. The prediction results using each set of features are reported in Table 5.1. For acoustic-based features, the highest CCC is achieved using functional-based eGeMAPS features in arousal prediction (ρ_c^a = 0.701), whereas the highest CCC is achieved using LLD-based MFCC in valence prediction (ρ_c^v = 0.331). In addition, the functional-based features outperform the LLD-based features in valence prediction, which is consistent with the knowledge that valence is more related to long-term temporal information. For auditory-based features, the highest CCC is achieved using MRCG features in arousal prediction (ρ_c^a = 0.734), whereas the highest CCC is achieved using MSF in valence prediction (ρ_c^v = 0.368). As we can see from the table that auditory-based features achieve better performance than acoustic-based features on the dimension prediction. The results indicate that auditory-based features perform better than acoustic-based features. It is worth mentioning that this is the first report on the MRCG for emotion recognition. The MRCG features are designed for speech separation [88] and recently applied for voice activity detection [100] and attitude recognition [101]. The experiment results show that the MRCG features perform better than the traditional acoustic features.

Table 5.2: Performance comparison (in term of ρ_c) of MRCG and MMCG under different window size and dynamic feature using plain LSTM networks on RECOLA database

Features	Window size	Dynamic feature	Arousal	Valence
MRCG	short	Non-delta	.734	.351
	short	delta	.744	**.423**
	long	Non-delta	.717	.306
	long	delta	**.749**	.322
MMCG	short	Non-delta	.742	.302
	short	delta	.766	.413
	long	Non-delta	.712	.349
	long	delta	**.768**	**.431**

5.5.3 Results of proposed features on RECOLA database

MMCG can capture both temporal information and spectral-temporal contexts at different scales. To investigate the effectiveness of temporal window size and dynamic delta features on emotion recognition in MMCG features, we still use the plain LSTM to train the emotion prediction model under different window sizes and dynamic features. In addition, we use the same strategy for MRCG features and compare the emotion recognition performance with that of MMCG features.

Short-time VS. Long-time features: When predicting dimensional emotion primitives continuously, we still do not know what is the best temporal window size for capturing the salient features [102]. In this chapter, we investigated two kinds of MMCG features: one is the short-time MMCG feature (MMCG_short), and the other is the long-time MMCG feature (MMCG_long). In the short-time MMCG feature, the window length of MCG1 is consistent with the dimension label length of the database (the window length of the RECOLA database is 40 ms, and the window length of the SEWA database is 100 ms), and the window length of MCG2 is 400 ms. In the long-time MMCG feature, the window length of MCG1 is 200 ms, and the window length of MCG2 is 2000 ms. The window shift of both MMCG features is consistent with the database dimension label length. The same parameter settings of MMCG are also used in the MRCG feature. Finally, short-time and long-time MRCG features (MRCG_short and MRCG_long) are obtained.

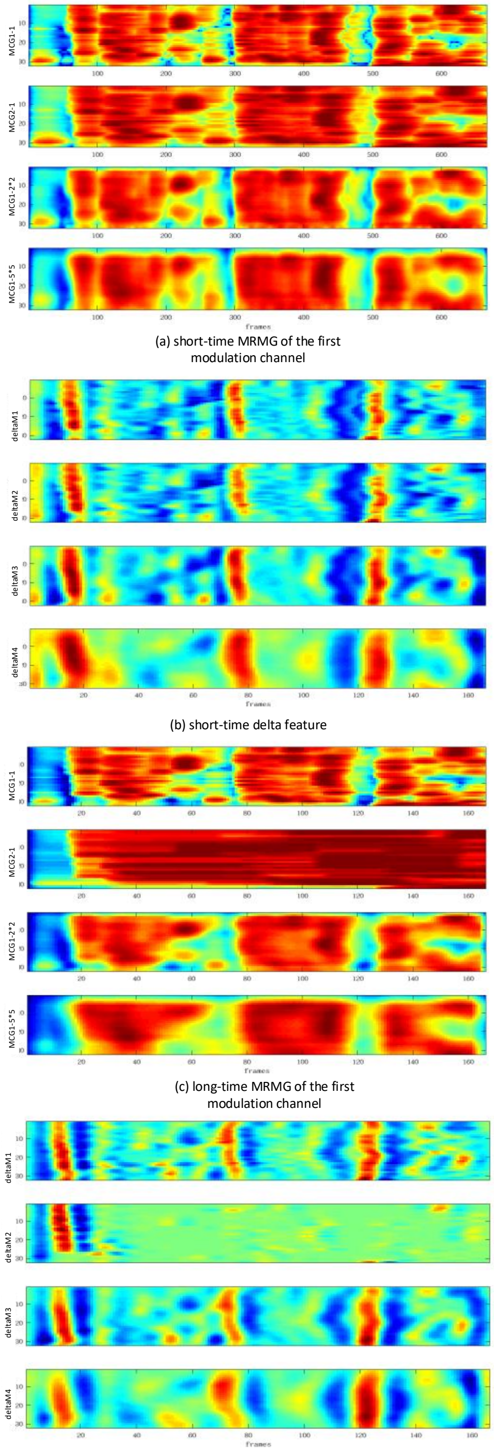
Delta VS. Non-delta features: In speech processing, delta and double-delta features are widely used to capture temporal dynamics. For example, MFCC with delta features gets better speech recognition results than MFCC alone. In this chapter, we also investigate the effectiveness of dynamic features of MMCG in improving the performance of the emotion prediction. The dynamic features of ΔMMCG is calculated as:

$$\Delta MMCG(n,m,i) = \frac{MMCG(n,m,i+1)-MMCG(n,m,i-1)+\left(MMCG(n,m,i+2)-MMCG(n,m,i-2)\right)}{2}. \quad (5.9)$$

Then 1152-dimension delta features are obtained from the formulas. The double-delta features are also obtained using a similar formula on ΔMMCG instead of MMCG. Lastly, we obtain a total of 3456 features from each frame. The prediction results of MRCG and MMCG under different window sizes and dynamic features are reported in Table 5.2.

Table 5.3: Performance comparison (in term of ρ_c) under different features using parallel LSTM networks on RECOLA database

	Arousal			Valence		
Features	raw	scaling	centering	raw	scaling	centering
MRCG-short	.703	.708	.781	.270	.303	.307
MRCG-long	.744	.751	.784	.256	.261	.323
MRCG-long +MSF	.753	.762	.830	.426	.463	.497
MMCG -short	.763	.777	.827	.417	.436	.451
MMCG -long	.778	.813	.824	.474	.494	.519
MMCG -long +MSF	**.812**	**.821**	**.865**	**.481**	**.502**	**.524**



(a) short-time MRMG of the first modulation channel

(b) short-time delta feature

(c) long-time MRMG of the first modulation channel

(d) long-time delta feature

Figure 5.5: The short-time and long-time MMCGs and its delta features. (a) and (c) show the short-time and long-time MMCGs (only the MMCG of the first modulation channel). (b) and (d) show the delta features corresponding to (a) and (b), respectively.

For MRCG features, the best arousal prediction result is obtained under long-time and delta feature conditions (ρ_c^a = 0.749), and the best valence prediction result is obtained under short-time and delta feature conditions (ρ_c^v = 0.423). For MMCG features, the best arousal and valence prediction results are obtained under long-time and delta feature conditions (ρ_c^a = 0.768, ρ_c^v = 0.431). Overall, this shows that MMCG has better results on both arousal and valence predictions than MRCG. However, we find that the recognition results of MMCG and MRCG in the long-time and short-time features are inconsistent. The short-time features of MRCG have higher CCC than the long-time features, but the long-time features have higher CCC than MMCG. This may be due to the fact that MMCG contains the temporal modulation clue of the envelope, and long-time modulation is more effective for emotion recognition than short-time modulation information.

Table 5.4: Performance comparison (in term of ρ_c) under different features and model on RECOLA dataset (upper four rows for performance comparison without delay compensation, lower nine rows for performance comparison with delay compensation)

Predictor	Features	Arousal	Valence
Zhang et al. [103]	eGeMAPS	.783	**.495**
Avila et al. [26]	MSFs	.795	.265
Proposed	MRCG	.753	.426
Proposed	MMCG	**.812**	.481
Brady et al. [104]	MFCC	.846	.450
Valstar et al. [45]	eGeMAPS	.796	.455
Zhang et al. [103]	eGeMAPS	.811	.519
Ouyang et al. [47]	eGeMAPS	.783	.467
Povolny et al. [105]	eGeMAPS	.832	.489
Le et al. [106]*	Log mel-filterbank	.855	.518
Proposed	MRCG	.830	.497
Proposed	MMCG	**.865**	**.524**

* Only list the results using the regression model and the CCC loss function in this study.

The delta feature gets better results on both the arousal and valence prediction than the non-delta feature, which shows that the delta feature reflecting the dynamic change of the signal can significantly improve the dimensional emotion recognition. However, the degree of CCC improvement is not the same under different conditions. The specific list is as follows: (i) The delta feature has a more significant improvement in valence prediction than arousal prediction. (ii) The delta feature has a more noticeable improvement under short-time conditions than under long-time conditions. As shown in Fig. 5.5, Fig 5.5(a) and 5.5(b) show the short-time and long-time MMCGs (only the MMCG of the first modulation channel). Fig. 5.5(c) and 5(d) show the delta features corresponding to Fig. 5.5(a) and 5.5(b), respectively. From the figures, one can see that the delta feature of the short-time MMCG contains more information reflecting the change of the modulation signal, so it can better reflect this feature is more conducive to identifying dimensional emotions than the long-time MMCG feature.

Overall, the MMCG feature performs significantly better than MRCG. Moreover, the MMCG features combining the original and dynamics with a long-time window can improve the prediction performance of dimensional emotion.

Each kind of modulation-filtered cochleagram in the MMCG contains different temporal modulation or spectral-temporal modulation information. Therefore, we use parallel LSTM to model the temporal dependencies of various resolution features (with delta and double-delta feature) from each resolution information. Table 5.3 reports the CCCs obtained by MRCG and MMCG using short-time and long-time parallel LSTM networks. In addition, the CCC obtained by combining these two features with MSF in the dense layer are also reported separately. The long-time MMCG features combined with MSF achieved the highest CCCs of 0.812 for arousal and 0.481 for valence, respectively. The upper four rows of Table 5.4 show a CCC comparison with the recent studies without delay compensation. It shows that the result of our approach is better results for arousal prediction, and get a comparable result for valence prediction. It is also worth mentioning that compared with the study [26] using the same auditory-based features, we got a relative improvement of 8% (0.795 to 0.812) for arousal prediction and a 29% improvement (0.265 to 0.481) for valence prediction.

Annotators often have reaction time delays to emotional cues when labeling consecutive emotions on a recording, which can cause a shift between the annotated emotion at a certain time step and its actual emotion. To compensate for delays in the ratings, some studies apply a chain of post-processing to the predictions obtained on the development set [23][45][92][98]. It includes: (i) median filtering (the window size ranging from 0.4 s to 20 s), (ii) time-shifting (by shifting the prediction forward in time with values ranging from 0.04 s to 10 s), (iii) scaling (using the ratio of standard-deviation of gold-standard and prediction as scaling factor) and (iv) centering (by computing the bias between gold-standard and prediction). In this chapter, we found that the median filtering and time-shifting cannot improve the predicting performance. This might be due to the reason that the MMCG features can represent stable temporal cues and the LSTM networks can effectively capture the temporal information. However, the scaling and centering are useful to improve the performance. Hence, we only report the post-processing results using scaling and centering ways.

Scaling the prediction primitives could help to attenuate some of the remaining noise. In this chapter, the scaling method with the standard-deviation ratio is adopted to scale the prediction primitives. A scaling output vector $y_{sacling}$ is calculated as:

$$y_{scaling} = \frac{\sigma_l}{\sigma_p} \otimes y \ , \tag{5.10}$$

where σ_p is the standard deviation of the predictions, σ_l is the standard deviation of the golden standard, $\otimes$ is the element-wise multiplication operation, and y is the prediction vector to be scaled.

The second post-processing method is centering, which is implemented by computing the bias between gold-standard and prediction. A centering output vector $y_{centering}$ is calculated as:

$$y_{centering} = \ y + \mu_l - \mu_p \tag{5.11}$$

where μ_l and μ_p are the mean values in the training labels and the prediction vector y, respectively.

Table 5.5: Performance comparison (in term of ρ_c) under different features on SEWA dataset (upper two rows for performance comparison using plain LSTM networks, lower six rows for performance comparison using parallel LSTM networks)

Features	Arousal	Valence
functional-based eGeMAPS	.396	.291
functional-based MFCC	.324	.310
MRCG-short	.338	.353
MRCG-long	.432	.464
MRCG-long +MSF	.507	.492
MMCG-short	.501	.483
MMCG-long	.523	.519
MMCG-long +MSF	**.572**	**.534**

The CCC results of MRCG and MMCG with scaling and centering are shown in Table. 5.3. The highest CCC is achieved by centering the raw MMCG, whose arousal and valence were 0.865 and 0.524, respectively. The lower nine rows of Table 5.4 show the comparison of evaluation methods CCC with delay compensation. The best performer of AVEC 2016, Brady et al. [104] used SVR trained on sparse-coded higher-level representations of various types of audio features. Many studies used eGeMAPS features to train different regression models to identify dimensional emotions. For example, Povolny et al. trained a set of linear regressors on eGeMAPS augmented with deep bottleneck features from deep neural network acoustic models [105]. Zhang et al. considered the difficulty of different data training models and the importance of contextual information for emotion recognition. They proposed Dynamic Difficulty Awareness Training combined with LSTM to predict dimensional emotion [103]. Le et al. proposed a discrete continuous emotion recognition based on Log Mel-filterbank coefficients. First, k-means is used to discretize all continuous labels, and then BLSTM is used for multi-task training and then decoded to make the model prediction continuous dimensional emotion [106]. Among the evaluation methods with delay compensation, the MMCG + parallel LSTM method achieves the best recognition effect on the degree of arousal and valence prediction. Secondly, compared with the MRCG feature, the arousal prediction is also relatively improved by 20% (from 0.830 to 0.865), and the valence prediction is relatively improved by 5% (from 0.497 to 0.524).

Table 5.6: Performance comparison (in term of ρ_c) under different features and models on SEWA database

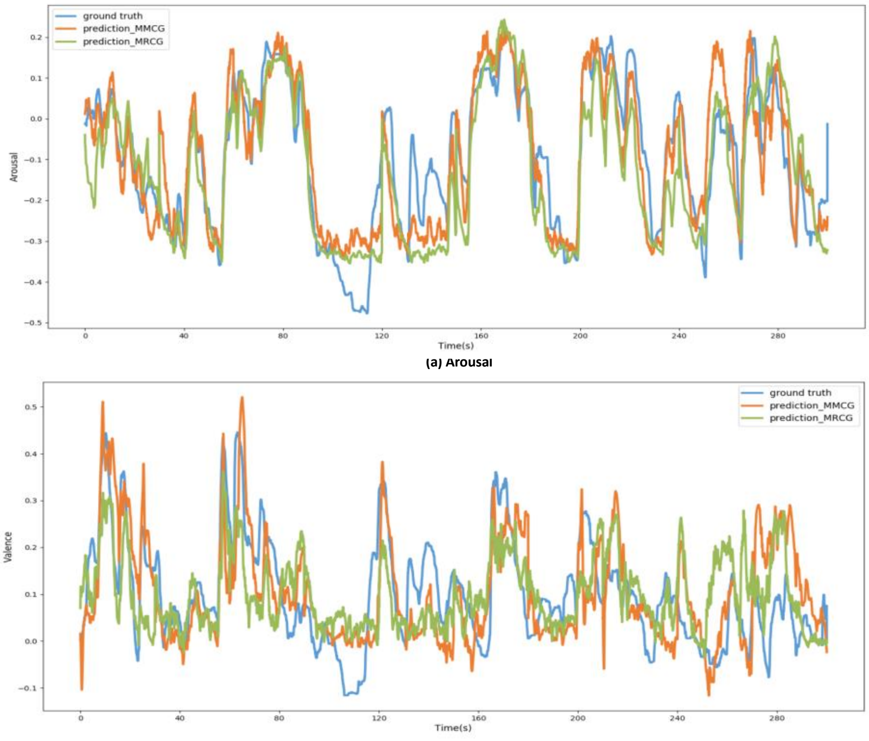
Predictor	Features	Arousal	Valence
Zhao et al[107] *	log-mel spectrogram	**.604**	.511
Schmitt et al [93] §	eGeMAPS	.586	.516
AVEC2017[46]	eGeMAPS	.344	.351
Schmitt et al [94] §	eGeMAPS	.571	.517
Han et al [108]	Is13_ComParE	.356	.396
Chen et al [109]	Is10_Paraling	.524	.504
Chen et al [109]	Soundnet	.527	.447
Ouyang et al [47]	eGeMAPS	.540	.502
Avila et al [26]	MSFs	.369	.308
Proposed	MRCG	.512	.493
Proposed	MMCG	.572	**.534**

* The results are obtained by removing the interlocutor’s speech segment by the official turn information. It is 0.479 for arousal and 0.447 for valence when containing both speaker’s and interlocutor’s audio signals.

§ The results are obtained by adding features the official turn information in the interlocutor’s speech segment.

5.5.4 Results of proposed features on SEWA database

In addition, we use a subset of the SEWA database to validate our proposed method. The upper two rows of Table 5.5 show the CCC (without considering feature compensation) obtained on the functional-based eGeMAPS and MFCC feature sets using the plain LSTM network. The remaining rows show the CCC obtained on the MRCG, MMCG, and combined features with MSF using parallel LSTM. The highest CCC obtained from the long-time combined MMCG feature is 0.572 for arousal and 0.534 for valence, respectively, which is consistent with the results obtained on RECOLA. Table 5.6 lists the results of state-of-the-art studies. In [107], the results are obtained by



(a) Arousal

(b) Valence

Figure 5.6: A prediction example of arousal (a) and valence (b) based on MRCG and MMCG features obtained for the subject P34 in RECOLA. The yellow and blue curves show the predicted and the ground truth for arousal and valence respectively.

removing the interlocutor’s speech segment by the official turn information. It is 0.479 for arousal and 0.447 for valence when containing both speaker’s and interlocutor’s audio signals. In [93], the results are obtained by adding features the official turn information in the interlocutor’s speech segment. The additional feature is added to the input feature sequence, derived from the speaker turn information, which was provided to all participants of AVEC 2017. For each timestamp, this feature is either 0 or 1, indicating whether the subject is audible or not. However, we used the mix of the target speaker and interlocutor to train the model without considering the turn information. It shows that the result of our approach is comparable to the studies [93,107] even under different experimental conditions, and has a significant improvement in valence prediction. In addition, as shown in Table 5.6, the results of arousal and valence prediction based on MMCG features are superior to the studies [46,47,93,94] based on eGeMAPS features, the study [108] based on IS13_ComParE features [51], and the study [109] based on Soundnet [110] and IS10_paraling features[50]. Moreover, compared with the study [26] using the same auditory-based features, we got a 33% improvement (0.369 to 0.572) for arousal prediction and a 32% improvement (0.308 to 0.534) for valence prediction.

5.6 General discussion

Experimental results show that multi-resolution modulation cochleagram features have achieved the best performance in dimensional emotion recognition. In this section, we will further discuss the effectiveness of each resolution modulation cochleagram feature in emotion recognition. Figure 5.6 shows an example of arousal and valence prediction based on MRCG and MMCG features (RECOLA database P42 record). The green curve represents the predicted sequence of arousal (Fig. 5.6 (a)) and valence (Fig. 5.6 (b)) of MRCG feature in the continuous speech signal, and the orange curve represents the arousal of MMCG feature, the blue curve represents the corresponding ground truth. It can be seen from the figure that the arousal prediction and ground truth curve fitting based on MRCG and MMCG features are very good (CCC is 0.84 and 0.88, respectively), and the MMCG feature can fit better than the MRCG feature. However, the results of valence prediction and ground truth curve fitting based on MRCG and MMCG features are relatively weak (CCC is 0.45 and 0.63, respectively), and MMCG features are significantly improved over MRCG features.

Table 5.7: Performance comparison (in term of ρ_c) under different resolution modulation-filtered cochleagram features on RECOLA dataset

Features (model)	Arousal	Valence
MCG1-Delta (plain LSTM)	.779	.348
MCG2-Delta (plain LSTM)	.761	.360
MCG3-Delta (plain LSTM)	.781	.359
MCG4-Delta (plain LSTM)	.791	.341
MMCG -Delta (plain LSTM)	.768	.431
MMCG -Delta (parallel LSTM)	.778	.474

The first four lines of Table 5.7 list the recognition results of the plain LSTM network on the four resolution features MCG1-MCG4. Both delta and double-delta features are included here. The experimental results show that each resolution modulation cochleagram feature has close to arousal and valence prediction. The fifth row lists the results obtained by the MMCG feature under the plain LSTM network (including delta and double delta features). The experimental results show that the MMCG feature combined by multi-resolution features has not improved arousal prediction, but the valence prediction has been significantly improved. The possible reason is that the modulation-filtered cochleagram at each resolution can characterize the amplitude envelope for arousal prediction, and the modulation-filtered cochleagram features of different resolutions contain the respective dynamic characteristics of the amplitude envelope, so that the multi-resolution features generated by the combination of features can be obtained more temporal dynamic information improves valence prediction ability. The sixth line lists the results obtained by the MMCG feature under the parallel LSTM network. The parallel method achieves higher results in both arousal and valence prediction. This shows that using multiple LSTMs in parallel can obtain different scale dependencies, thus obtain more temporal and contextual information than the non-parallel mode. This indicates that each modulation-filtered cochleagram contains various features from other modulation-filtered cochleagrams, and multi-resolution features generated by combined features are more helpful for valence prediction.

5.7 Summary

In this chapter, we proposed the multi-resolution modulation-filtered cochleagram (MMCG) features for predicting the valence and arousal emotional primitives. This feature is constructed by combining four modulation-filtered cochleagrams at different resolutions to capture various spectral and temporal features. In addition, a parallel LSTM network architecture is designed to extract more temporal dynamics from each resolution information of modulation-filtered cochleagram features. The experimental results show that the MMCG feature can significantly improve the performance of emotion recognition compared with the acoustic-based features and other auditory-based features, especially the performance improvement in valence prediction. In summary, the MMCG feature can effectively extract high-level features for emotion recognition. In addition, we plan to investigate this feature for other acoustic scene analysis, such as categorical emotion recognition, speech separation, voice activity detection, etc.

Chapter 6

Conclusion and future work

6.1 Summary

The purpose of this study is to explore auditory-based emotion features and deep learning methods to improve the performance in categorical and dimensional emotion recognition. To this end, the following works have been done to solve the challenges of speech emotion recognition.

This study first investigated the multi-channel acoustic frequency components obtained from auditory filterbank and temporal modulation cues obtained from modulation filterbank. As temporal modulation cues play an important role in speech perception and contain multi-dimensional spectral-temporal information, this study proposed a 3D CNN architecture to obtain discriminative auditory representations from the temporal modulation cues by joint spectral-temporal feature learning. The experimental results show that the joint spectral-temporal auditory representations can be extracted using 3D CNN from temporal modulation cues. The results demonstrate that the performance of emotion recognition based on joint spectral-temporal representation can exceed the recognition accuracy compared to that of the methods based on the acoustic feature. It is confirmed that the 3D convolution model based on temporal modulation cues can extract discriminative auditory representation to recognize the emotions effectively.

The high-level auditory representation of sequence data is divided into non-overlapping segments in 3D CNN architecture. These non-overlapping segment-level features cannot fully reflect the dynamic changes of real emotions. This study proposed an attention-based sliding recurrent neural network (ASRNN) to continuously track spectral-temporal representation and capture salient emotion regions for categorical emotion recognition. In the ASRNN model, the sliding window is used to extract the continuous segment level internal representation, and the temporal attention model is used to capture salient regions of emotion representation. The experimental results show that the proposed method could capture the salient emotion regions and exhibit better recognition accuracy. In addition, to explore the relationship between temporal attention model and human auditory attention, a subjective evaluation experiment is designed to analyze the correlation between them, and the results show that they have a strong correlation. This implies that the proposed attention model can reflect human selective attention to a large extent.

It is a new trend to recognize dimension emotion continuously in human-robot interaction. It can help the robot capture the temporal dynamics of the speaker's emotion in real-time. In dimension emotion recognition, we need to track the change of arousal and valence value of short-term frames in a long sequence. Therefore, this study proposed a multi-resolution modulation-filtered cochleagram feature (MMCG) to capture the temporal and contextual modulation cues and used a parallel LSTM to track the emotion dynamics of emotion. This feature can encode temporal modulation cues into different resolution modulation-filtered cochleagram to capture temporal and contextual information. The parallel LSTM network architecture is designed to model the temporal dependence of each resolution modulation-filtered cochleagram feature and track the temporal dynamics of emotion. The experimental results showed that the MMCG feature could effectively extract high-level auditory representations for emotion recognition and the parallel LSTM network can model auditory representation sequence to track the temporal dynamics of emotion for dimensional emotion recognition. Compared with other evaluated features, experimental results showed that MMCG features could significantly improve the performance of emotion recognition.

In conclusion, to categorical emotion recognition, this study proposed 3D CNN to extract joint spectral-temporal auditory representations and ASRNN to capture the salient from auditory representation sequence. Experiments showed that the 3D convolution model based on temporal modulation cues could extract discriminative auditory representation and the ASRNN model can effectively model human selective attention to capture the salient regions of emotion representation. To dimensional emotion recognition, this study proposed the MMCG feature and the parallel LSTM to capture the temporal dynamics for both valence and arousal prediction. Experiments showed that this method could effectively predict dimensional emotion. It is confirmed that using auditory representation from the auditory model and deep learning methods lead to better results in categorical and dimensional emotion recognition. This indicates that the auditory representation can provide spectral-temporal representations, and deep learning frameworks can effectively extract emotional information from such representation for emotion recognition.

6.2 Future work

Identifying emotional states based on human auditory characteristics using deep learning methods is a perspective way. This study mainly focuses on joint spectral-temporal analysis, continuous tracking of salient emotion regions, and capturing temporal dynamics of emotion. In the future, the research of speech emotion recognition based on auditory characteristics mainly considers the following aspects:

1) Application of multi-resolution modulation-filtered cochleagram feature for other acoustic scene analysis

The experimental results of this study show that the MMCG feature can significantly improve the valence and arousal prediction performance compared with the acoustic-based features. This feature should also be effective for other acoustic scene analysis. For future work, we plan to investigate this feature for other acoustic scene analysis, such as categorical emotion recognition, speech separation, voice activity detection, etc.

2) Further exploration of auditory characteristics in speech perception

The modulated signal in each band can be regarded as a temporal amplitude envelope with a carrier (temporal fine structure). Our existing methods mainly extract temporal modulation cues from time-domain envelope signals for emotion recognition. It is also necessary to understand the contribution of temporal fine structure for speech perception. In addition, the effect of frequency-domain processing of the auditory system on speech perception is also worth further investigation.

3) Robustness analysis of auditory-based features

Although many studies on auditory features have come to a common conclusion that auditory feature representation is noise-robust, however, there is no quantitative analysis on the noise robustness of the categorical emotion recognition method based on 3D spectral-temporal modulation representation and the dimension emotion recognition method based on the MMCG features. Therefore, the feature robustness in different environmental noise should be analyzed in future work.

4) Exploration of the multimodal emotion recognition method

The content of the interaction in natural HRI should be multimodal. It may use speech, spoken text, emoticons, facial expressions, and other information at the same time. It should be able to extract feature recognition emotion from different modal data to realize natural human-computer interaction so that users' emotional state can be recognized even when some modes are missing.

Bibliography

[1] M. Minsky, *Society of mind*. Simon and Schuster, 1988.

[2] K. Bahreini, R. Nadolski, and W. Westera, “Towards multimodal emotion recognition in e-learning environments,” *Interact. Learn. Environ.*, vol. 24, no. 3, pp. 590–605, 2016.

[3] A. Popescu, J. Broekens, and M. Van Someren, “Gamygdala: An emotion engine for games,” *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 32–44, 2013.

[4] M. Jiang and L. Zhang, “Big data analytics as a service for affective humanoid service robots,” *Procedia Comput. Sci.*, vol. 53, pp. 141–148, 2015.

[5] J. S. K. Ooi, S. A. Ahmad, H. R. Harun, Y. Z. Chong, and S. H. M. Ali, “A conceptual emotion recognition framework: stress and anger analysis for car accidents,” *Int. J. Veh. Saf.*, vol. 9, no. 3, pp. 181–195, 2017.

[6] H. Park, J.-B. Kim, S.-G. Bae, and M.-S. Kim, “A Study on the Lie Detection of Telephone Voices Using Support Vector Machine,” in *International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, 2018, pp. 89–99.

[7] L. Joseph, S. Pramod, and L. S. Nair, “Emotion recognition in a social robot for robot-assisted therapy to autistic treatment using deep learning,” in *2017 International Conference on Technological Advancements in Power and Energy (TAP Energy)*, 2017, pp. 1–6.

[8] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, “The design and implementation of XiaoIce, an empathetic social chatbot,” *Comput. Linguist.*, no. Just Accepted, pp. 1–62, 2018.

[9] K.-J. Oh, D. Lee, B. Ko, and H.-J. Choi, “A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation,” in *2017 18th IEEE International Conference on Mobile Data Management (MDM)*, 2017, pp. 371–375.

[10] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

[11] T. Dau, B. Kollmeier, and A. Kohlrausch, “Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration,” *J. Acoust. Soc. Am.*, vol. 102, no. 5, pp. 2906–2919, 1997.

[12] A. R. Møller, “Unit responses in the rat cochlear nucleus to tones of rapidly varying

frequency and amplitude," *Acta Physiol. Scand.*, vol. 81, no. 4, pp. 540–556, 1971.

[13] N. Suga, "Analysis of information-bearing elements in complex sounds by auditory neurons of bats," *Audiology*, vol. 11, no. 1–2, pp. 58–72, 1972.

[14] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 887–906, 2005.

[15] F. E. Theunissen, K. Sen, and A. J. Doupe, "Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds," *J. Neurosci.*, vol. 20, no. 6, pp. 2315–2331, 2000.

[16] R. Drullman, "Temporal envelope and fine structure cues for speech intelligibility," *J. Acoust. Soc. Am.*, vol. 97, no. 1, pp. 585–592, 1995.

[17] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP J. Appl. Signal Processing*, vol. 2003, no. 7, pp. 668–675, 2003.

[18] M. Unoki and Z. Zhu, "Relationship between contributions of temporal amplitude envelope of speech and modulation transfer function in room acoustics to perception of noise-vocoded speech," *Acoust. Sci. Technol.*, vol. 41, no. 1, pp. 233–244, 2020.

[19] J. H. McDermott and E. P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis," *Neuron*, vol. 71, no. 5, pp. 926–940, 2011.

[20] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Contributions of temporal cue on the perception of speaker individuality and vocal emotion for noise-vocoded speech," *Acoust. Sci. Technol.*, vol. 3, pp. 234–242, 2018.

[21] N. Moritz, J. Anemueller, and B. Kollmeier, "An Auditory Inspired Amplitude Modulation Filter Bank for Robust Feature Extraction in Automatic Speech Recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 11, pp. 1–1, 2015.

[22] H. Yin, V. Hohmann, and C. Nadeu, "Acoustic features for speech recognition based on Gammatone filterbank and instantaneous frequency," *Speech Commun.*, vol. 53, no. 5, pp. 707–715, 2011.

[23] R. V Sharan and T. J. Moir, "Acoustic event recognition using cochleagram image and convolutional neural networks," *Appl. Acoust.*, vol. 148, pp. 62–66, 2019.

[24] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Contribution of modulation spectral features on the perception of vocal-emotion using noise-vocoded speech," *Acoust. Sci. Technol.*, vol. 6, pp. 379–386, 2018.

[25] S. Wu, T. H. Falk, and W. Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Commun.*, vol. 53, no. 5, pp. 768–785, 2011.

[26] A. R. Avila, Z. A. Momin, J. F. Santos, D. OShaughnessy, and T. H. Falk, “Feature Pooling of Modulation Spectrum Features for Improved Speech Emotion Recognition in the wild,” *IEEE Trans. Affect. Comput.*, vol. 3045, no. c, pp. 1–1, 2018.

[27] T. H. Falk and W. Y. Chan, “Spectro-temporal features for robust far-field speaker identification,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 18, no. 1, pp. 634–637, 2008.

[28] O. Chapelle and V. Vapnik, “Model selection for support vector machines,” in *Advances in neural information processing systems*, 2000, pp. 230–236.

[29] Y. LeCun, Y. Bengio, and others, “Convolutional networks for images, speech, and time series,” *Handb. brain theory neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[30] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] A. Graves, A. Mohamed, and G. Hinton, “Speech Recognition with Deep Recurrent Neural Networks,” no. 3, 2013.

[32] G. Valenza, S. Member, and A. Lanata, “The Role of Nonlinear Dynamics in Affective Valence and Arousal Recognition,” *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 237–249, 2012.

[33] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, “Spectro-temporal modulation transfer functions and speech intelligibility,” *J. Acoust. Soc. Am.*, vol. 106, no. 5, pp. 2719–2732, 1999.

[34] Z.-Q. Wang and I. Tashev, “Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks,” in *Proc. 42nd IEEE Int. Conf. Acoust., Speech and Signal Process.*, 2017, pp. 5150–5154.

[35] J. Lee and I. Tashev, “High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition,” *Proc. INTERSPEECH 2015 16th Annu. Conf. Int. Speech Commun. Assoc.*, pp. 1537–1540, 2015.

[36] C. Stevens and D. Bavelier, “The role of selective attention on academic foundations : A cognitive neuroscience perspective,” *Dev. Cogn. Neurosci.*, vol. 2, pp. S30–S48, 2012.

[37] S. Chen, “Multi-modal Dimensional Emotion Recognition using Recurrent Neural Networks,” *AVEC ’15 Proc. 5th Int. Work. Audio/Visual Emot. Chall.*, pp. 49–56, 2015.

[38] P. Ekman, “An argument for basic emotions,” *Cogn. Emot.*, vol. 6, no. 3–4, pp. 169–200, 1992.

[39] R. Cowie *et al.*, “Emotion recognition in human-computer interaction,” *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, 2001.

[40] J. T. F. L. M. Zhang and H. Jia, “Design of speech corpus for mandarin text to speech,” in *The Blizzard Challenge 2008 workshop*, 2008.

[41] C. Busso *et al.*, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[42] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, “MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception,” *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, 2017.

[43] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, “Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions,” *2013 10th IEEE Int. Conf. Work. Autom. Face Gesture Recognition, FG 2013*, no. i, 2013.

[44] J. Kossaifi *et al.*, “SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild,” *arXiv Prepr. arXiv1901.02839*, 2019.

[45] M. Valstar *et al.*, “AVEC 2016 - Depression, Mood, and Emotion Recognition Workshop and Challenge,” pp. 3–10, 2016.

[46] F. Ringeval *et al.*, “Avec 2017: Real-life depression, and affect recognition workshop and challenge,” in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 3–9.

[47] A. Ouyang, T. Dang, V. Sethu, E. Ambikairajah, and E. Engineering, “Speech Based Emotion Prediction : Can a Linear Model Work ?,” pp. 2813–2817, 2019.

[48] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proc. 18th ACM int. conf. Multimedia*, 2010, pp. 1459–1462.

[49] B. Schuller, S. Steidl, and A. Batliner, “The interspeech 2009 emotion challenge,” in *Proc. INTERSPEECH 2009 10th Annu. Conf. Int. Speech Commun. Assoc.*, 2009.

[50] B. Schuller *et al.*, “The INTERSPEECH 2010 paralinguistic challenge,” in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[51] B. Schuller *et al.*, “The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism,” in *Proc. INTERSPEECH 2013 14th Annu. Conf. Int. Speech Commun. Assoc.*, 2013.

[52] F. Eyben *et al.*, “The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing,” *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2015.

[53] W. Lim, D. Jang, and T. Lee, “Speech emotion recognition using convolutional and Recurrent Neural Networks,” *2016 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, pp. 1–4, 2016.

[54] G. Keren and B. Schuller, “Convolutional RNN: An enhanced model for extracting features from sequential data,” *Proc. Int. Jt. Conf. Neural Networks*, vol. 2016-Octob, pp. 3412–3419, 2016.

[55] M. Neumann and N. T. Vu, “Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7390–7394.

[56] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, “Learning salient features for speech emotion recognition using convolutional neural networks,” *IEEE Trans. Multimed.*, vol. 16, no. 8, pp. 2203–2213, 2014.

[57] B. C. J. Moore and B. R. Glasberg, “Suggested formulae for calculating auditory-filter bandwidths and excitation patterns,” *J. Acoust. Soc. Am.*, vol. 74, no. 3, pp. 750–753, 1983.

[58] R. D. Patterson, M. H. Allerhand, and C. Giguere, “Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform,” *J. Acoust. Soc. Am.*, vol. 98, no. 4. pp. 1890–1894, 1995.

[59] T. Irino and R. D. Patterson, “A dynamic compressive gammachirp auditory filterbank,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 6, pp. 2222–2232, 2006.

[60] Zeng and F.-G., “Trends in Cochlear Implants,” *Trends Amplif.*, vol. 8, no. 1, pp. 1–34, 2004.

[61] F. E. Theunissen, K. Sen, and a J. Doupe, “Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds.,” *J. Neurosci.*, vol. 20, no. 6, pp. 2315–2331, 2000.

[62] I. K. Lin, “A Concordance Correlation Coefficient to Evaluate Reproducibility,” *Biometrics*, vol. 45, no. 1, pp. 255–268.

[63] V. Chernykh, G. Sterling, and P. Prihodko, “Emotion Recognition From Speech With Recurrent Neural Networks,” 2017.

[64] K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” in *Proc. INTERSPEECH 2014 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014.

[65] M. Chen, X. He, J. Yang, and H. Zhang, “3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition,” *IEEE Signal*

Process. Lett., vol. 25, no. 10, pp. 1440–1444, 2018.

[66] J. Kim, K. P. Truong, G. Englebienne, and V. Evers, “Learning spectro-temporal features with 3D CNNs for speech emotion recognition,” *2017 7th Int. Conf. Affect. Comput. Intell. Interact. ACII 2017*, vol. 2018-Janua, pp. 383–388, 2018.

[67] Y. Kim and E. M. Provost, “Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 3677–3681.

[68] E. M. Provost, “Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 3682–3686.

[69] B. R. Glasberg and B. C. J. Moore, “Derivation of auditory filter shapes from notched-noise data,” *Hear. Res.*, vol. 47, no. 1–2, pp. 103–138, 1990.

[70] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, “Representation Learning for Speech Emotion Recognition.,” in *Interspeech*, 2016, pp. 3603–3607.

[71] M. Neumann and N. T. Vu, “Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech,” *Proc. INTERSPEECH 2017 18th Annu. Conf. Int. Speech Commun. Assoc.*, vol. 2017-Augus, no. 3, pp. 1263–1267, 2017.

[72] S. Rosen, “Temporal Information in Speech: Acoustic, Auditory and Linguistic Aspects,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 336, no. 1278. pp. 367–373, 1992.

[73] R. D. Patterson, M. Unoki, and T. Irino, “Extending the domain of center frequencies for the compressive gammachirp auditory filter,” *J. Acoust. Soc. Am.*, vol. 114, no. 3, pp. 1529–1542, 2003.

[74] M. Unoki, T. Irino, B. Glasberg, B. C. J. Moore, and R. D. Patterson, “Comparison of the roex and gammachirp filters as representations of the auditory filter,” *J. Acoust. Soc. Am.*, vol. 120, no. 3, pp. 1474–1492, 2006.

[75] G. Zhu, L. Zhang, P. Shen, and J. Song, “Multimodal Gesture Recognition Using 3-D Convolution and Convolutional LSTM,” *IEEE Access*, vol. 5, pp. 4517–4524, 2017.

[76] Y. Li, T. Zhao, and T. Kawahara, “Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning,” *Proc. Interspeech 2019*, pp. 2803–2807, 2019.

[77] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, “Dilated Residual Network with Multi-head Self-attention for Speech Emotion Recognition,” in *ICASSP 2019-2019 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6675–6679.

[78] C.-C. Chang and C.-J. Lin, “LIBSVM: a library for support vector machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, 2011.

[79] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: an update,” *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[80] A. J. E. Kell *et al.*, “A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy,” *Neuron*, vol. 98, no. 3, pp. 1–15, 2018.

[81] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, “Direct Modelling of Speech Emotion from Raw Speech,” *arXiv Prepr. arXiv1904.03833v3*.

[82] Z. Aldeneh and E. M. Provost, “Using regional saliency for speech emotion recognition,” *Proc. 42th IEEE Int. Conf. Acoust. Speech Signal Process.*, no. 3, pp. 2741–2745, 2017.

[83] R. Xia and Y. Liu, “A Multi-Task Learning Framework for Emotion Recognition Using 2D Continuous Space,” *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 3–14, 2017.

[84] Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, and C. Li, “Exploring Spatio-Temporal Representations by Integrating Attention-based Bidirectional-LSTM-RNNs and FCNs for Speech Emotion Recognition,” *Proc. Interspeech 2018,* no. September, pp. 272–276, 2018.

[85] C. Z. Seyedmahdad Mirsamadi, Emad Barsoum, “Automatic speech emotion recognition using recurrent neural networks with local attention,” *Proc. 42nd IEEE Int. Conf. Acoust. Speech, Signal Process. ICASSP 2017*, pp. 2227–2231, 2017.

[86] C.-W. Huang and S. S. Narayanan, “Attention Assisted Discovery of Sub-Utterance Structure in Speech Emotion Recognition,” *Interspeech 2016,* pp. 1387–1391, 2016.

[87] S. Sukittanon, L. E. Atlas, J. W. Pitton, and K. Filali, “Improved modulation spectrum through multi-scale modulation frequency decomposition,” in *Proceedings.(ICASSP’05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, 2005, vol. 4, pp. iv--517.

[88] J. Chen, Y. Wang, and D. Wang, “A feature study for classification-based speech separation at low signal-to-noise ratios,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1993–2002, 2014.

[89] R. Santoro *et al.*, “Encoding of Natural Sounds at Multiple Spectral and Temporal

Resolutions in the Human Auditory Cortex," *PLoS Comput. Biol.*, vol. 10, no. 1, 2014.

[90] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-modeling of continuous emotions in an audiovisual affect recognition framework," *Image Vis. Comput.*, vol. 31, no. 2, pp. 153–163, 2013.

[91] Z. Peng, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Dimensional Emotion Recognition from Speech Using Modulation Spectral Features and Recurrent Neural Networks," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 524–528.

[92] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Reconstruction-error-based learning for continuous emotion recognition in speech," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2367–2371.

[93] M. Schmitt and B. Schuller, "Deep recurrent neural networks for emotion recognition in speech," in *Proc. DAGA*, 2018, vol. 44, pp. 1537–1540.

[94] M. Schmitt and N. Cummins, "Continuous Emotion Recognition in Speech – Do We Need Recurrence ?," *Proc. Interspeech 2019*, pp. 2808–2812, 2019.

[95] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, Springer, 2005, pp. 181–197.

[96] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," *Pap. Present. Meet. IOC Speech Gr. Audit. Model. RSRE*, vol. 2341, no. December, pp. 14–15, 1987.

[97] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5089–5093.

[98] G. Trigeorgis *et al.*, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," *Proc. 41st IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 5200–5204, 2016.

[99] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-modeling of continuous emotions in an audiovisual affect recognition framework," *Image Vis. Comput.*, vol. 31, no. 2, pp. 153–163, 2013.

[100] X.-L. Zhang and D. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *Fifteenth annual conference of the international speech communication association*, 2014.

[101] F. Haider and S. Luz, "Attitude Recognition Using Multi-resolution Cochleagram

Features," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3737–3741.

[102] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. Synth. Emot.*, vol. 1, no. 1, pp. 68–99, 2010.

[103] Z. Zhang, J. Han, and E. Coutinho, "Dynamic difficulty awareness training for continuous emotion prediction," *IEEE Trans. Multimed.*, vol. 21, no. 5, pp. 1289–1301, 2019.

[104] K. Brady *et al.*, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 97–104.

[105] F. Povolny *et al.*, "Multimodal emotion recognition for AVEC 2016 challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 75–82.

[106] D. Le, Z. Aldeneh, E. M. Provost, and A. Arbor, "Discretized Continuous Speech Emotion Recognition with Multi-Task Deep Recurrent Neural Network," *Interspeech2017*, pp. 1108–1112, 2017.

[107] J. Zhao, R. Li, S. Chen, and Q. Jin, "Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, 2018, pp. 65–72.

[108] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 890–897.

[109] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 19–26.

[110] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning Sound Representations from Unlabeled Video," no. Nips, 2016.

List of Publications

Journal Paper

[1] Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, “Speech Emotion Recognition Using 3D Convolutions and Attention-Based Sliding Recurrent Networks With Auditory Front-Ends,” IEEE Access, vol. 8, pp. 16560–16572, 2020.

[2] Z. Peng, J. Dang, M. Unoki, and M. Akagi, “Multi-resolution modulation-filtered cochleagram features for LSTM-based dimensional emotion recognition from speech,” Neural networks. (Under review)

International Conference

[1] Z. Peng, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, “Speech emotion recognition using multichannel parallel convolutional recurrent neural networks based on Gammatone auditory filterbank,” in 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017, pp. 1750–1755.

[2] Z. Peng, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, “Auditory-inspired end-to-end speech emotion recognition using 3d convolutional recurrent neural networks based on spectral- temporal representation,” In: 2018 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2018. p. 1-6.

[3] Z. Peng, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, “Dimensional Emotion Recognition from Speech Using Modulation Spectral Features and Recurrent Neural Networks,” in 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019, pp. 524–528.

[4] S. Peng, Q. Hu, J. Dang, and Z. Peng. "Stochastic Sequential Minimal Optimization for Large-Scale Linear SVM." International Conference on Neural Information Processing. Springer, Cham, 2017, pp. 279-288

Domestic Conference:

[1] Z. Peng, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, “End-to-end speech emotion recognition using 3D convolutional recurrent neural networks based on modulation spectral features,” Acoustic Society of Japan, Spring, 2018, 2-Q-10.

[2] Z. Peng, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, “Auditory-inspired end-to-end speech emotion recognition using 3d convolutional recurrent neural networks based on spectral-temporal modulation” in JAIST world conference (JWC), 2018