TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

School **School of Computer Science & Robotics**
Academic program **09.04.01 Computer Science and Engineering**
Division **Division for Information Technology**

MASTER'S GRADUATION THESIS

Topic of research work
Recognizing emotion in human speech using deep learning techniques

UDC 004.934.1:004.85

Student

Group	Full name	Signature	Date
8BM03	Chen Jin		

Scientific supervisor

Position	Full name	Academic degree, academic rank	Signature	Date
Associate Professor	Botygin I.A.	PhD		

SECTION ADVISERS:

Section «**Financial Management, Resource Efficiency and Resource Saving**»

Position	Full name	Academic degree, academic rank	Signature	Date
Associate Professor	Bylkova T.V.	PhD		

Section «**Social Responsibility**»

Position	Full name	Academic degree, academic rank	Signature	Date
Full Professor	Fedorenko O.Yu.	PhD		

ADMITTED TO DEFENSE:

Director of program	Full name	Academic degree, academic rank	Signature	Date
Artificial intelligence and machine learning	Spitsyn V.G.	PhD		

Tomsk – 2022

LEARNING OUTCOMES

Expected learning outcomes

Code competencies	Learning outcome (a graduate should be ready)
Universal competencies	
UK(U)-1	Able to critically analyze problematic situations using a systematic approach, to develop a strategy of action
UK(U)-2	Able to manage a project through all stages of its life cycle
UK(U)-3	Able to organize and manage a team, develop a team strategy to reach the set target
UK(U)-4	Able to use modern communication technologies, also in foreign language(s), for academic and professional interactions
UK(U)-5	Able to analyze and take into account the diversity of cultures in the process of intercultural interaction
UK(U)-6	Able to identify and implement priorities of their own activities and ways to improve them on the basis of self-assessment
General professional competencies	
GPC(U)-1	Able to independently acquire, develop and apply mathematical, natural-science, socio-economic and professional knowledge to solve non-standard tasks, including in a new or unfamiliar environment and in an interdisciplinary context
GPC(U)-2	Able to develop original algorithms and software tools, including those using modern intellectual technologies, to solve professional tasks
GPC(U)-3	Capable of analyzing professional information, summarizing, structuring, presenting in analytical reviews with substantiated conclusions and recommendations
GPC(U)-4	Capable of applying new scientific principles and research methods in practice
GPC(U)-5	Capable of developing and upgrading software and hardware for information and automated systems
GPC(U)-6	Capable of developing components of hardware-software complexes for information processing and computer-aided design
GPC(U)-7	Capable of adapting foreign data processing and CAD systems to the needs of domestic enterprises

GPC(U)-8	Capable of managing effectively the development of software tools and designs
Professional competencies	
PC(U)-1	Capable of creating software for analysis, recognition and processing of information, digital signal processing systems (06.042 "Big Data Specialist", 06.001 "Programmer")
PC(U)-2	Capable of designing complex user interfaces (06.025 "Graphic and User Interface Designer")
PC(U)-3	Capable of managing processes and projects for creation (modification) of information resources (06.017 "Software Development Manager")
PC(U)-4	Capable of managing the development of complex projects at all stages and phases of work (40.008 "Specialist in the organization and management of scientific is able to manage complex projects at all stages of work performance (40.008 "Specialist in the organization and management of research and development activities")
PC(U)-5	Capable of designing and organizing the educational process of the educational programmes with the use of modern educational technologies (01.002 "Educational psychologist (psychologist in education)")

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

School **School of Computer Science & Robotics**
Academic program **09.04.01 Computer Science and Engineering**
Division **Division for Information Technology**

APPROVED BY:
Director of program
_____ _____ Spitsyn V.G.
«_____» _____2022 г.

ASSIGNMENT
for the Graduation Thesis completion

In the form:

Master's thesis

For a student:

Group	**Full Name**
8BM03	Chen Jin

Topic of research work:

Recognizing emotion in human speech using deep learning techniques	
Approved by the rector's order (date, ID)	№ 34-63/c от 03.02.2022

Deadline for completion of Master's Graduation Thesis	10.06.2022

TERMS OF REFERENCE:

Initial data for research work *(name of the object of research or design; productivity or load; mode of operation (continuous, periodic, cyclic, etc.); type of raw materials or product material; requirements for the product, product or process; special requirements for the features of operation (operation) of the object or product in terms of operational safety, impact on the environment, energy costs; economic analysis, etc.)*	The object of research and development is a deep convolutional neural network algorithm model for speech emotion recognition

List of issues to be researched, designed and developed *(analytical review of literary sources in order to clarify the achievements of world technology science in the field under consideration; statement of the problem of research, design, construction; content of the procedure of research, design, construction; discussion of the results of the work performed; the name of additional sections to be developed; conclusion of the work).*	1. Overview of approaches to the problem of recognizing emotion from a speech signal. 2. Basic Theory of Speech Emotion Recognition. 3. Emotion Recognition Based on MFCC Features. 4. Speech Emotion Recognition Based on MS-ResNet. 5. Financial management, resource efficiency, and resource saving. 6. Social responsibility.

Advisors to the sections of the Master's Graduation Thesis
(with indication of sections)

Section	Advisor
Financial Management,Resource Efficiency and Resource Saving	Associate Professor Bylkova T.V.
Social Responsibility	Full Professor Fedorenko O.Yu.
English Language	Senior Lecturer Anufrieva T.N.

Recognizing emotion in human speech using deep learning techniques	10.03.2022

The assignment was given by the scientific supervisor:

Position	Full name	Academic degree, academic rank	Signature	Date
Associate Professor	Botygin I.A.	PhD		

Assignment accepted for execution by a student:

Group	Full name	Signature	Date
8BM03	Chen Jin		

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

School **School of Computer Science & Robotics**
Academic program **09.04.01 Computer Science and Engineering**
Division **Division for Information Technology**

Form of presentation of the work:

Master's Thesis

SCHEDULED ASSESSMENT CALENDAR
for the Master's Graduation Thesis completion

Student:

Group	Full Name
8BM03	Chen Jin

Topic of research work:

Recognizing emotion in human speech using deep learning techniques

Deadline for completion of Master's Graduation Thesis:	10.06.2022

Assessment date	Title of the section (module) / type of work (research)	Maximum score of a section (module)
1.03.2022	Drawing up and approving the terms of reference	10
10.03.2022	Selection and study of materials on the topic	10
30.03.2022	Research of subject area	15
20.04.2022	Conducting experiments	25
15.05.2022	Analysis and description of results	25
01.06.2022	Preparing for thesis defense	15

COMPILED BY:
Scientific supervisors:

Position	Full name	Academic degree, academic rank	Signature	Date
Associate Professor	Botygin I.A.	PhD		

AGREED BY:
Director of the program

Director of program	Full name	Academic degree, academic rank	Signature	Date
Artificial intelligence and machine learning	Spitsyn V.G.	PhD		

Student

Group	Full Name	Signature	Date
8BM03	Chen Jin		

TASK FOR
«FINANCIAL MANAGEMENT, RESOURSE EFFICIENCY AND RESOURSE SAVING» SECTION

To student:

Group	Name
8BM03	Chen Jin

Institute	School of Computer Science & Robotics	**Department**	Division for Information Technology
Education level	Master Degree Program	**Academic program**	Computer Science and Engineering

Initial data to «Financial management, resource efficiency and resource saving » chapter:	
1. Costs of research, including technical, financial, energy, information and human costs	Use the current price tags and contractual prices for the consumed material and information resources, as well as the value of the tariff for e-mail specified in the MU. energy
2. Norms of expenditure of resources	…
3. The taxation system used, the rates of taxes, discounting and lending	Current rates of the unified social tax and VAT, discount rate = 4%
List of tasks:	
1. Evaluation of commercial and innovative potential	To characterize the existing and potential consumers (buyers) of the WRC results, the expected scale of their use
2. Development of the charter of the technical project	Develop a draft of such a charter if the implementation of the results of the WRC requires the creation of a separate organization or a separate structural unit within an existing organization
3. Planning of management process: structure and schedule, budget, and risks	Construction of a schedule for the implementation of WRC, drawing up an appropriate cost estimate, calculation of the price of the WRC result.
4. Estimation of resource, financial and economic efficiency	Evaluation of the economic efficiency of using the results of the WRC, characterization of other types of effect
List of graphical data:	
1. «Portrait» of consumer *2. Market segmentation* *3. Assessment of the competiveness of solution* *4. SWOT matrix* *5. Calendar and budget of the research* *6. Assessment of resource, financial and economic efficiency* *7. Potential risks*	

Date of task obtaining	

The task was given by the adviser:

Position	Name	Academic degree	Signature	Date
Associate Professor	Bylkova Tatyana Vasilievna	PhD		

The task was accepted by the student:

Group	Name	Signature	Date
8BM03	Chen Jin		

TASK FOR
«SOCIAL RESPONSIBILITY» PART

To student:

Group	Name		
8BM03	Chen Jin		
Institute	School of Computer Science & Robotics	**Department**	Division for Information Technology
Educational level	Master Degree Program	**Academic program**	Computer science and engineering

Subject BKP:

Speech Emotion Recognition Based on Convolutional Neural Networks	
Initial data to «Social responsibility» chapter:	
Introduction – Characteristics of the object of study (substance, material, device, algorithm, technique) and the scope of its application. – Description of the working area (workplace) when developing a design solution / during operation	*Object of study :Neural network algorithm* *Application area :Recognize emotions in speech* *Working area : office* *Room dimensions :8*4m^2* *Quantity and name of equipment of the working area: Desktop and personal computer.* *Indoor temperature: 20℃-28℃.* *Ventilation condition: good.* *Lighting condition: 300lx* *Work processes associated with the object of study, carried out in the working area: Design the relevant algorithm model, select the algorithm model data set, write the algorithm code, and run the algorithm model.*
List of items to be investigated and to be developed:	
1. Legal and organizational issues to provide safety when developing a design solution: - Special (specific for operation of objects of investigation,designed workplace) legal rules of labor legislation; - Organizational activities for layout of workplace.	- GOST 12.2.032-78 SSBT. Workplace when performing work while sitting. General ergonomic requirements. - Labor Code: Federal Law No. 197-FZ of December 30, 2001 (as amended on March 9, 2021). - GOST 12.0.003-2015 Hazardous and harmful production factors. Classification. List of dangerous and harmful factors. - GOST 22269-76 “Operator's workplace. Mutual arrangement of workplace elements. - GOST R 50923-96. Displays. Operator's workplace. General ergonomic and work environment requirements. Measurement methods. - GOST 21889-76 "Man-machine" system. Operator's chair. General ergonomic requirements”. - SanPiN 1.2.3685-21 Hygienic standards and requirements for ensuring the safety and (or) harmlessness of environmental factors for humans. - GOST 12.1.005-88 System of labor safety standards (SSBT). General sanitary and hygienic requirements for the air of the working area. - SP 52.13330.2016 Natural and artificial lighting. Updated edition of SNiP 23-05-95. - GOST 12.1.003-83 Occupational safety standards system (SSBT). Noise. General safety requirements. - GOST 12.1.030-81 System of labor safety standards (SSBT). Electrical safety. Protective ground. Zeroing.

	- GOST 12.1.038-82 Occupational safety standards system (SSBT). Electrical safety. Maximum allowable values of touch voltages and currents. - GOST 12.1.004-91 Occupational safety standards system (SSBT). Fire safety. General requirements. - GOST 17.4.3.04-85 Nature Protection (SSOP). Soils. General requirements for control and protection against pollution.
2. Work Safety when developing a design solution: -Analysis of identified harmful and dangerous factors. -Justification of measures to reduce probability of harmful and dangerous factors	Harmful factors: - Increased levels of electromagnetic radiation. - Insufficient lighting in work area. - Excessive workplace noise. - Eyestrain. - The labor process is monotonous. Dangerous factors: - Electric shock Harm reduction methods: Collective Protection: - Adding a protective coating reduces ionizing radiation levels. -Increase light sources, install lights to improve lighting conditions - Add sound insulation and sound absorption devices to reduce noise - Add insulation and protection devices to prevent electric shock Personal protection : - Wear anti-noise headphones to prevent noise Calculations will be made for factors such as noise, lighting, etc.
3. Ecological safety when developing a design solution	- There is no impact on the residential area, hydrosphere and atmosphere. - Lithosphere: when disposing of fluorescent lamps and office equipment.
4. Safety in emergency situations when developing a design solution:	Possible emergencies: Natural disasters (floods, hurricanes, etc.); Geological impacts (earthquakes, landslides) The most typical emergency: Fire
Assignment data for section according to schedule	

The task was issued by consultant:

Position	Full name	Scientific degree, rank	Signature	date
Full Professor	Fedorenko O.Yu.	PhD		

The task was accepted by student:

Group	Full name	Signature	date
8BM03	Chen Jin		

Summary

The Master's Graduation Thesis contains: 85 pages, 25 images, 23 tables, 51 references.

Keywords: Speech emotion recognition, Deep learning, Convolutional neural network, ResNet, GoogLeNet.

The research object is the emotion of human speech.

The purpose of the work is to design a deep learning algorithm to recognize emotions in human language by building a deep convolutional neural network model.

In a research project, learn how to build a new type of deep convolutional neural network model and use Python to build that model, training it with a database, to build a program that recognizes emotion in human speech.

The results show that using Mel-frequency cepstral coefficients as emotion features and using multi-scale residual networks can improve the accuracy of emotion recognition. The conclusion is that the algorithm model of speech emotion recognition is constructed. Using this model and the Mel frequency cepstral coefficients can accurately identify human speech emotion.

The research results are mainly used in the field of speech recognition. The recognition of human speech emotions can reduce the error rate in speech recognition, enabling machines to understand human speech more humanely.

List of abbreviations

MFCC	Mel Frequency Cepstral Coefficients
AER	Automatic Emotion recognition
HCI	Human-Computer Interaction
HRI	Human-Robot interaction
SVM	Support Vector Machine
KNN	K-Nearest Neighbor
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
CNN	Convolution Neural Network
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
LPCC	Linear Prediction Cepstral Coefficients
FFT	Fast Fourier transform
RBM	Restricted Boltzmann Machines
DNN	Deep Neural Network
BN	Batch Normalization

Table of contents

INTRODUCTION

Information processing and decision-making in human-computer interaction have received considerable attention in recent years. Intelligent information technologies and, in particular, human-machine interaction systems have developed significantly. Effectiveness of such systems depends largely on the quality of recognition of information coming from the user of automated system and purposefulness of human influence on objects of study. Achievement of the purpose of dialogue interaction of the computer and the user is possible only at the account of the majority of the aspects characterizing speech streams arising in the course of communication. One of the directions for improving the quality of information processing is to define human emotional responses.

The application of computerized speech recognition and the determination of its emotionality is primarily of interest to companies introducing robotic systems into people's daily lives, as well as to companies working with a large number of customers and wishing to move to a new level of communication with them. Effective communication in natural language has an important role to play in the multimedia society of the future, with easy-to-handle human-computer interfaces. The use of such interfaces, which leave customers feeling comfortable and satisfied when accessing information or services in self-service mode, makes it possible, with the quality of the software already achieved, to create socially relevant systems whose implementation will make public access to services and data cheaper, more convenient and round the clock.

One source of determining emotional reactions is speech. For example, Russian contains about 40% of emotionally colored words. In addition, emotions are encoded by certain acoustic parameters in the speech signal. Understanding these features of acoustic coding of emotions will make it possible to understand the very mechanism of emotion perception and expression.

Experts on the problem of recognizing emotions through the acoustic characteristics of speech have revealed that the emotional state of a speaker is naturally reflected in the acoustic characteristics of his speech and voice, which, in turn, is an objective basis for an adequate subjective perception of the speaker by the listener.

Despite the work done, a number of questions remain open, in particular those related to the level of accuracy of human recognition of basic emotions based on acoustic parameters of speech. It has been proved that the information about the conveyed emotion is conveyed by intonation characteristics, but the question of how these characteristics allow the recognition of emotional expression in human speech remains poorly understood. Solving this question is of significant, as it allows us to separate the semantic and intonational components of a speech message.

Thus, the development of methods for the automatic detection of human emotional reactions by voice is an urgent task, allowing solving a number of economic, social and domestic problems and playing an important role in security issues.

1 Overview of approaches to the problem of recognizing emotion from a speech signal

1.1 Understanding of emotions in philosophy and modern psychology

There are many different views in the scientific community on the nature of emotional processes. No single, generally accepted theory has yet been developed. In this connection, there is also no universal definition of the emotional process. Thus, in psychology, emotional processes are understood as having both mental and physiological components. They are distinguished from other psychophysiological processes by the fact that they reflect the meaning of something for the subject, and regulate his behavior, thinking and even perception in a corresponding way. According to 1, an emotional process (emotional phenomenon, emotional state) is a psychophysiological process that motivates and regulates activity, reflects the subjective meaning of objects and situations, and is represented in consciousness in the form of experience.

The practice of classifying emotional processes into affects, emotions and feelings according to their psychological features and patterns of flow is widespread. Often mood is also identified as a separate class. The result is the following classification:

· Affects – are short-term and intense emotional processes accompanied by pronounced movement and changes in the functioning of internal organs. For example, fright.

· Emotions – are longer-lasting and less intense emotional processes than affects, reflecting the subjective significance of situations, but not specific objects in themselves. For example, anxiety.

· Feelings – longer lasting and less intense than affect, emotional processes that reflect the subjective significance of specific objects in themselves. For example, hatred.

· Moods – reasonably prolonged emotional processes of low intensity. For example, boredom.

One of the key components of emotional processes is emotions and the emotional intelligence associated with them. In the course of development, emotions lose their direct instinctive basis, acquire a complex conditioned character, differentiate and form a variety of types of so-called higher emotional processes (social, intellectual, aesthetic), which in humans constitute the main content of their emotional life. Emotional intelligence is a person's ability to recognize emotions, understand the intentions, motivations and desires of others and his own, as well as the ability to manage his own emotions and those of others in order to solve practical problems.

Emotion, as an object of study, has existed practically as long as human beings have existed as an object of study in philosophy. Attitudes towards emotions have changed as philosophy has developed. Classical and New European philosophy was dominated by the idea of the «primacy of reason» over the senses and bodily life in general. Emotions and affective life were described as low-ranking phenomena relative to cognitive processes, or as «nuisances». Modern (postmodern) philosophy, in contrast, focuses on corporeality and sensuality, which are valued above reason and thought.

It should be noted that emotional orientation, unlike rational orientation, is much more short-term, involuntary, poorly understood, does not require special training, and is possible under conditions of information deficit. There are two sides to

emotions – the relationship to the situation (the objective world) and the relationship to the subjective state (needs, motivation).

In accordance with 2, we will adopt the following definition of emotion. An emotion is a mental process of medium duration that reflects a subjective evaluative attitude towards existing or possible situations and the objective world. This mental process is characterized by three components:

· The experience or awareness in the psyche of an emotion.

· Processes occurring in the nervous, endocrine, respiratory, digestive and other systems of the body.

· The observed expressive complexes, including changes in the face, gestures, voice patterns, etc.

Emotion, like many other mental phenomena, is understood in different ways by different authors, so the above definition cannot be considered either precise or universally accepted.

Interest in the problem of emotion grows along with the development of research into artificial intelligence. Increasingly, there is a need for human-like behavior, which is very difficult to reproduce without attempting to model the emotional apparatus. Emotion modelling becomes particularly important in the context of creating agents whose functionality is related to communication with humans. For many practical tasks and problems (e.g., recognizing emotions, realizing the effects or consequences of emotions), the development of bioinformatics and machine learning technologies is promising. However, solving individual problems (such as emotion recognition from photos/texts, etc.) has not yet led to a qualitative breakthrough in the modelling of emotional systems. Moreover, researchers are increasingly refusing to create a separate emotional system, appealing to the fact that the effects of emotions are implemented in the agent's behavior automatically if they have been incorporated into the data used to train the agent.

The number of computational models of emotion and architectures of affective agents has increased significantly over the past five years. Researchers in cognitive science, artificial intelligence, bioengineering, software, and robotics have long been engaged in creating "models of emotion". The reasons that motivate scientists and

developers to explore the nature of emotions and develop theoretical modelling of emotions are driven by the demands of practice - the need to qualitatively improve human-computer interaction, and the need to create more plausible and efficient artificial intelligence systems and robots.

1.2 The relevance of automatic emotion recognition in human speech

Emotion recognition has long been of interest to researchers, but in the past two decades, this research has evolved markedly due to improvements in hardware and software, as well as the growing demand for intelligent interlocutors.

Automatic emotion recognition (AER) is the process of identifying human emotions using a computational model that uses input data. Like many other machine-learning tasks, AER can be based on different modalities such as speech, facial expression, textual data, user behavior, etc. One of the distinguishing features of AER compared to other machine learning tasks (automatic speech recognition, age and gender recognition) is the high subjectivity of emotion evaluation.

Artificial Intelligence is making significant steps forward every year, embracing more and more applications and expanding its presence in our daily lives. In an effort to improve the quality of life and harness experiences, artificial intelligence is taking over routine tasks. System intelligence is determined by many factors, including the ability to communicate with users at a high level. And here's the emotion. Since they are closely related to someone's mood and behavior, they play an important role. Because they change the meaning of words and affect the user's decision-making ability. Here are some promising applications for such systems.

(1) Human-Computer Interaction (HCI)

Language interface is the most natural way for humans to interact with a computer system. The progression from simple voice commands to meaningful conversation is an important step in the integration of contextualized emotion recognition into everyday life. Without this component, it is difficult to imagine the system as an intelligent and humanoid agent.

Human-robot interaction (HRI). As a complement to HCI, intelligent robots are of particular interest. Such robots can interact with humans on demand, such as a cleaning robot or a humanoid robot used as an interactive partner and artificial listener. A robot can help elderly people cope with loneliness, stay active, exercise, prevent health problems caused by low brain activity and learn to use modern technological advances without the need for high computer literacy (just by using a universal language interface). The emotion recognition module is indispensable in such systems for achieving high standards of quality of life.

Health monitoring in hospitals is a specific application area for computer systems, especially robots. Such robots can continuously monitor the physiological and psychological state of the patient and provide basic care, reducing the workload of nurses. Assigned to a specific ward, the robots can act as a personal nurse, issuing specific commands to hospital staff as needed and providing round-the-clock monitoring of patients' conditions.

Information gathered from patients can be extremely useful in building a complete picture of health, preventing stress and depression, helping with domestic problems, alerting relatives or specialists (if necessary).

In addition to reducing workload, such robots can protect doctors and nurses from infection by eliminating direct contact with the patient.

Of course, for complex cases such as care of the elderly or psychological condition monitoring, emotion recognition systems need to be very robust, allowing for continuous operation over time and taking into account contextual information affecting the emotional state of the user.

(2) Recommender systems.

Integrating a motion recognition component into a multimedia or entertainment application can serve as an additional source of information. This can help improve the quality of the recommendation system by suggesting similar products to the user, based not only on their purchase history, but also on their emotional reactions and behavior.

(3) Online learning.

Many educational courses are now available online, and this area is growing steadily. It offers a huge opportunity for people all over the world to acquire knowledge

in almost any field. The introduction of an emotion recognition component can provide useful guidance to teachers on how to structure lessons, as well as monitor student interest and engagement.

Mood monitoring in a smart environment can also be an additional feature that enhances the usability of the environment. It is possible to integrate this component into every level of the environment: smart room or office, smart home, smart car, smart city.

Emotional voice scanner is needed in transport companies and dispatch services for automated introduction of restrictions or outright prohibition of access to job duties for persons in unstable or inadequate emotional state. Such monitoring systems would also allow for additional screening of passengers on flights as part of counter-terrorism activities.

Thus, knowing the set of words that make up a speech signal is not sufficient to reveal the meaning of a message. In addition to the main verbal channel, a person often uses paralinguistic means of communication, such as phonation (tone of voice, tempo and volume) or kinetic (gestures, facial expressions). Correct interpretation of the context is not possible without taking into account the non-verbal cues that accompany verbal constructions. Therefore, the recognition of the emotional state of the speaker is a key aspect in the analysis of spoken language.

The scientific field that aims to develop and create systems capable of recognizing, interpreting, processing, and simulating human emotions is called affective computing. It is an interdisciplinary field of knowledge that incorporates artificial intelligence, psychology and cognitive science technologies. It is clear that in everyday life, affective interfaces used in human-machine communication lack the ability to measure the full range of emotional attributes using sensors. Multimodal systems are being developed that use both video and audio channels to obtain measurements.

This paper considers an approach to recognizing the emotional state of a person from a set of acoustic characteristics of the speech signal.

1.3 A review and comparative analysis of research in the field of language emotion

The initial research methods of speech emotion recognition use machine learning methods, including discriminative models (SVM, KNN) or generative models (GMM, HMM). There is almost no difference between the two in practical effect, they are on the same level, and they were still the main method until the beginning of the 20th century. However, for these methods, different feature selection will have a great impact on the accuracy of model recognition, which leads to most researchers need to spend a lot of time on feature set selection. With the increase in research on speech emotion recognition, and due to the improvement of computing power, deep learning methods have been introduced into speech emotion recognition. Various speech emotion recognition methods based on RNN, DNN and CNN have gradually emerged.

1.3.1Speech emotion recognition based on traditional methods

Traditional methods are mainly machine learning methods, which are divided into discriminative models and generative models. The discriminative model outputs the posterior probability P(T|x) on the label T for a given input x.

In speech emotion recognition, the first model used is KNN. In 1996, Dellaert et al. carried out the pioneering work based on KNN speech emotion recognition 3. The experimental results proved that the KNN method has certain emotion recognition ability for speech features. The KNN classifier method is simple and requires little training, but differences in feature selection can have dramatic effects on the results.

Another more popular discriminative model is SVM, which mainly finds a classification hyperplane in the emotional feature space. The earliest research in speech emotion recognition was conducted by Yu et al. in 2001 4. They sentiment-marked more than 2000 speech segments, then extracted prosodic features and used a Gaussian kernel-based SVM for classification judgment. Of course, feature selection still has a great impact on SVM.

Different from the discriminative model, the generative model establishes the joint probability distribution P(T,x) between the sentiment feature x and the sentiment label T, which is used in parallel with the discriminative model in the same period.

Commonly used generative models include Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs).

In 1998, Slaney et al. used the GMM model 5 to collect 509 sentences by recording adults speaking to infants, and defined three emotion types, and trained a Gaussian mixture containing 10 Gaussians for each emotion type. Model. After that, similar methods were also used on different datasets (Schuller et al. 6, Wang et al. 7). After that, Neiberg et al. borrowed the UBM-GMM method 8 in the dataset containing 3 emotions. A good recognition effect has been achieved on the data set of the category 9. ZHOU et al. further established a GMM model for each sentence in the sentiment dataset through MAP 10, and then concatenated the means of all Gaussian distributions in each sentence to form a supervector. Finally, the supervector is used as the sentence-level feature, and the SVM classifier is used for classification, which is significantly improved compared to using the GMM model alone.

Another class of generative models is the Hidden Markov Model (HMM). The process of training the HMM is equivalent to the process of estimating the model with known observed states. In the field of speech recognition, the recognition model of HMM usually uses frame-level speech features, trains an HMM separately for each emotion, and then calculates the probability of the observed state through the forward-reverse algorithm.

From the 1990s to the early 2000s, the vast majority of speech emotion recognition methods were based on discriminative or generative models, which appeared at the same time. In contrast, because the classification algorithm of KNN is relatively simple and relies too much on distance and feature selection, it is difficult to achieve better performance. The input of HMM-based methods is frame-level features. However, since emotional information is usually hidden in long-term speech segments, the effect of HMM model in speech emotion recognition is not as significant as it is in the field of speech recognition. The performance of methods based on GMM and SVM is generally better than the first two methods. Due to the powerful capabilities of SVM classifiers, SVMs alone can usually achieve good performance 11 Before the advent of deep learning methods, SVM-based recognition systems have been the mainstream

systems in speech emotion recognition, until today they are also compared as baseline systems in many studies.

However, most of the problems faced by these methods are that they are highly sensitive to features, so most of the research needs to go through a complex feature selection process before training the classifier. Therefore, the process of emotional discrimination information extraction is to a certain extent, feature selection through algorithms. In addition, most of the non-deep learning algorithms are difficult to achieve end-to-end training, which introduces a lot of artificial interference, which adversely affects the practical application of speech emotion recognition systems.

1.3.2Deep learning-based speech emotion recognition method

In recent years, due to the improvement of computer computing power and in-depth research on deep learning, it has become possible to train more complex neural networks. The method of speech emotion recognition using deep learning methods has gradually come into the field of researchers.

Inspired by the excellent performance of convolutional neural networks (CNNs) in image recognition and their good ability to capture high-level representations in the spatial domain, some researchers have started using spectrograms as input data for speech signals. It represents the frequency points on the spectrogram in the form of image pixels, and the vertical and horizontal coordinates represent time and frequency respectively, which not only represent the frequency-time characteristics of speech, but also reflect language characteristics of the speaker. The appearance of the spectrogram makes it possible to avoid the singularity of the parameters of the algorithm analysis, and becomes a very important tool in the field of speech emotion recognition. In recent years, many researchers have combined deep learning with spectrograms to extract relevant spectral features from spectrograms and used them for speech emotion recognition, and achieved good results. At the same time, since the neural network cannot only be used as a classification model, but also can directly perform feature extraction, combined CNN and LSTM to extract features from spectrograms for sentiment classification 12. They divided the raw speech signal into speech segments no longer than 3 seconds and extracted a spectrogram of each speech segment. Then, a 3-layer convolutional neural network plus LSTM method is used to extract deep speech

emotion features from speech emotion in the IEMOCAP database, and the average recognition rate is determined to be 66%. CapsNets 13 fully considers the spatial relationship of actions in speech features, and the routing algorithm recursively associated with the capsule layer can consider time information. Therefore, Xixin Wu et al. 14 proposed a CNN+GRU+SeqCap speech emotion recognition system that is more efficient than the basic system provide better results.

Using the spectrogram as input, the speech signal is first converted into a spectrogram, followed by feature extraction and classification, so the learning process is divided into two stages. Therefore, some researchers prefer to use waveform signals as input, which do not require speech signal conversion and can be trained directly. George et al. 15 proposed a convolutional recurrent neural network that operates on raw waveform signals to perform the end-to-end task of spontaneous emotion prediction based on speech data. Neumann et al.16 proposed a convolutional neural network with an additional attention mechanism for emotion recognition. And Mirsamadi et al. 17 adopted a local attention based RNN.

Deep learning methods have powerful feature expression capabilities and can learn some attributes that humans do not know, which may include the aforementioned prosodic features, sound quality features, etc. These unknown properties are more useful in identifying speech emotion. Teng Zhang and Ji Wu 18 proposed a recurrent neural network to extract i-vector features from speech. After matching features with sentiment labels using the proposed recurrent neural network, and then using a database of real call center conversations, an F-Score result of 48.9% was obtained, which was 10.6% higher than the SVM model using only prosodic symbols.

2 Basic Theory of Speech Emotion Recognition

The research in the field of speech emotion recognition has a history of 40 years. After continuous research and exploration by a large number of researchers around the world, the development of speech emotion recognition is remarkable, and various technical means emerge one after another. But no matter how it is developed, a complete speech emotion recognition system is composed of various parts as shown in Figure 2.1.

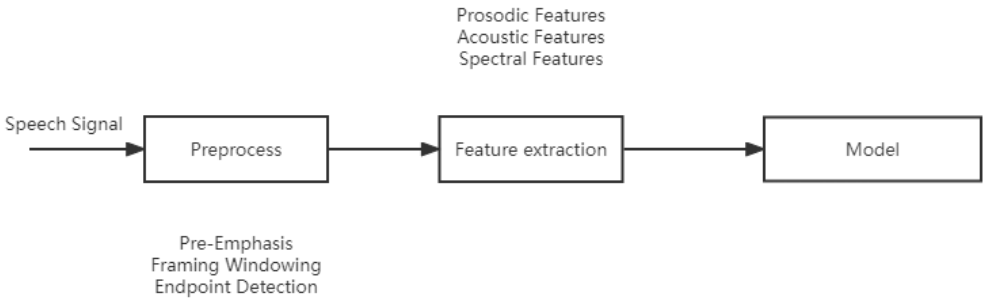

Figure 2.1 Speech emotion recognition process

In speech emotion recognition, the speech signal is usually first correlated with the speech signal through pre-processing techniques (including pre-emphasis, frame-by-frame windowing, etc.). Then the required emotional features (such as prosodic features or spectral features, etc.) are extracted from the processed speech. Finally, a recognition model that can be recognized is selected, and the features are sent to the model for training, and then the fully trained recognition model is used to calculate certain rules, and finally the recognition result is obtained.

2.1 Speech Preprocess

Since the voice signal may be affected by various factors such as external factors or hardware factors, the received voice signal must be preprocessed to reduce the pollution of the voice signal. There are many techniques for preprocessing, which will be briefly introduced in this section.

2.1.1Pre-Emphasis

When people speak, the speech is interfered by some external factors, which will cause the signal above 8kHz to show a 6db/frequency range drop. Therefore, when calculating the spectrum of the speech signal, the spectrum value corresponding to the high frequency end is relatively small, which makes it difficult to obtain the high frequency part. In order to solve such problems, pre-emphasis techniques are proposed. In fact, the pre-emphasis technique actually passes a high-pass filter, as shown in the following equation (2.1):

$$H(Z) = 1 - \mu Z^{-1} \tag{2.1}$$

In formula (2.1), μ is the pre-emphasis coefficient, and its range is [0.9, 1], and generally takes 0.97. After pre-emphasis, the entire spectrogram will become flat and there will be no fluctuation.

2.1.2Framing and Windowing

The speech signal is a one-dimensional continuous signal, and usually the speech signal will appear "short-term characteristics", which will change with time and have a stable periodicity within a very short period of time (usually 10ms-30ms). And due to the macroscopic instability of the speech signal, it can be divided into some short signals of 10-30ms for processing. These short signals are called frames. Usually the number of frames per second is about 33-100 frames. The above 10-30ms corresponds.

Framing methods are generally divided into two types: continuous framing and overlapping framing. Framing cannot disconnect the speech signal since the speech signal is continuous. In order to make the transition between frames smooth and ensure their continuity, overlapping frame processing is generally adopted, that is, there needs to be a partial overlap between each frame, and the entire overlap is generally referred to as frame shift. The frame shift is generally 0-1/2 of the frame length. As shown in Figure 2.2:

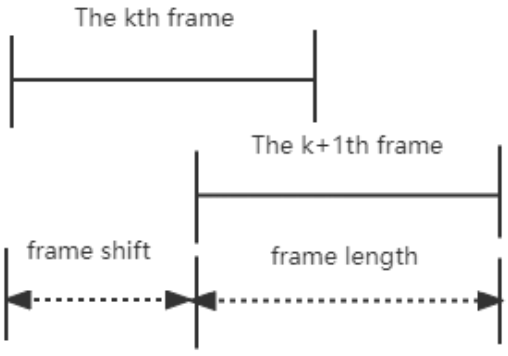

Figure 2.2 Frame shift diagram

In order to meet the requirements of framing, the voice signal is still a continuous signal, so it is necessary to multiply a data of the same length, and this data is the window function. Therefore, the windowing process is to convolve the signal with the speech signal $s(n)$ through the appropriate window function ω(n), so as to obtain the windowed speech signal $s_\omega(n)$, $s_\omega(n) = s(n) * \omega(n)$.

There are two most commonly used windows, one is the rectangular window and the other is the Hamming window. Its introduction is as follows:

（1）Rectangular Window

The rectangular window can be expressed in the time domain as:

$$w(n) = \begin{cases} 1\ , & 0 \le n \le (N-1) \\ 0, & others \end{cases} \tag{2.2}$$

In the frequency domain, it can be expressed as:

$$W_R\left(e^{i\omega T}\right) = e^{-j\left(\frac{N-1}{2}\right)\omega T} \frac{\sin\left(\frac{\omega NT}{2}\right)}{\sin\left(\frac{\omega T}{2}\right)} \tag{2.3}$$

（2）Hamming Window

The representation of the Hamming window in the time domain is:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), n = 1,2,\dots.,N-1 \\ 0, \qquad\qquad others \end{cases} \tag{2.4}$$

In the frequency domain, it can be expressed as:

$$W_H\left(e^{j\omega}\right) = 0.54W_R(\omega) + 0.23\left[W_R\left(\omega - \frac{2\pi}{N-1}\right) + W_R\left(\omega + \frac{2\pi}{N-1}\right)\right] \tag{2.5}$$

where $W_R(\omega)$ represents the amplitude-frequency characteristic function of the rectangular window function.

In the process of speech signal analysis, the selection of the window function W has a great influence on some parameter characteristics of the speech signal, so it is very important to choose a suitable window function. In this paper, the Hamming window is selected as the window function.

2.1.3Fast Fourier Transform

Perform fast Fourier transform on the windowed signal, convert it to the frequency domain, and obtain the spectrum of the signal. The formula is shown in (2.6):

$$Y(l) = \sum_{m=0}^{M-1} y(m)e^{-i\left(\frac{2\pi}{M}\right)ml}(l = 0,1,\dots,M-1) \tag{2.6}$$

Where $Y(l)$is the frequency domain sample, $y(m)$ is the time domain sample, M is the size of the fast Fourier transform, and the value of m is the same as l, from 0 to $M-1$.

2.2 Emotion

To study the speech emotion recognition system, we first need to define or quantify emotion. Due to the complexity of human emotions, the classification of emotions has always been a hot issue in the field of emotions, and there is still no fully unified standard. At present, the mainstream emotion description models can be divided into dimensional and discrete models.

Discrete emotion describes emotion as a single, independent label in the form of an adjective. Mainly based on concentrated emotions, other emotions are formed by combining certain basic emotions to varying degrees. The most commonly used concept is the six major categories, that is, there are six basic emotions, but these six basic emotions have not been unified by researchers. The following is a list of some of the definitions of emotion by different scholars.

Table 2.1 Definitions of basic emotions by different scholars 19

Person	Basic Emotions
Ekman, Friesen &Ellsworth(1982)	anger, disgust, fear, joy, sadness, surprise
Arnold(1960)	anger, aversion, courage, dejection, desire, despair, dear, hate, hope, love, sadness
Frijda (personal communication, September 8, 1986)	desire, happiness, interest, surprise, wonder, sorrow
Gray(1982)	rage and terror, anxiety, joy
Izard (1971)	anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
Mowrer(1960) Oatley & Johnsonlaird (1987)	pain, pleasure, anger, disgust, anxiety, happiness, sadness
Panksepp (1982)	expectancy, fear, rage, panic
Plutchik (1980)	acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise

Tomkins(1984)	anger, interest, contempt, disgust, distress, fear, joy, shame, surprise
Watson(1930)	fear, love, rage

Among these numerous studies, the American psychologist Ekman, who proposed the following six basic emotion types, proposed the most widely used emotion type: happiness, anger, fear, sadness, surprise and disgust 20. Most public emotional speech databases are built based on these 6 basic emotions, so these 6 basic emotion types are also called original emotion types.

Another widely recognized affective model is the dimensional affective model 21. Some psychologists and artificial intelligence researchers believe that there are no strict boundaries between emotions and should be represented by a continuous spatial dimension. They believe that several features that describe emotions should be selected for dimensional modeling, so as to further describe emotions. Among them, there are two typical models. The first is the emotional two-dimensional model, and the two dimensions are Valence and Arousal. As shown in Figure 2.3, it is Lange's two-dimensional sentiment classification model22, which uses a variable dimension to describe whether an emotion is positive or negative, which is between unpleasant and pleasant.

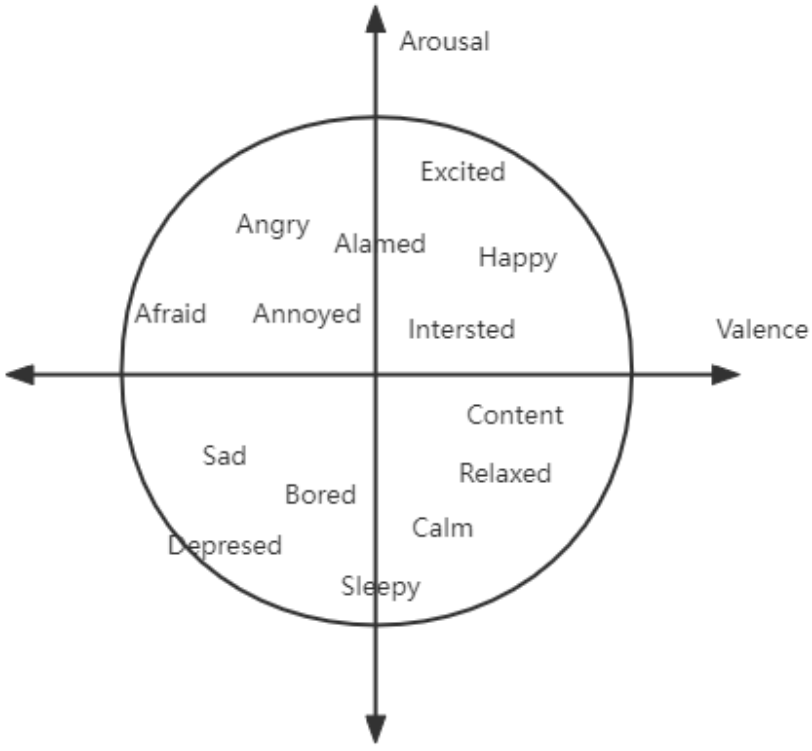

Figure 2.3 Lang's 2D Emotion Map

The second dimensional affective model is the three-dimensional PAD affective model. This model is similar to the two-dimensional emotional model, which believes that human emotions can be described from three aspects: pleasure, activation and priority. Among them, the pleasure degree is similar to Valence, which is used to describe whether the emotion is positive or negative, which is expressed as P. Activation is similar to Arousal but slightly different, and is used to indicate the intense state of emotion, which is related to the degree of collective energy activity corresponding to the emotion, denoted by A. The degree of preference is used to indicate the influence state of the emotional subject on the object and the situation, and it is denoted by D.

Both of the above-mentioned two-dimensional emotions have good emotional expressiveness and can perfectly show the degree of difference between emotions. However, dimensional emotion is too complex to make or obtain emotional datasets in a targeted manner. Therefore, the selection of emotional models is mainly based on the selection of the most appropriate emotional models according to different scenarios. This paper mainly selects discrete emotion as the experimental model.

2.3 Emotion Database

Today, many professional research teams have produced professional datasets for different language environments to adapt to emotion recognition research and work in different environments and scenarios. The datasets used in this study are shown below, and as described in the previous section, this paper mainly focuses on the research and elaboration of discrete emotion models.

（1） Surrey Audio-Visual Expressed Emotion(SAVEE) database

The SAVEE database contains recordings of seven different emotions of anger, disgust, fear, neutral, happiness, sadness and surprise from four native English-speaking men and women aged 27 to 31 from the University of Surrey (University of Surrey) document. In addition to neutral emotions, each emotion has a total of 15 sentences, and 30 neutral sentences at the same time. A total of 480 audio files.

（2） RAVDESS Emotional Speech Library 23

The database is available in audio-only, video-only, and audio-video formats, and contains 24 professional actors (12 women, 12 men), speaking in North American accents. Emotions include 7 expressions including calm, happy, sad, angry, fearful, surprised and disgusted, each generated at two levels of emotional intensity (normal, intense), with an additional neutral expression .

（3） EMO-DB database24

EMO-DB is a speech database established by the Technical University of Berlin. This database selects 5 male and 5 female actors to record 10 sentences, each of which is neutral, angry, scared, happy, sad, disgusted, bored. Read aloud for 7 emotions. In the process of recording, the actors are guided to record real emotions. In the end, a total of 535 sentences were retained, with 233 male sentences and 302 female sentences. The sampling rate was 48 kHz and the quantification was 16 bits.

（4） CASIA Chinese Speech Database 25

CASIA is a Chinese corpus constructed by the Institute of Automation, Chinese Academy of Sciences, which contains 9600 pure speeches. It was recorded by 2 males and 2 females in a pure recording environment (signal-to-noise ratio is about 35dbs), and it was deduced through certain emotions and situational guidance. The database contains 6 emotions: angry, happy, scared, sad, surprised, and neutral. Each of them recorded 400 voices for each emotion. Among the 400 voices, 300 are voices with the same sentence content expressed through different emotions, and the remaining 100 are sentences with different content and corresponding emotions.

2.4 Speech emotion feature

Feature extraction of speech signal is the most important part of speech emotion recognition research. Its function is to use feature extraction tools to extract the hidden feature data from the preprocessed speech audio signal through certain algorithms, and then send it to the corresponding classification model for model training and speech emotion classification. The most important part of the emotion recognition process.

In general, the emotional features in speech emotion recognition research can be divided into the following three types 26.

2.4.1Prosodic Features

Prosody features mainly include the structural expression features of speech in speech signals, which are phonetic features that are different from the semantics of speech itself, and generally include time-related features, fundamental frequency-related features, and energy-related features. As shown in the table below, the classification relationship of various features and some commonly used features.

Table 2.2 Classification of prosody features

Classification / Features	Feature classification
Time-related features	Speech rate, etc.
Fundamental frequency-related features	Mean square error, Fundamental frequency, etc.
Energy-related features	Amplitude average rate of change, Short-term average energy, etc.

It can be seen from the above table that prosody features have rich types of features, and such features are also widely used in the field of speech emotion recognition at this stage.

2.4.2Spectral Correlation Features

At present, the mainstream features based on spectral correlation are Linear Prediction Cepstral Coefficients (LPCC) and Mel Frequency Cepstral Coefficients (MFCC). During the research process, it was found that LPCC is derived from the power spectrum after smooth processing, while MFCC is more suitable for the auditory characteristics of human ears, and the recognition effect is better. Therefore, only the MFCC is mainly introduced here.

MFCC is a short-time Fourier transform of the original sound, through a set of Mel filter banks, its role is to change the linear frequency to a nonlinear frequency, so as to suit human hearing. Finally, the discrete cosine transform is used to calculate the cepstrum, and the correlation in the features is removed at the same time. The MFCC represents the speech signal as the short-term power spectrum of the speech. MFCC can give better frequency resolution in the low frequency region. Therefore, it is suitable for all types of signals and has no noise effects.

2.4.3Sound Quality Features

The sound quality characteristics are determined by the physical properties of the sound, which can use characteristics such as vibration, harmonic-to-noise ratio, etc. to distinguish emotions. Therefore, the experimenters believe that the sound quality characteristics have a close relationship with emotional expression.

2.5 Result Evaluation

Since the discrete emotion recognition model corresponds to the classification problem in machine learning, the evaluation indicators for the experimental results of speech emotion recognition can be: accuracy, precision, recall, and F1-score. The formulas for these four metrics are shown below.

(1) Accuracy:

$$Accuracy = \frac{T_P+T_N}{T_P+T_N+F_P+F_N} \tag{2.7}$$

(2) Precision:

$$Precision = \frac{T_P}{T_P+F_N} \tag{2.8}$$

(3) Recall:

$$Recall = \frac{T_p}{T_P+F_N} \tag{2.9}$$

(4) F1-score:

$$F1_{score} = 2\frac{Precision*Recall}{Precision+Recall} \tag{2.10}$$

In the above formula, TP is the number of True Positives, FP is the number of False Positives, TN is the number of True Negatives, and FN is the number of False Negatives.

2.6 Chapter Summary

This chapter mainly introduces the basic knowledge of speech emotion recognition. First, it introduces the basic process of speech emotion recognition, including speech signal preprocessing, feature extraction and emotion recognition. The preprocessing process of the speech signal is explained in detail. After the preprocessing of the speech signal, the feature extraction can be carried out. The feature of the speech signal is the most important part of the speech emotion recognition, and the quality of the feature directly affects the accuracy of the recognition. After that, the

classification of emotion models is introduced, including discrete and dimensional speech emotion models, and a simple analysis is made. Then, some details of the emotion dataset used in this experiment are introduced. Finally, the evaluation criteria for the model in this experiment are introduced, including accuracy, recall, precision and F1-score. This chapter builds the research foundation for the following research.

3 Emotion Recognition Based on MFCC Features

In the research of speech emotion recognition, the selection of speech emotion features is always a crucial part. According to the current research status and the comparison of the advantages and disadvantages of the effect, the MFCC feature is more direct to the representation of speech emotion, and it performs well in the field of speech emotion recognition. However, there are two types of 13-dimensional Mel-frequency cepstral coefficients and 39-dimensional Mel-frequency cepstral coefficients that have good performance in the recognition effect. Therefore, the main research content of this chapter is to build a VGG-like one-dimensional convolutional neural network to conduct speech emotion analysis on the 13-dimensional Mel frequency cepstral coefficients and the 39-dimensional Mel frequency cepstral coefficients extracted from the speech signal respectively. Identify the experiments, compare the performance differences between the two, and analyze the experimental results. Then, the speech features with better performance among the two are selected as the features of subsequent experiments.

3.1 Mel Frequency Cepstral Coefficients (MFCC)

MFCC is a feature that accurately describes the variation of vocal tract deformation in the short-term power spectrum of speech. It is a feature that is more suitable for the human hearing mode, because the human ear has a certain difference in the auditory inspiration in sensing sound waves of different frequencies, and the MFCC has a linear change in the logarithmic energy spectrum of the nonlinear Mel scale. The characteristics are closer to the human auditory system than the spectrum, so MFCC can better characterize the sound signal from multiple angles.

3.1.1Feature extration

The MFCC feature is the cepstral data obtained by performing a series of calculations under the Mel scale after a series of processing of the speech signal. The extraction process is: pre-emphasis, framing, windowing, fast Fourier transform, Mel filtering, logarithmic operation, discrete cosine transform, and finally the MFCC feature is obtained.

Among them, pre-emphasis, framing and windowing have been described in detail above, and the subsequent steps are mainly described in this chapter.

（1） fast Fourier transform(FFT)

After the speech signal is preprocessed, because the characteristics of the time domain are not obvious, the speech signal is converted from the time domain to the frequency domain signal through the fast Fourier transform. Compared with the time domain features, the frequency domain signal is more star-like, and its energy distribution can be observed through the energy spectrum, and the power spectrum after the frequency domain is modulo squared can be effectively analyzed. Therefore, many researchers also use the spectrogram as one of the research characteristics. The function expression for conversion is (3-1):

$$S_i(k) = \sum_{n=1}^{N} s_i(n)\omega(n)e^{\frac{-j2\pi kn}{N}}, 0 \le \mathrm{k} \le \mathrm{K} \tag{3-1}$$

（2） Mel filtering

Mel filtering and the acoustic frequency (f) of the human ear exhibit a nonlinear relationship, which can be approximated by (3-2):

$$Mel(f) = 2595 * \lg\left(1 + \frac{f}{700}\right) \tag{3-2}$$

According to the characteristics of the human ear to the speech signal, when extracting features, the low-frequency part should be denser, and the high-frequency part should be relatively sparse. Mel filtering is implemented by triangular filtering. The energy spectrum after passing the FFT is sent to a Mel-scale triangular filter bank (generally consists of 26 triangular filters), as shown in Figure 3.1.

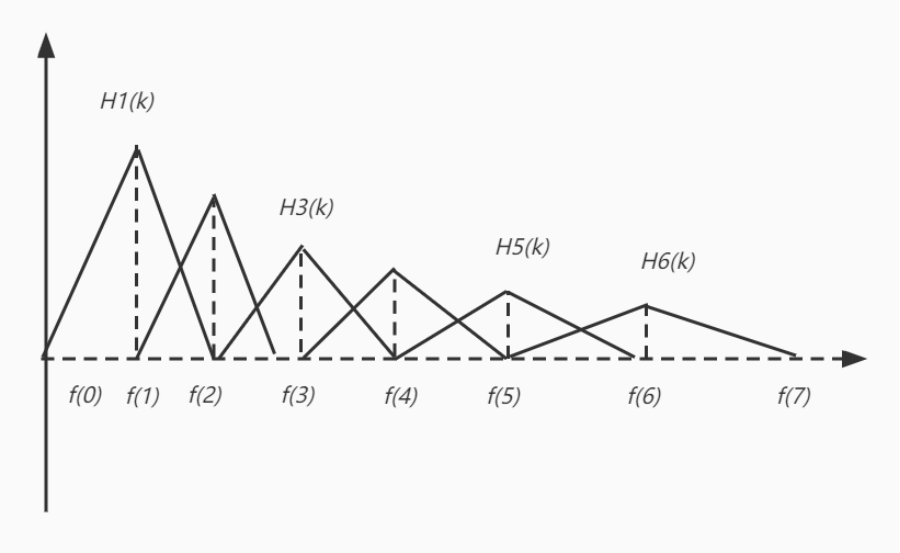

Figure 3.1 Schematic of a triangular filter bank

The triangular filter expression is as follows:

$$H_n = \begin{cases} 0\,, k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)}, f(m-1) \le k \le f(m) \\ 1,\ k = f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)}, f(m) \le k \le f(m+1) \\ 0, k > f(m+1) \end{cases} \tag{3-3}$$

And $\sum_{m=0}^{M-1} H_m(k) = 1,\ f(m)$ is the center frequency of the mth triangular filter.

A triangular filter will smooth out the frequencies, making the original formants more pronounced. Therefore, the triangular filter will reduce the interference of the pitch level on the emotional characteristics of the MFCC, and avoid the objective influence of the pitch on the emotion.

（3） log energy

After the smoothing of the triangular filter, the logarithmic calculation of the result of each filter is required, and the logarithmic energy can be obtained from the formula (3-4).

$$s(m) = \ln\left(\sum_{k=0}^{N-1} |X_a(k)|^2 H_m(k)\right), 0 \le m \le M \tag{3-4}$$

（4） Discrete Cosine Transform

After the logarithmic energy is obtained, a discrete cosine transform can be performed to obtain the MFCC, the formula is as (3-5)

$$C(n) = \sum_{m=0}^{N-1} s(m) \cos\left(\frac{\pi n(m-0.5)}{M}\right), n = 1,2,3, \dots L \tag{3-5}$$

M is the number of triangular filters, and L is the order of Mel cepstral coefficients.

In general, L is usually 12, but in order to ensure data integrity, a frame of logarithmic energy spectrum and a 12-dimensional MFCC will be spliced to obtain a 13-dimensional MFCC.

（5） High-dimensional MFCCs

The 12-dimensional and 13-dimensional MFCC features obtained above are only the static features of the original voice signal, and the differential spectrum obtained by the low-dimensional MFCC through differential can describe the corresponding dynamic features of the voice signal. Therefore, high-dimensional MFCC features (including first-order and second-order differences) are generated, and the difference parameter extraction formula is shown in (3-6).

$$d(t)\begin{cases} C_{t+1}-C_t \ , t<K \\ \dfrac{\sum_{k=1}^{K} k(C_{t+k}-C_{t-k})}{\sqrt{2\sum_{k=1}^{K} k^2}}, \ others \\ C_t-C_{t-1}, \ \ t \geq Q-K \end{cases} \tag{3-6}$$

3.2 Introduction to Convolutional Neural Networks

Convolutional Neural Network (CNN) is a kind of feedforward neural network with deep structure that includes convolutional computation and is one of the representative algorithms of deep learning 27. Convolutional neural network can perform shift-invariant classification of input information according to its hierarchical structure 28, so it is also called "Shift-Invariant Artificial Neural Networks". According to the characteristics of CNN, it is often used for further feature extraction and representation, in order to better complete the training of deep learning.

The structure of CNN is mainly divided into: input layer, hidden layer and output layer. The input layer is used to receive multi-dimensional feature data in the form of tensors, and the output layer is considered to be classified, and it is usually connected with the hidden layer and the input layer by a fully connected layer. Its structure is similar to the traditional artificial neural network, the only difference is that the hidden layer of CNN introduces the concept of convolution layer. The hidden layer of CNN mainly includes the following parts:

（1） Convolutional Layer

The convolutional layer receives the data of the input layer and performs feature extraction on the input data. Of course, the feature extraction here is not the same as the feature extraction for speech data in the previous article. In the convolution layer, it mainly refers to the feature extraction of feature data using convolution calculation, which does not change the mapping between input and output. relationship, and its mapping relationship is shown in (3-7):

$$x_j^l = f_c(\sum_{i \in M_j} x_i^{l-1} * k_{i,j}^l + \theta_j^l) \tag{3-7}$$

$x_j^l$is the jth feature set of the lth convolutional layer, x_i^{l-1}represents the ith feature set of the $l-1$th convolutional layer, $k_{i,j}^{\ l}$represents the convolution kernel between the two feature sets, and * represents the two-dimensional Convolution processing, θ_j^l represents the dimensional addition bias.

（2） Pooling Layer

After the input data is extracted by the features of the convolution layer, it will enter the pooling layer for further screening of data information and the second selection of features, and at the same time, the features obtained by convolution are subjected to dimensionality reduction processing. The pooling function set in the pooling layer is shown in formula (3-8), and the result of a single point in the feature map is replaced by the statistics of the feature map of adjacent regions.

$$x_j^l = f_p(\beta_j^l down(x_i^{l-1}) + \theta_j^l \tag{3-8}$$

$f_p(\cdot)$ is the activation function of the pooling layer, and $down(\cdot)$ is the pooling method used in layers $l-1$ to l. Pooling methods are usually divided into two types, the first is the mean pooling method, and the second is the maximum pooling method. In equation (3-8), β_j^l represents the multiplication bias, and θ_j^lrepresents the additive bias.

（3） Fully Connected Layer

The fully connected layer is equivalent to the hidden layer in the traditional feedforward neural network. The fully connected layer is located in the last part of the hidden layer of the convolutional neural network and only passes signals to other fully

connected layers. It usually uses the Softmax model to solve the multi-classification problem. The loss function of Softmax is shown in (3-9):

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m}\sum_{j=1}^{k} l\{y^{(i)} = j\} log \frac{e^{\theta_j^l}}{\sum_k e^{\theta_k^l}} \tag{3-9}$$

θ_j^l represents the input of the jth neuron node in the lth layer, $\sum_k e^{\theta_k^l}$ represents the sum of the inputs of all neuron nodes in the whole l layer, and $\frac{e^{\theta_j^l}}{\sum_k e^{\theta_k^l}}$ is the input of the jth neuron in the lth layer. $l(\cdot)$ is a judgment function. When the result in the parentheses is true, the result of the function is 1, and when the result in the parentheses is false, the result of the function is 0. In order to prevent the local optimization of $J(\theta)$, the $\frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=0}^{n}\theta_{i,j}^2$ weight decay idea is introduced to punish the excessively large parameters in the training process. The specific expression is shown in (3-10):

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m}\sum_{j=1}^{k} l\{y^{(i)} = j\} log \frac{e^{\theta_j^l}}{\sum_k e^{\theta_k^l}} + \frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=0}^{n}\theta_{i,j}^2 \tag{3-10}$$

In addition to the above three network layers, there are also batch normalization layers (BN) and dropout layers that are usually used after convolutional layers. The role of the BN layer is that the CNN is more sensitive to the data near the zero value, and the BN layer can reset the offset data to the range of 0-1, thereby obtaining more effect and better convergence. The function of the Dropout layer is to discard some neurons, so that some neural networks do not participate in training to achieve the purpose of suppressing overfitting and improving network stability to a certain extent. These two network layers are means of regularization and can be used to prevent overfitting.

3.3 CNN speech emotion recognition network model

This section is mainly to test the performance of 39-dimensional MFCC and 13-dimensional MFCC in the neural network of CNN architecture, so a relatively basic VGG-like CNN network is used for experiments. The experimental results will play a certain role in the selection of subsequent speech features.

The schematic diagram of the CNN network structure of the VGG-like structure used in this experiment is shown in Figure 3.2.

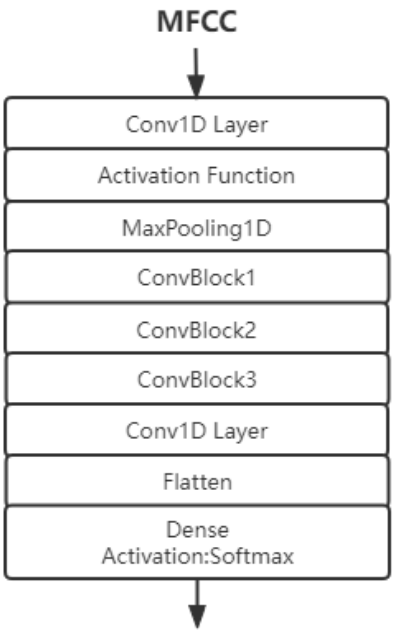

Figure 3.2 CNN model

The ConvBlock in Figure 3.2 is a convolution block, and its composition is shown in Figure 3.3.

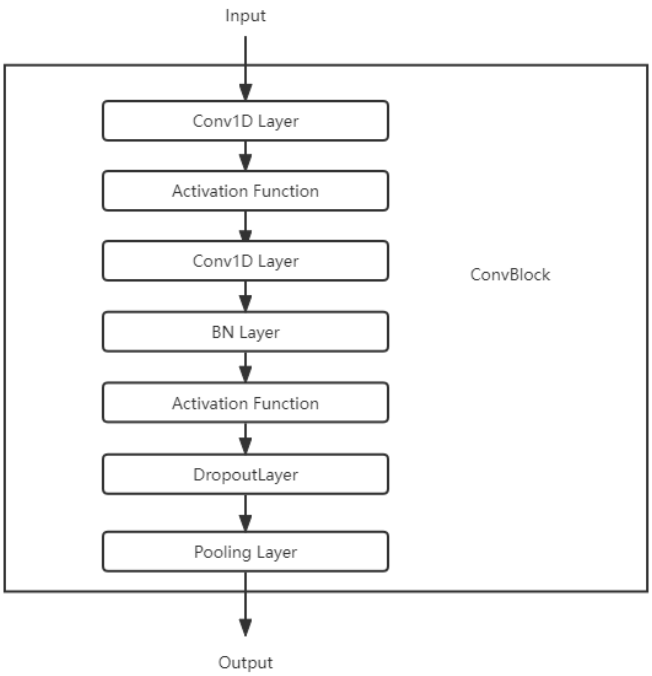

Figure 3.3 ConvBlock

Since the processing of speech is carried out according to the frame, the corresponding feature vector can be obtained after each frame is processed. Assuming

that a piece of speech has N frames, after processing this piece of speech, an MFCC feature matrix with M rows and N columns can be obtained (M represents the feature dimension, and N represents the number of frames). Therefore, in this CNN model, a 1-dimensional convolution Conv1D layer is used, the activation function is ReLu, and the pooling layer is MaxPooling1D. The construction of a simple speech emotion recognition network is realized through the CNN network similar to VGG structure. A comparison experiment of speech emotion recognition is carried out on the 13-dimensional MFCC and the 39-dimensional MFCC.

3.4 Experimentation and Analysis

3.4.1Experimental environment

All the proposed models and improvements in this paper are experimented in the environment shown in Table 3.1.

Table 3.1 Experimental environment

Item	Version
CPU	Intel(R) Core(TM) i7-10510U
CPU Frequency	1.80GHz
RAM	8.00G
Graphics Card	NVIDIA GeForce MX250
Compute Capability	6.1
Operating System	Windows10
Develop Software	Anaconda(jupyter notebook)
Development Environment	Python 3.7.0
Frame	Tensorflow 1.13.0
Packages	Keras 2.2.4, numpy 1.20.3, librosa 0.8.1, pandas 0.23.4

In this experiment, the dataset is selected as the SAVEE sentiment dataset. The selected data set is randomly divided into training set and test set according to the ratio of 8:2, and cross-validation is performed.

The data for the network model are shown in Table 3.2.

Table 3.2 Network parameters

Item	Parameters
Conv1D Layer	Filters =256, kernel_size=5
Activation Function	ReLu
Maxpooling1D	4
Dropout	0.25
ConvBlock1	Filters=256,kernel_size=5
ConvBlock2	Filters=128,kernel_size=5
ConvBlock3	Filters =64,kernel_size=5

The training optimizer uses Adam, and its parameters are: lr=0.001, beta_1=0.9, beta_2=0.99. The loss function used the categorical cross-entropy function.

3.4.2Code

The librosa library is used here to extract MFCC features, and the code is as follows:

```
data = pd.DataFrame(columns=['feature'],dtype=object)
for i in tqdm(range(len(Savee_df))):
    X, sample_rate = librosa.load(Savee_df.Path[i], res_type='kaiser_fast',duration=input_duration,sr=22050*2,offset=0.5)
#     X = X[10000:90000]
    sample_rate = np.array(sample_rate)
    mfccs = np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=13), axis=0)
    feature = mfccs
    data.loc[i] = [feature]
```

Figure 3.3 13-dimensional MFCC extraction code

```
data = pd.DataFrame(columns=['feature'],dtype=object)
for i in tqdm(range(len(Savee_df))):
    X, sample_rate = librosa.load(Savee_df.Path[i], res_type='kaiser_fast',duration=input_duration,sr=22050*2,offset=0.5)
#     X = X[10000:90000]
    sample_rate = np.array(sample_rate)
    mfccs = np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=39), axis=0)
    feature = mfccs
    data.loc[i] = [feature]
```

Figure 3.4 39-dimensional MFCC extraction code

3.4.3Results and Analysis

For the results of this experiment, multiple evaluation criteria were selected. In addition to the accuracy, precision, recall and F1-score mentioned above, there is also a confusion matrix.

Confusion matrix is an evaluation benchmark that directly expresses experimental data results in matrix form. Since the two dimensions of the confusion matrix are represented as the true label category and the predicted result category, the confusion matrix is usually represented as a two-dimensional N-order matrix with N rows and N columns. The advantage of the confusion matrix is that the recognition

situation of each classification can be observed intuitively, and it is easy to draw more accurate judgments and make corresponding adjustments quickly.

（1） Results

The average accuracy, precision, recall and F1-score of its 13-dimensional MFCC and 39-dimensional MFCC are shown in Table 3.3.

Table 3.3 Comparison of experimental standards

Category Standard	13-dimensional MFCC	39-dimensional MFCC
Accuracy	85.94%	84.90%
Precision	87.63%	84.81%
Recall	84.90%	84.38%
F1-score	86.23%	84.59%

Among them, the mixture matrix of the verification set of speech emotion recognition performed by 13-dimensional MFCC features is shown in Table 3.4.

Table 3.4 13-dimensional MFCC mixing matrix

Predicted / True	neutral	sad	angry	surprise	disgust	fear	happy	Precision
neutral	22	0	0	0	0	0	2	91.67%
sad	0	22	0	0	0	2	0	91.67%
angry	0	0	22	0	0	0	2	91.67%
surprise	2	0	0	22	0	0		91.67%
disgust	2	3	2	0	39	2	0	81.25%
fear	3	0	1	0	2	18	0	75.00%
happy	0	0	2	2	0	0	20	83.33%
Recall	75.86%	88.00%	81.48%	91.67%	95.12%	81.81%	83.33%	84.90%/87.63%

The mixture matrix for speech emotion recognition on 39-dimensional MFCC features is shown in Table 3.5.

Table 3.5 39-dimensional MFCC mixing matrix

Predicted / True	neutral	sad	angry	surprise	disgust	fear	happy	Precision
neutral	21	0	0	0	0	1	2	87.50%
sad	0	20	0	0	3	1	0	83.33%
angry	0	2	20	0	0	2	0	83.33%
surprise	0	0	0	22	0	2	0	91.67%
disgust	4	2	0	2	40	0	0	83.33%
fear	0	0	0	0	4	20	0	83.33%
happy	0	0	2	0	2	0	20	83.33%
Recall	84.00%	83.33%	90.91%	91.67%	81.63%	76.92%	90.91	84.38%/84.81%

（2） Analysis

First, for the results shown in Table 3-3. From the four evaluation benchmarks of accuracy, precision, recall, and F1-score, it is obvious that the 13-dimensional MFCC may be more suitable for emotion recognition than the 39-dimensional MFCC. Except for the recall rate of these four indicators, the other three evaluation benchmarks, the 13-dimensional MFCC is 1-3% higher than the 39-dimensional MFCC.

It can be seen from Table 3-4 that the 13-dimensional MFCC has higher recognition accuracy in the four emotions of neutral, sad, angry, and surprise, but has lower recognition accuracy for disgust, fear, and happy. At the same time, the recall rate of sad, surprise and disgust is higher, while the other emotions are lower.

From the experimental data of the 39-dimensional MFCC in Table 3-5, it can be clearly seen that the accuracy of the seven emotion recognition is not very different. Compared with the 13-dimensional MFCC, the accuracy of various emotions is more average.

In terms of the recognition accuracy of various emotions, the 39-dimensional MFCC performs better on the SAVEE dataset than the 13-dimensional MFCC, and can identify various emotions more accurately. The reason should be the advantages brought by the dynamic characteristics represented by the first-order difference spectrum and the second-order difference spectrum in the 39-dimensional MFCC. However, from the overall accuracy, precision, recall and F1-score of the four evaluation benchmarks, the 13-dimensional MFCC is better. It may be due to the better

representation of the static features characterized by the 13-dimensional MFCC. Therefore, the next experiments will select 13-dimensional MFCC as the feature of speech emotion recognition.

3.5 Chapter Summary

In this chapter, Mel Frequency Cepstral Coefficients (MFCC) are first introduced, including the basic knowledge and extraction process of MFCC. Use Python's librosa package to extract MFCC features of speech signals, and then conduct experiments on 13-dimensional MFCC and 39-dimensional MFCC through a VGG-like CNN neural network. Through the experimental results, it is concluded that the static features of the 13-dimensional MFCC contribute more to speech emotion recognition than the dynamic features of the 39-dimensional MFCC, and the static features of the 13-dimensional MFCC are more powerful in representation. In view of this conclusion, 13-dimensional MFCC is selected as the emotion recognition feature for the next experiment.

4 Speech Emotion Recognition Based on MS-ResNet

With the deepening of research in the field of deep learning, more and more excellent deep learning neural network models have been proposed one by one. For the model improvement and innovation of neural networks, there are two main ideas for building network models at this stage, one is to deepen the number of network layers, and the other is to widen the network structure. The representatives of these two methods are ResNet 29 and GoogLeNet 30. These two models have outstanding performances in image processing, text recognition and speech recognition, so many researchers use these two models The model serves as the model basis for related research.

The main research contents of this chapter are as follows: First, build the ResNet neural network and the GoogleNet neural network, and complete the speech emotion recognition comparison experiment of the corresponding network. Second, according to the reason of MFCC as a frame-level feature, it is guessed that classifying the features of MFCC through neural networks at different scales may extract the hidden feature information in the MFCC. In this way, a MS-ResNet network with

multi-scale and residual network is proposed to conduct related speech emotion recognition experiments and obtain a certain improvement in recognition accuracy.

4.1 The theoretical foundation of deep learning

A deep neural network is essentially a special kind of multilayer perceptron (MLP). It generally includes one or even more hidden layers, which are composed of multiple restricted Boltzmann machines (RBMs) stacked on each other 31. The model is shown in Figure 4.1.

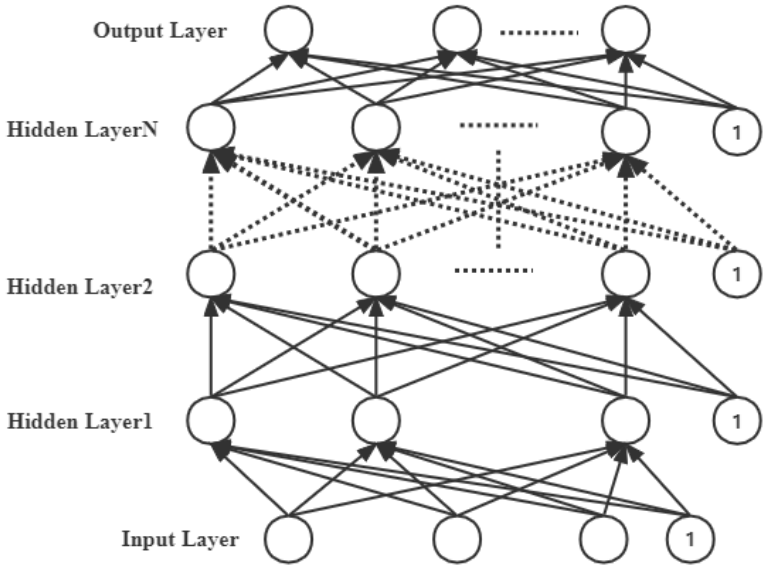

Figure 4.1 Deep Neural Network General Model

Restricted Boltzmann machine is essentially a kind of generative neural network model, its network model is usually based on energy, and has an important feature, which is randomness [31]. It can be regarded as an undirected graph model (as shown in Figure 4.2), which consists of a visible layer and a hidden layer, and can learn data from unknown distributions. Each RBM is a Markov with a two-layer structure. random field. Compared with the traditional Boltzmann machine, the restricted Boltzmann machine is very different. These differences are mainly manifested in that its neuron nodes are not connected within the same layer, but only between layers. . Usually in a restricted Boltzmann machine, the value types of the neuron nodes in the visible layer and the hidden layer are also different. In each RBM, the visible layer vector v exists in the visible layer, and the hidden layer vector h also exists in the hidden layer. According to the energy theorem, a certain relationship can be drawn

between them. The researchers as give this relationship the energy value function, which is defined as:

$$E(v,h|\theta) = -\sum_{i=1}^{n}\sum_{j=1}^{m} W_{i,j}v_i h_j - \sum_{i=1}^{n} b_i v_i - \sum_{j=1}^{m} a_j h_j \quad (4\text{-}1)$$

Where $v \in \{0,1\}^{N_v * 1}$, $h \in \{0,1\}^{N_h * 1}$, θ represent the parameters $\theta = \{W, a, b\}$ of the entire RBM model, W represents the weight matrix, a_i and b_j represent the thresholds the visible layer unit i and hidden layer unit j, respectively, and W_{ji} represents visible layer unit i and hidden layer unit. The weight value between j.

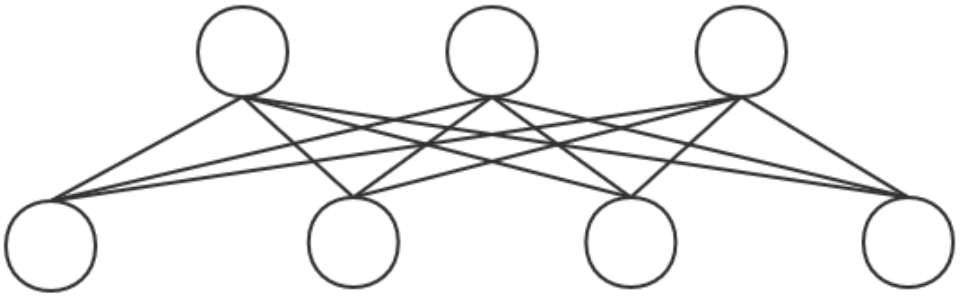

Figure 4.2 Restricted Boltzmann Machine Structure Diagram

The probability function between them can be obtained by formula (4.1):

$$p(v,h|\theta) = \frac{e^{-E(v,h|\theta)}}{Z(\theta)} \quad (4\text{-}2)$$

where $Z(\theta) = \sum_{v,h} e^{-E(v,h|\theta)}$, represents a normalization factor. From the perspective of probability theory, the expression formula of the probability distribution function $p(v|\theta)$ of the visible layer v can be obtained:

$$p(v|\theta) = \frac{1}{Z(\theta)}\sum_{h} e^{-E(v,h|\theta)} \quad (4\text{-}3)$$

From the structural characteristics of RBM, it can be concluded that when the state of the neurons in the visible layer is determined, the state of the neurons in the hidden layer is subject to conditional independence, and the probability distribution function of the hidden layer can be expressed as:

$$p(h|v,\theta) = \prod_j p\left(h_j|v,\theta\right) \quad (4\text{-}4)$$

Among them, the conditional probability can be expressed as:

$$p\left(h_j|v,\theta\right) = sigm(b_j + \sum_i w_{ij}v_{i)}, \quad j = 1,2,\ldots,m \quad (4\text{-}5)$$

$$p(v_i|h,\theta) = sigm(a_i + \sum_j w_{ij}v_j), \quad i = 1,2,\ldots,m \quad (4\text{-}6)$$

Among, $sigm(x) = \frac{1}{\exp(-x)}$ (4-7)

The training of RBM is essentially to learn the entire model parameters during the training process to make it fit the input data as much as possible. Generally, a DNN is composed of multiple RBMs through a combination of certain rules. In order to fully train the entire DNN, the RBM of each layer must be trained, and each RBM must be trained layer by layer. The basic idea of this training method is to send the hidden layer of the current layer to the next adjacent layer as the input of the visible layer of the next layer. In this way, the training process of the entire network can be completed, and each RBM in the network is superimposed according to the hierarchical relationship, and finally a complete Deep Belief Network (DBN) is constructed.

The training of DNN consists of two parts, an unsupervised training process and a supervised fine-tuning process. The purpose of unsupervised training is to maximize the fit to the internal structure of the input data. When an RBM is trained, it can be used to represent the data, and the hidden layer neuron h can be calculated for each visible layer unit v. Then use h to reconstruct the visual layer. And based on the difference between the visible layer and the reconstruction layer, the weights between the visible layer and the hidden layer are iteratively updated. Taking the obtained hidden layer as the next visible layer, and stacking layer by layer, features can be extracted layer by layer from the original input data. If the training process of the RBM is stopped, the weights of the hidden layers of the DNN will be initialized, and after training, the network parameters of the DNN can be obtained. Then, label data is added to the top layer of the DNN, and then supervised training is performed, that is, the parameters of the DNN are fine-tuned with the back-propagation algorithm (BP). This paper adopts the classical cross-entropy (CE) quasi-measure in BP. In the ideal case, the parameter optimization of DNN is carried out using Eq. (4.8):

$$J_{CE}(W, b, S) = \frac{1}{M}\sum_{m=1}^{M} J_{CE}(W, b; o^m, y^m) \quad (4\text{-}8)$$

Equation (4.8) can be called the minimization expected loss function, where M is the number of training samples, $S = \{(o^m, y^m)|0 \le m \le M\}$, besides:

$$J_{CE}(W, b; o^m, y^m) = -\sum_{i=1}^{C} P_{emp}(i|o) log v_i^L \quad (4\text{-}9)$$

Here v_i^L is the estimated output probability of the Lth layer, and $P_{emp}(i|o)$ is the empirical output probability. Equation (4.9) shows that the distance between the empirical distribution and the estimate needs to be minimized. Based on the negative log-likelihood criterion, Equation (4.9) can be degenerated into:

$$J_{NLL}(W, b: o, y) = -kogv_c^L \tag{4-10}$$

The parameter optimization under the BP algorithm is:

$$W_{t+1}^{\xi} \leftarrow W_t^{\xi} - \varepsilon \Delta W \tag{4-11}$$

$$b_{t+1}^{\xi} \leftarrow b_t^{\xi} - \varepsilon \Delta b_t^{\xi} \tag{4-12}$$

Here W_t^{ξ} and b_t^{ξ} represent the bias and network weight of the ξ-th layer after the t-th iteration training, respectively, and e represents the learning rate. The average bias gradient and weight gradient of DNN after t iterations of training are:

$$\Delta b_t^{\xi} = \frac{1}{M_b} \nabla_{b_t^{\xi}} J(W, b;\ o^m, y^m) \tag{4-13}$$

$$\Delta W_t^{\xi} = \frac{1}{M_b} \nabla_{W_t^{\xi}} J(W, b;\ o^m, y^m) \tag{4-14}$$

The form of the weight gradient is determined by the specific training criterion, usually the CE criterion is used. Under the condition of the CE criterion, the weight matrix gradient and the bias gradient are expressed as:

$$\nabla_{W_t^L} J_{CE}(W, b; o, y) = \nabla_{Z_t^L} J_{CE}(W, b; o, y) \frac{\partial Z_t^L}{\partial W_t^L} = (v_t^L - y)(v_t^{L-1})^T \tag{4-15}$$

$$\nabla_{b_t^L} J_{CE}(W, b; o, y) = (v_t^L - y) \tag{4-16}$$

The fine adjustment of the DNN model parameters is performed using the back-propagation of the error, and the error of the output layer can be expressed as:

$$e_t^L \triangleq \nabla_{Z_t^L} J_{CE}(W, b; o, y) = (v_t^L - y) \tag{4-17}$$

According to these theories, we can use the BP algorithm to fine-tune the parameters in the DNN during the training process, and finally obtain a fully trained DNN model.

4.2 ResNet

Although ResNet and traditional neural network retain the main structure of traditional convolutional neural network in terms of structure, there is a big structural difference between ResNet and traditional neural network architecture due to the introduction of its residual block. ResNet network introduces residual network

structure into neural network. Through this structure, the problem of gradient disappearance or gradient explosion can be dealt with to a certain extent. The core idea is to directly skip one or more layers by introducing a residual structure. This residual structure of the current layer and the previous layer is generally described as "shortcuts". Among them, the residual module is shown in Figure 4.3.

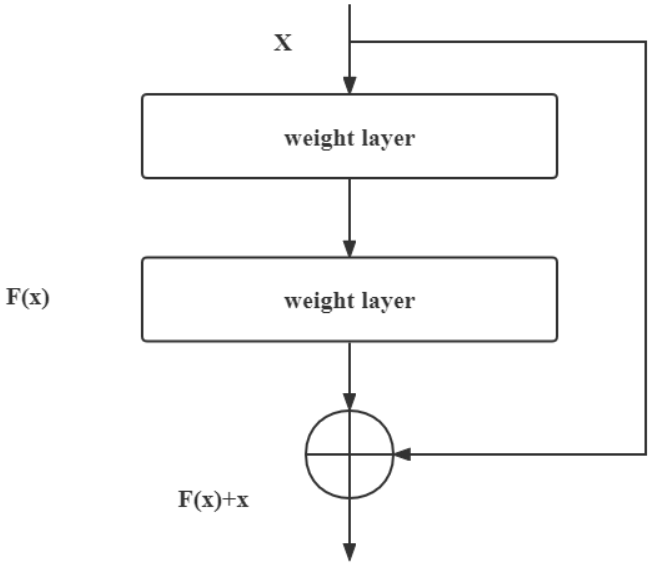

Figure 4.3 Residual block diagram

In the figure, the input signal X is first sent to the first weight layer and passed through the activation function to obtain the intermediate original equation of the neural network, which is $F(x) = H(x) - x$, and the value of $F(x)$s shown in formula (4-18)

$$F(X) = W_2 * Relu(W_1 * X) \tag{4-18}$$

Through the residual block, the output function is obtained as:

$$H(X) = F(X) + X \tag{4-19}$$

which is:

$$H(X) = W_2 * Relu(W_1 * X) + X \tag{4-20}$$

In the residual block, it should be noted that after the input x passes through the first weight and activation function, after passing through the second weight, it is first added to x and then input to the second activation function. Through the residual block structure H(x) in Figure 4-3, the adaptive force of the network can be improved, thereby solving the problem of network degradation.

The following table shows the configuration of the ResNet network structure proposed in 29. The residual structure in the table gives the size of the convolution kernel and the number of convolution kernels on the main branch. *N in the table indicates that the residual network structure is repeated. N times.

Table 4.1 Residual Structure Parameters

Layer name	Output size	18-layer	34-layer	50-layer	101-layer	152-layer
Conv1	112*112	7*7,64,stride 2				
Conv2_x	56*56	3*3 max pool, stride 2				
		$\begin{bmatrix} 3*3 & 64 \\ 3*3 & 64 \end{bmatrix}$ *2	$\begin{bmatrix} 3*3 & 64 \\ 3*3 & 64 \end{bmatrix}$ *3	$\begin{bmatrix} 1*1 & 64 \\ 3*3 & 64 \\ 1*1 & 64 \end{bmatrix}$ *3	$\begin{bmatrix} 1*1 & 64 \\ 3*3 & 64 \\ 1*1 & 256 \end{bmatrix}$ *3	$\begin{bmatrix} 1*1 & 64 \\ 3*3 & 64 \\ 1*1 & 256 \end{bmatrix}$ *3
Conv3_x	28*28	$\begin{bmatrix} 3*3 & 128 \\ 3*3 & 128 \end{bmatrix}$ *2	$\begin{bmatrix} 3*3 & 128 \\ 3*3 & 128 \end{bmatrix}$ *4	$\begin{bmatrix} 1*1 & 128 \\ 3*3 & 128 \\ 1*1 & 128 \end{bmatrix}$ *4	$\begin{bmatrix} 1*1 & 128 \\ 3*3 & 128 \\ 1*1 & 512 \end{bmatrix}$ *4	$\begin{bmatrix} 1*1 & 128 \\ 3*3 & 128 \\ 1*1 & 256 \end{bmatrix}$ *8
Conv4_x	14*14	$\begin{bmatrix} 3*3 & 256 \\ 3*3 & 256 \end{bmatrix}$ *2	$\begin{bmatrix} 3*3 & 256 \\ 3*3 & 256 \end{bmatrix}$ *6	$\begin{bmatrix} 1*1 & 256 \\ 3*3 & 256 \\ 1*1 & 256 \end{bmatrix}$ *6	$\begin{bmatrix} 1*1 & 256 \\ 3*3 & 256 \\ 1*1 & 1024 \end{bmatrix}$ *23	$\begin{bmatrix} 1*1 & 256 \\ 3*3 & 256 \\ 1*1 & 1024 \end{bmatrix}$ *36
Conv5_x	7*7	$\begin{bmatrix} 3*3 & 512 \\ 3*3 & 512 \end{bmatrix}$ *2	$\begin{bmatrix} 3*3 & 512 \\ 3*3 & 512 \end{bmatrix}$ *3	$\begin{bmatrix} 1*1 & 512 \\ 3*3 & 512 \\ 1*1 & 512 \end{bmatrix}$ *3	$\begin{bmatrix} 1*1 & 512 \\ 3*3 & 512 \\ 1*1 & 2048 \end{bmatrix}$ *3	$\begin{bmatrix} 1*1 & 512 \\ 3*3 & 512 \\ 1*1 & 2048 \end{bmatrix}$ *3
	1*1	Average pool, 1000-d fc, sofrmax				

It can be seen from the above table that ResNet has various network structures ranging from 18 layers to 152 layers. From the theoretical analysis, with the increase of the number of layers, the corresponding network structure is gradually deepened, and the more thorough the feature extraction is, the better the effect is obtained. However, as the number of layers increases and the number of training parameters increases, the resources and time required for training will gradually increase. Therefore, when selecting the required number of network layers, it is necessary to select the best depth within a reasonable range of resources. In this paper, ResNet-18

is selected as one of the comparison models for speech emotion recognition comparison experiments.

The code in this ResNet-18 is as follows:

```
class BasicBlock(tf.keras.layers.Layer):

    def __init__(self, filter_num, stride=1):
        super(BasicBlock, self).__init__()
        self.conv1 = tf.keras.layers.Conv1D(filters=filter_num,kernel_size=3,strides=stride,padding="same")
        self.bn1 = tf.keras.layers.BatchNormalization()
        self.conv2 = tf.keras.layers.Conv1D(filters=filter_num,kernel_size=1,strides=1,padding="same")
        self.bn2 = tf.keras.layers.BatchNormalization()
        if stride != 1:
            self.downsample = tf.keras.Sequential()
            self.downsample.add(tf.keras.layers.Conv1D(filters=filter_num,kernel_size=1,strides=stride))
            self.downsample.add(tf.keras.layers.BatchNormalization())
        else:
            self.downsample = lambda x: x

    def call(self, inputs, training=None, **kwargs):
        residual = self.downsample(inputs)

        x = self.conv1(inputs)
        x = self.bn1(x, training=training)
        x = tf.nn.relu(x)
        x = self.conv2(x)
        x = self.bn2(x, training=training)

        output = tf.nn.relu(tf.keras.layers.add([residual, x]))

        return output
```

Figure 4.4 Residual block code

```
def make_basic_block_layer(filter_num, blocks, stride=1):
    res_block = tf.keras.Sequential()
    res_block.add(BasicBlock(filter_num, stride=stride))

    for _ in range(1, blocks):
        res_block.add(BasicBlock(filter_num, stride=1))

    return res_block
```

Figure 4.5 Residual layer code

```
def ResNet18():
    input = Input((X_train.shape[1],1))
    x = Conv1D(filters = 64, kernel_size = 7,strides = 2,padding = 'same',activation = 'relu')(input)
    x = BatchNormalization()(x)
    x = Activation('relu')(x)
    x = tf.keras.layers.MaxPooling1D(pool_size=3, strides = 2, padding = 'same')(x)
    x = make_basic_block_layer(filter_num = 64, blocks = 2)(x)
    x = make_basic_block_layer(filter_num = 128, blocks = 2, stride = 2)(x)
    x = make_basic_block_layer(filter_num = 256, blocks = 2, stride = 2)(x)
    x = make_basic_block_layer(filter_num = 512, blocks = 2, stride = 2)(x)
    x = tf.keras.layers.GlobalAveragePooling1D()(x)
    x = Flatten()(x)
    x = Dense(units=14, activation='softmax')(x)
    model = Model(inputs=input, outputs = x)

    return model
```

```
def resnet_18():
    return ResNetType(layer_params=[2, 2, 2, 2])
```

Figure 4.6 ResNet-18 code

4.3 GoogLeNet

GoogLeNet is a high-performance neural network model proposed by Google in 2014. GoogLeNet is the same as ResNet, and also introduces the idea of basic block, that is, the Inception module. The core structure of the Inception module GoogLeNet can not only improve the performance of the neural network while increasing the depth and width of the network, but also ensure the efficiency of computing resources. Since then, the Inception structure has been continuously updated and optimized for the main problems that limit the performance of deep neural networks, and different versions have been iterated, which are introduced below.

（1） Inception V1

Google first proposed the original version of Inception, as shown in Figure 4.4. Its main idea is to design a parallel network structure, using multiple convolution kernels and pooling kernels of different sizes in each layer to process the input. This structure can increase the adaptability of the network to a certain extent. However, the parameter quantity of each layer of Inception module is the sum of all branch parameters. If the multi-layer Inception structure is superimposed, the final parameter quantity of the model will be too large, which will result in greater dependence on computing resources.

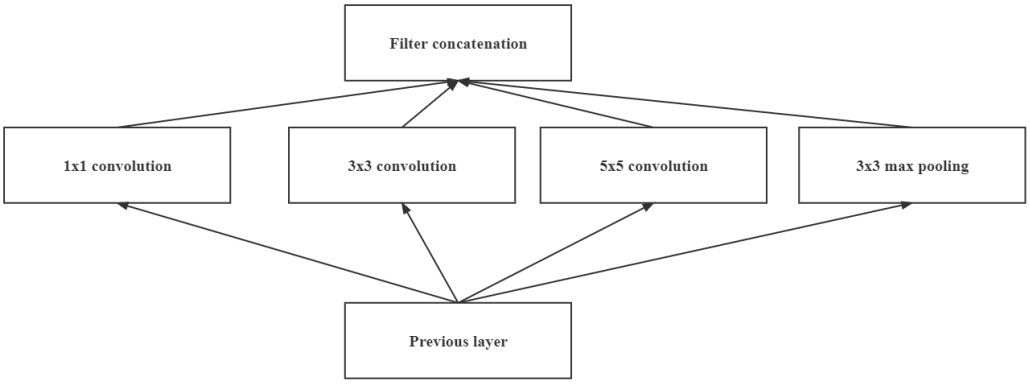

Figure 4.7 Original Inception version

In order to reduce the number of parameters, a reduced-dimensional version of the Inception structure was proposed later. The main difference between the two is that the latter performs this 1x1 convolution operation before the 3x3, 5x5 convolution kernel and after max pooling, respectively. The operation can ensure that the parameter

quantity and complexity of the model are reduced without losing the representation ability of the model. The specific structure is shown in Figure 4.8.

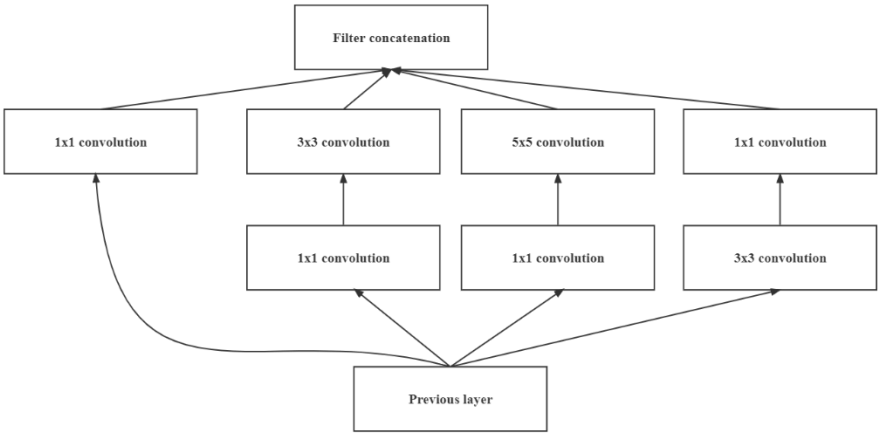

Figure 4.8 Inception V1

A 22-layer GoogLeNet is constructed based on this Inception structure. The network adopts the Inception modular structure to facilitate changes to the model.

（2） Inception V2

Once the Inception structure was proposed, it has attracted widespread attention due to its excellent performance. Google has further improved it and proposed Inception V2, V33233. Its core ideas have two main points. First, for The Batch Normalization (BN) method is proposed to solve the Internal Covariate Shift problem in the process of neural network training. During neural network training, the input distribution of each layer is always changing, making it difficult to train the model. BN is an effective regularization method. It performs a normalization operation on the data in a mini-batch to ensure the output. It is N(0,1), which can increase the robustness of the model. Second, the convolution kernel is decomposed, and a large convolution kernel is decomposed into multiple small convolution kernels. For example, two 3x3 convolution kernels are used to replace the 5x5 convolution kernels in the Inception module. A large number of experiments have proved that this approach will not lead to a decline in the model's expressive ability. In addition, the convolution kernel is asymmetrically decomposed, and a larger 2D convolution is split into two smaller 1D convolutions, as shown in Figure 4.9. Asymmetric decomposition of convolution enables the network to handle richer spatial features.

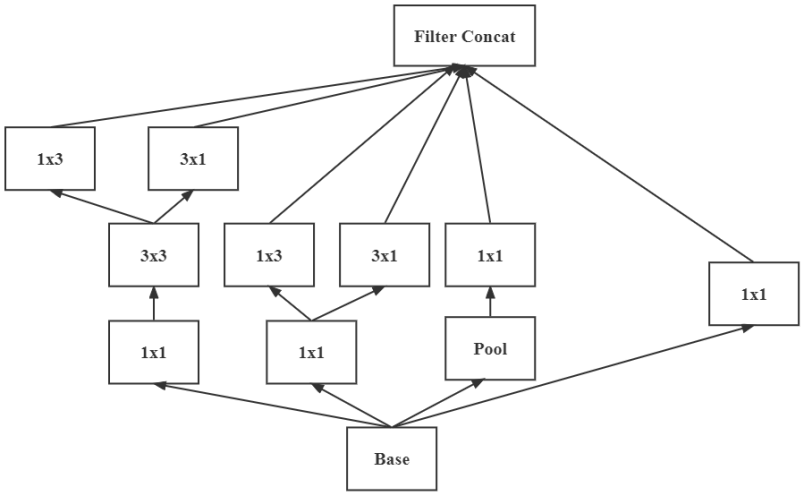

Figure 4.9 Inception V2

The GoogLeNet model construction code is as follows:

```
def inception_model(input, filters_1x1, filters_3x3_reduce, filters_3x3, filters_5x5_reduce, filters_5x5, filters_pool_proj):
    conv_1x1 = Conv1D(filters=filters_1x1, kernel_size=1, padding='same', activation='relu')(input)

    conv_3x3_reduce = Conv1D(filters=filters_3x3_reduce, kernel_size=1, padding='same', activation='relu')(input)

    conv_3x3 = Conv1D(filters=filters_3x3, kernel_size=3, padding='same', activation='relu')(conv_3x3_reduce)

    conv_5x5_reduce  = Conv1D(filters=filters_5x5_reduce, kernel_size=1, padding='same', activation='relu')(input)

    conv_5x5 = Conv1D(filters=filters_5x5, kernel_size=5, padding='same', activation='relu')(conv_5x5_reduce)

    maxpool = MaxPooling1D(pool_size=3, strides=1, padding='same')(input)

    maxpool_proj = Conv1D(filters=filters_pool_proj, kernel_size=1, strides=1, padding='same', activation='relu')(maxpool)

    inception_output = keras.layers.concatenate([conv_1x1, conv_3x3, conv_5x5, maxpool_proj])  # use tf as backend

    return inception_output
```

Figure 4.10 Code for the Inception block

4.4 Speech emotion recognition based on MS-ResNet model

As mentioned above, MFCC is a frame-level feature based on speech signals, so the features contained in each frame may be different. In response to this problem, MS-ResNet is proposed, a multi-scale recognition model based on residual network, which fuses the features of multiple scales to obtain the latent features of MFCC.

4.4.1Multi-scale mechanism

The relevant knowledge of convolutional neural network has been introduced above. Convolutional neural network extracts target features through layer-by-layer abstraction. One of the more important concepts is receptive field.

The receptive field refers to the input area "seen" by neurons in the neural network. In the convolutional neural network, the calculation of an element on the

feature map is affected by a certain area on the input image, and this area is the element's Feel the wild. As shown in Figure 4.11.

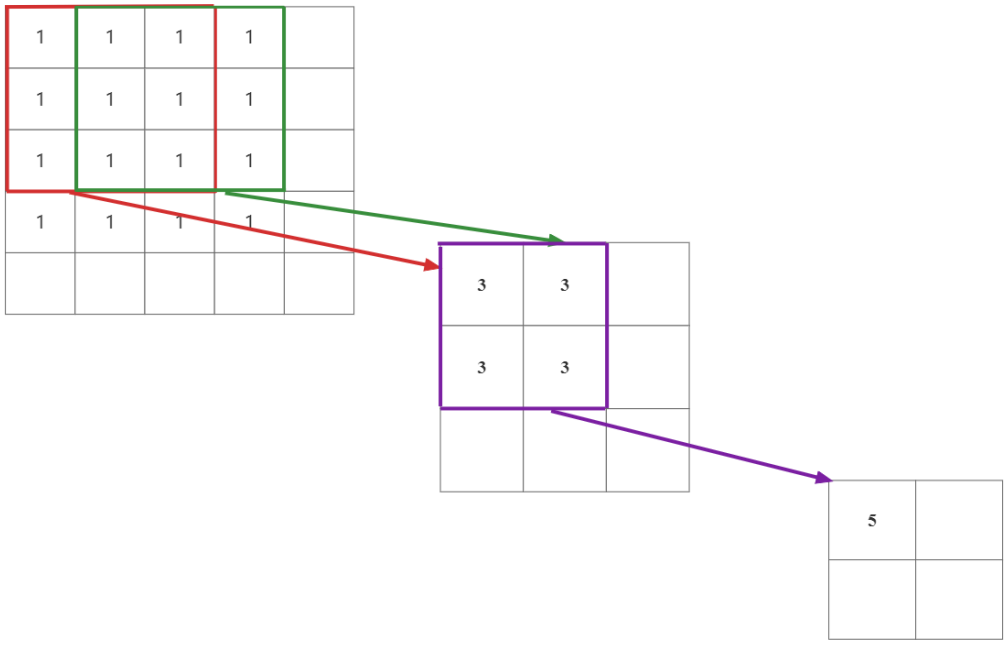

Figure 4.11 Receptive Field

As can be seen from the figure, the original image range that each unit in the first convolution can see is 3*3, while for each unit of the second convolution is the first range of 2*2. A convolution is formed, but in fact, the original image range of 5*5 can be seen.

Since speech is a one-dimensional signal, the convolution used is also one-dimensional convolution, but the principle is the same. Therefore, according to the concept of the convolutional receptive field, the size of the receptive field will affect the characteristics of the target extracted by the convolutional neural network. If the receptive field is too small, local features will be observed intelligently; if the receptive field is too large, too much invalid information will be obtained.

In order to solve the problem brought by the receptive field, a multi-scale mechanism is proposed. In general, multi-scale is the use of multi-scale for learning. The multi-scale network is relatively flexible and has no clear boundaries. Roughly, the network structure can be divided into three types: multi-scale input, multi-scale feature fusion, and multi-scale output. Among them, multi-scale input is typically represented by the use of image pyramids. There are two common multi-scale feature

fusion networks, one is a parallel multi-branch network represented by GoogLeNet, and the second is a serial multi-branch structure represented by FCN 34 and U-Net 35.

4.4.2MS-ResNet model

This paper proposes the MS-ResNet model based on the parallel multi-branch network of GoogLeNet. At the same time, the extension of the depth and width of the neural network is taken into account. Based on the ResNet-18 network model, a multi-scale mechanism is introduced, and a parallel multi-branch network is constructed. The network model is shown in Figure 4.12.

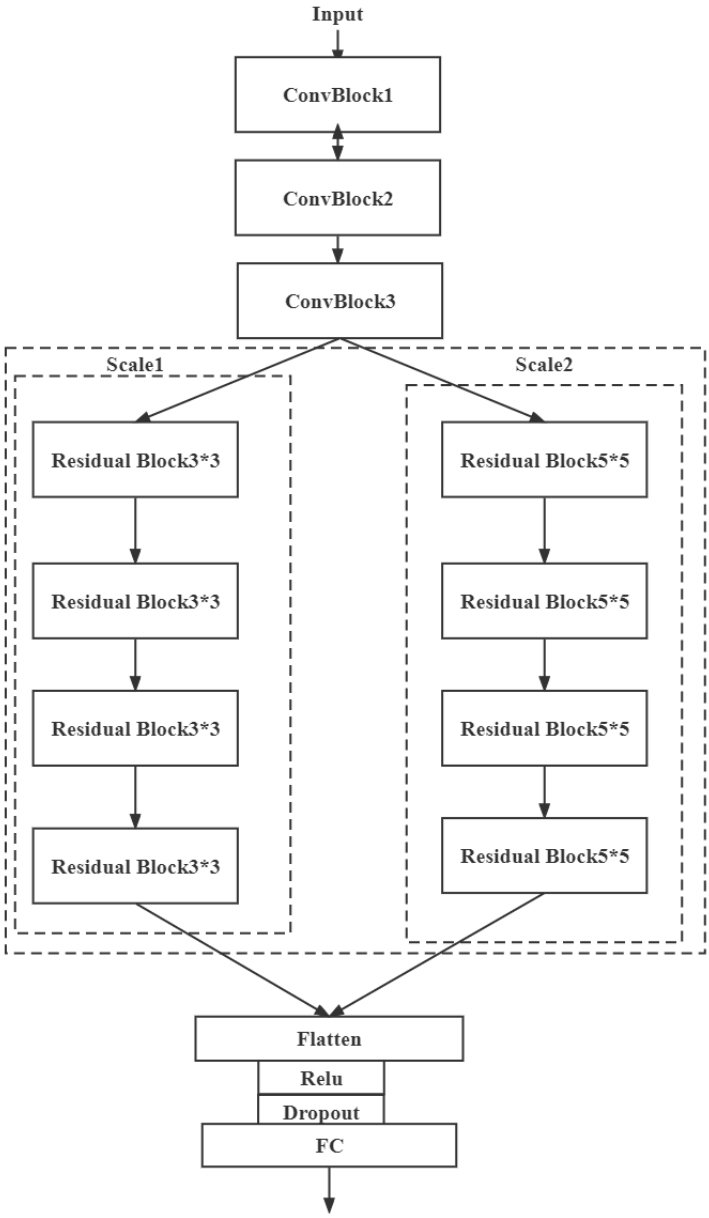

Figure 4.12 MS-ResNet model

In Figure 4.8, ConvBlock is a 1-dimensional convolution block, and its model is shown in Figure 3.3, where ConvBlock1, ConvBlock2, and ConvBlock3 are

convolution blocks with filters of 256, 128, and 64, respectively, and the convolution kernel of 7*7. Scale1 and Scale2 are based on ResNet-18, respectively, and the changed convolutional network, the convolution kernel is 3*3 and 5*5 respectively.

Among them, the residual network adopts the method of "shortcut connection", and the same feature map will have different expressions in two different scale spaces to achieve the purpose of information complementation. Therefore, the two scale spaces are fused in this way to obtain feature parameters with better emotional information, and global information can be obtained. The correlation between adjacent frames can also be obtained. The correlation between non-adjacent frames can be obtained. If the output of the scale1 network is $f^{s_1}(x)$ and the output of the scale2 network is $f^{s_2}(x)$, the output of the network after fusion is $f^{s_1}(x) + f^{s_2}(x)$.

4.5 Result

The experimental environment used in this chapter is exactly the same as the experimental environment in the experimental part of the third chapter of this paper. The datasets are compared with SAVEE and EMO-DB. In addition, the selected dataset is divided into training set and test set in a ratio of 8:2. At the same time, use the StratifiedShuffleSplit function in the sklearn library for cross-validation. The object is the merger of StratifiedKFold and ShuffleSplit, and returns a hierarchical random stack. where stacking is done by the percentage of samples in each class, the percentage is 20%.

This chapter mainly builds MS-ResNet network for speech emotion recognition. In addition, the ResNet structure and the GoogLeNet structure are introduced in this chapter, so this chapter will use the 13-dimensional MFCC described above to conduct comparative experiments around the above models.

Combined with hardware factors such as the experimental environment, ResNet-18 has relatively few parameters and can perform better experiments. Therefore, this paper selects ResNet-18 as the benchmark model for experiments.

（1） On SAVEE database

Table 4.2 Experimental results on the SAVEE dataset

	Accuracy	Precision	Recall	F1-score
GoogLeNet	81.25%	81.12%	80.73%	80.92%

ResNet-18	84.90%	85.38%	84.90%	85.14%
MS-ResNet	86.46%	87.40%	86.46%	86.91%

The line chart of the training accuracy of MS-ResNet on the two datasets is as follows.

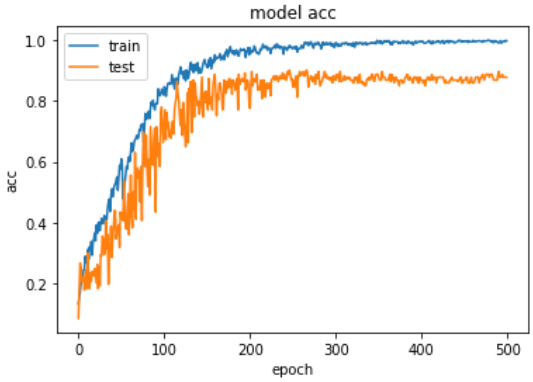

Figure 4.13 On SAVEE acc curve graph

（2） On Emo-DB database

Table 4.3 Experimental results on Emo-DB dataset

	Accuracy	Precision	Recall	F1-score
GoogLeNet	87.85%	87.85%	87.85%	87.85%
ResNet-18	85.98%	86.20%	86.20%	86.20%
MS-ResNet	87.85%	88.57%	88.19%	88.97%

The line chart of the training accuracy of MS-ResNet on the two datasets is as follows.

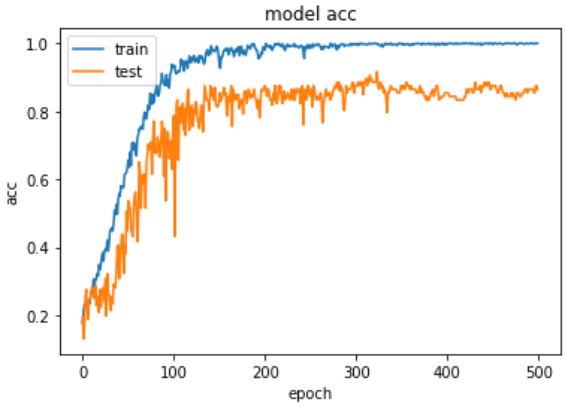

Figure 4.14 On EMO-DB acc curve graph

From the comparison of experiments on the two datasets, it can be clearly seen that MS-ResNet has improved compared with GoogLeNet and ResNet-18 in the four evaluation indicators of accuracy, precision, recall and F1-score.

Combined with the experiments in this paper, it can be seen that for MS-ResNet, compared with the GoogLeNet and ResNet benchmark models, the performance on the MFCC feature set has a certain improvement. From the results, on the SAVEE dataset, the four evaluation indicators, MS-ResNet has an improvement of 1%-2% compared with the ResNet network model, and an improvement of about 5%-6% compared with GoogLeNet. Compared with the benchmark ResNet-18, the improvement of MS-ResNet on Emo-DB is more obvious.

It can be seen from this that the multi-scale learning mechanism can more effectively extract the hidden information in the speech information, and fuse the speech features between adjacent frames and non-adjacent frames, compared with the single-scale deep convolutional neural network. The network model has improved. Therefore, the multi-scale residual network provides a research idea for the extension of the deep learning network model, depth and width.

4.6 Summary of this chapter

This chapter is the main part of this experiment, taking two development ideas of deep learning neural network as the starting point of network construction. In view of the characteristics of MFCC as a frame-level feature, a multi-scale learning mechanism is introduced. Based on this, the MS-ResNet network structure is proposed, and for this model, on the SAVEE and Emo-DB datasets, the MFCC features are compared with the ResNet and GoogLeNet networks. The experimental results prove that the MS-ResNet network structure has a positive effect on the improvement of recognition accuracy. It provides a follow-up research idea for the speech emotion recognition network model.

5 Financial management, resource efficiency and resource saving

This paper aims to create a convolutional neural network algorithm for speech emotion recognition. The purpose of this chapter is an analysis of the financial and economic aspects of the work performed. Carry out a scientific management planning

process, calculate the total capital cost of the project, and improve the efficiency of evaluating economic work on this basis, and provide convenience for subsequent work.

5.1 SWOT analysis

The analysis of maximum competitiveness is carried out with the method of SWOT analysis: S (strength), W (weakness), O (opportunity), T (threat).

The analysis is divided into two stages: first, describe the strengths and weaknesses of the project, and identify opportunities and threats that have emerged or may appear in the project; second, determine the compatibility of the project's strengths and weaknesses with external conditions, and what needs to be done strategic changes. As shown in Table 5.1.

Table 5.1 SWOT analysis

	Strengths: S1. Ability to recognize emotional information in speech signals. S2. Improve the accuracy of speech recognition.	**Weaknesses:** W1. Lack of sentiment-accurate classification W2. There is a lack of emotional voice database data, and the data may not accurately express emotions
Opportunities: O1. Very high commercial value O2. Speech recognition is an important part of artificial intelligence.	Strengths and Opportunity-Based Strategies: 1.Reducing the Accessibility of Voice	Strategies based on weaknesses and opportunities: 1. Add voice database
Threats: T1. This project requires a large amount of voice data support. When the emotion classification is inaccurate, it may lead to voice recognition errors.	Strengths and Threats-Based Strategies 1.Determine sentiment classification	Weakness and threat based strategy: 1.Real-time feedback, step-by-step revision

5.2 Organization and planning of work

In the process of organizing the implementation work, it is necessary to reasonably plan the division of labor and timing of each participant in the process. A complete work list was constructed for this purpose as shown in Table 5.2 below.

Among them, E - engineer, author of BKP. T – Tutor.

Table 5.2 - List of works and duration of their implementation

Stage	Work	Participant	Workload
1	Set goals and objectives	E	E – 100%
2	Domain Analysis	E, T	E – 100%, T – 50%
3	Specified time schedule	E, T	E – 20% , T – 100%
4	Select references	E, T	E – 100%, T – 10%
5	Determine the method to achieve the goal	E, T	E – 100%, T – 50%
6	Analysis and selection of various algorithms for speech emotion recognition	E, T	E – 100%, T – 25%
7	Algorithm code building and testing	E	E – 100%
8	Writing of conclusions and explanations	E, T	E – 100% , T – 30%
9	Prepare report documents	E	E – 100%
10	The act of obtaining the results of the implementation of the master thesis	T	T- 100%
11	Defend	E	E – 100%

According to the data obtained in the table, a Gantt chart is constructed to visually display the calendar schedule of all work, and the result is shown in Figure 5.1.

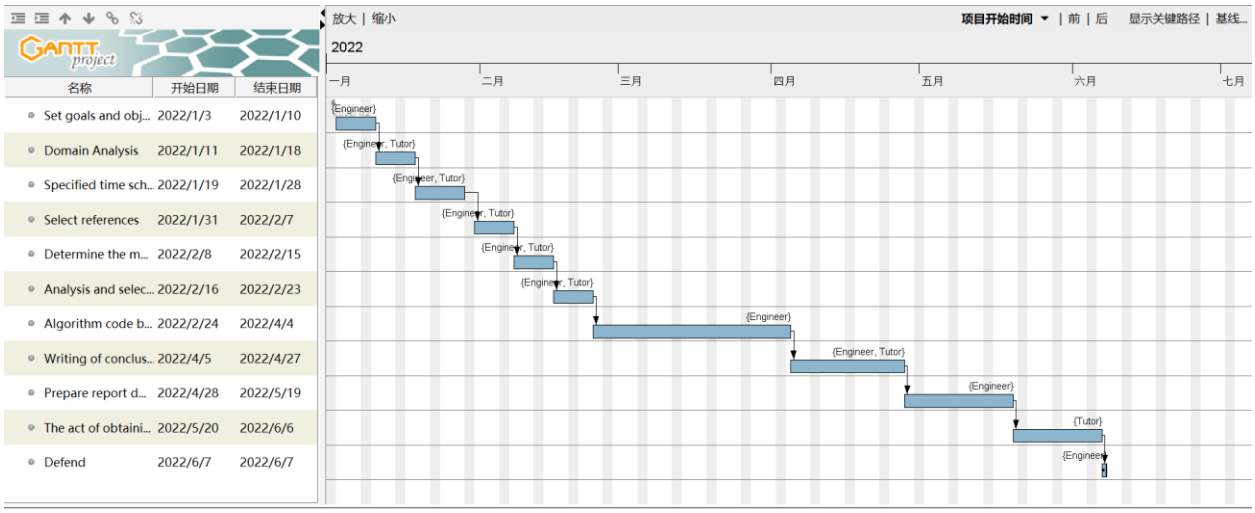

Figure 5.1 Work Schedule

Work duration

Due to the low degree of reproducibility of this work, in order to calculate the duration of the work phase, the experimental static method is chosen here. On the basis of the simulation method, according to the experimental static method and retrograde calculation, the possible value of the construction period is determined as t_{wt}.

$$t_{wt} = \frac{3t_{min}+2t_{max}}{5} \quad (5.1)$$

Where t_{min} is the minimum duration of work, days. t_{max} - maximum duration of work, days.

To specify a work schedule, calculate the duration of each phase in working days (Equation 5.2), then use Equation 5.3 to convert the result to calendar days. Calculated as follows.

The calculation of the duration of each stage (T_{wd}) in working days is carried out according to the following formula:

$$T_{wd} = \frac{t_{wt}}{K_{co}} K_{add} \quad (5.2)$$

Where t_{wt} is the duration of work, days; K_{co} - coefficient of work performance, taking into account the influence of external factors on compliance with predetermined durations, in particular (K_{co} = 1); K_{add} - coefficient taking into account additional time for compensation of unforeseen delays and coordination of work (K_{add} = 1.2).

The calculation of the duration of the stage in calendar days is carried out according to the formula:

$$T_{CD} = T_{wd}T_C \quad (5.3)$$

Where T_{CD} is the duration of the stage in calendar days; T_C is the calendar coefficient that allows you to switch from the duration of work in working days to their counterparts in calendar days, and is calculated by the formula:

$$T_c = \frac{T_{cad}}{T_{cad} - T_{do} - T_{ph}} = \frac{365}{365 - 52 - 14} = 1{,}22$$

Where T_{cad}- calendar days ($T_{cad} = 365$); VD – days off ($T_{do} = 52$); TPD - public holidays ($T_{ph} = 14$).

Table 5.3 describes the work phases and labor intensity of the participants in each phase. A linear schedule for project implementation was established based on the labor intensity values of the participants at each stage. As shown in Table 5.3

Table 5.3 - Labor costs for the dissertation

Stage	Participants	Duration of work, days			Participant labor intensity (person-day)			
					T_{wd}		T_{CD}	
		t_{min}	t_{max}	t_{wt}	E	T	E	T
1	2	3	4	5	6	7	8	9
Set goals and objectives	E	3	8	5	6	-	7.3	-
Domain Analysis	E, T	10	15	12	14.4	7.2	17.6	8.8
Specified time schedule	E, T	1	2	1.4	1.7	0.3	2.1	0.4
Select references	E, T	5	6	5.4	6.5	0.7	8.0	0.9
Determine the method to achieve the goal	E, T	2	4	2.8	3.4	1.7	4.1	2.1
Analysis and selection of various algorithms for speech emotion recognition	E, T	20	30	24	28.8	7.2	35.1	8.8
Algorithm code building and testing	E	40	60	48	57.6	-	70.3	-
Writing of conclusions and explanations	E, T	4	6	4.8	5.8	1.7	7.1	2.1
Prepare report documents	E	5	7	5.8	7	-	8.5	-
The act of obtaining the results of the implementation of the master thesis	T	10	20	14	-	16.8	-	20.5
Defend	E	1	2	1.4	1.7	-	2.1	-
Total				276	132.9	35.6	162.2	43.6

5.3 Scientific and technical research budget

In the process of budgeting, the following grouping of costs by items is used:

- Material costs of scientific and technical research;
- costs of special equipment for scientific work (Depreciation of equipment used for design);
- basic salary;
- additional salary;
- labor tax;
- overhead.

5.3.1Calculation of material costs

The calculation of material cost is shown in Table 5.4.

Table 5.4 - Calculation of the cost of materials

Name of materials	Unit price, rub.	Quantity	Amount, rub.
A4 papers	5	200	1000
Total			1000

5.3.2Costs of special equipment

This article mainly includes the salaries of two participants, Tutor and Engineer. The calculation of the basic salary is based on the labor cost of each labor stage of the project implementation. The average daily salary is calculated according to the following formula:

$$S_d = S_m/21 \quad (5.4)$$

where the number of working days in a year is taken into account ≈ 247 and, therefore, there are an average of 21 working days in a month (with a five-day working week).

Time spent by each participant during the workday, rounded to the nearest whole number, taken from Table 5.2 (T - 35.6=36, E - 132.9=133 workdays). At the same time, mentors and engineers work 5 days a week.

A number of factors are used to include bonuses, additional wages and regional allowances as part of full wages. Therefore, the integral coefficient is calculated as:

$$K = K_p K_e K_R$$

Among them, K_p is premium share (K_p=1.1); K_e-extra wage (K_e=1.188 for six-day work week, K_e =1.113 for five-day work week); K_R -regional allowance (K_R=1.3)

Therefore, to transfer the wage (base) amount from the participant's earnings to the full earnings (estimated wage portion), we multiply the average rate, time cost, and points factor.

Table 5.5 - Wage Costs.

Participant	Salary, rub./month	Average daily rate ,rub./working day	Time spent, working days	Coefficient	Salary fund, rub.
T	52312	2491	36	1.59	142684
E	12364	589	133		124555
Total					267240

5.3.3Labor tax

The cost of the Uniform Social Tax (UST), which includes contributions to pension funds, social insurance and health insurance, is a compulsory expenditure item and accounts for 30% of the total salary of staff. Therefore, the cost of UST is calculated using the following formula:

$$M_S = M_T * 0.3 \quad (5.4)$$

Where M_T is the total payroll cost of the project.

The total costs of social tax are equal to $M_S = 267240 * 0.3 = 80172$

5.3.4Overhead costs

This section mainly focuses on the cost of electricity used during the operation of the equipment during the project. The electricity bill during the operation of the equipment is calculated according to the following formula:

$$M_e = P_{eq} t_e M_t \quad (5.5)$$

where P_{eq} is the power consumed by the equipment, kW; t_e – equipment operation time, hour; M_t - tariff for 1 kWh. Current tariff for Tomsk – 2.73.

At the same time, according to the data of an engineer in Table 5.2 (Td=133 days), the operating time of the equipment is calculated according to 8 hours of working days:

$$t_e = T_d * K_t \quad (5.6)$$

where T_d is the total duration of the stages of work, days; K_t is the coefficient of equipment utilization in time, equal to the ratio of its operation time to T_d.

where $K_t \leq 1$ is the equipment utilization factor in time, equal to the ratio of the time of its work in the process of project implementation to T_d. The time spent at a personal computer is equal to 2/3 of the total time of work on the project.

In turn, the power consumed by the equipment is determined by the formula:

$$P = P_{eq} * K_{load} \quad (5.7)$$

P_{eq} - rated power of the equipment, kW (indicators for a PC - 0.5 kW, for a printer - 0.1 kW); $K_{load} \leq 1$ - load factor, depending on the average degree of use of the rated power. Let's take the value of $K_{load} = 1$.

The cost calculation is presented in Table 5.6 below.

Table 5.6 - Electricity bill calculation

Name of equipment	Equipment operation time t_e, hour	Power consumption P, kW	Tariff for 1 kW h M_t, rub	Electricity cost, rubles
Personal Computer	133days * 8hours*2/3=709	0.5	2.73	968
Laser printer	10	0.1		2.73
Total:				971

5.3.5Calculation of depreciation expenses

The purpose of this section is to describe the depreciation of equipment that was used during the conduct of the study. The following formula is used to calculate depreciation cost:

$$M_{de} = \frac{R_a * M_{eq} * t_w * n}{M_y} \quad (5.8)$$

Where R_a is the annual depreciation rate for a piece of equipment; M_{eq} - book value of a unit of equipment, rub; M_y - the actual annual fund of the operating time of the relevant equipment, hour; t_w is the actual operating time of the equipment during the project, hour; n is the number of involved equipment of the same type.

R_a for a personal computer is the reciprocal of the depreciation period of 2.5 years, hence $R_a(PC) = 1 / 2.5 = 0.4$. The nominal value of one PC is 50 thousand rubles. To calculate $My\ (PC)$, we will take into account that the number of working days in 2022 with a five-day working week is 247 days, the working day lasts 8 hours. Thus $M_y(PC) = 247 * 8 = 1976\ hours$. However, the PC is occupied only 2/3 of the design time spent by the engineer (Td from Table 5.3), therefore, $t_w = 133\ days * 8\ hours * 2/3 = 709\ hours$.

Thus, the depreciation accrued on the PC is equal to:

$$M_{de}(PC) = \frac{0.4 * 50000RUB * 709hours * 1}{1976} = 7176RUB$$

The cost of a simple printer is 7800 rubles, its $M_y = 500\ hours; R_a = 1/2\ years = 0.5;\ t_w = 10\ hours$ (from Table 5.5).

$$M_{de}(Priter) = \frac{0.5 * 7800RUB * 10}{500} = 780RUB$$

Total depreciation costs are $M_{de} = 7956\ rubles$, where most of the amount goes to cover the wear and tear of a personal computer.

5.3.6Calculation of other expenses

Other expenses include the research and development expenses not included in the preceding paragraph, which are calculated at 10% of the sum of the original expenses, and calculated as follows:

$$M_{other} = (M_m + M_T + M_s + M_e + M_{de}) * 10\% \quad (5.9)$$

So other expenses are: $M_{other} = (1050 + 267240 + 80172 + 971 + 7956) * 10\% = 35739RUB$.

5.3.7Formation of budget costs

The calculated cost of research is the basis for budgeting project costs. Determining the budget for the scientific research is given in the Table 5.7.

Table 5.7 Calculation of development cost budget

Name	**Cost**, rub.
Materials	1050
Basic salary	267240
Contributions to social funds	80172

Electricity costs	971
Depreciation deductions	7956
Other expenses	35739
Total	393128

5.4 Conclusion

This chapter plans the various stages of scientific and technological research, calculating the labor intensity of all works and the number of performers employed for each work. Visual charts are built on the basis of these data.

In addition, the total cash cost of project development is calculated and the terms and conditions of return are determined. Secondly, the user information of consumers was analyzed, and then a SWOT table was created, and then the benefits of the project were evaluated.The cost of the project is equal to its development cost $M = 393128\,rubles$.

6 Social responsibility

The project is aimed to develop the program to study the recognition and classification of emotion in speech. The program development is carried out only with the help of a computer.

In this chapter, the following issues will be discussed:

Identify and study harmful and dangerous production factors when using PC;

Identify ways to reduce the effects of said factors to safe limits or, where possible, eliminate them entirely;

Environment safety;

Security in emergencies that may arise while the PC is running.

6.1 Legal and organizational issues of security

The legal side of the issue of labor safety when working on a master's thesis is regulated by the Labor Code of the Russian Federation. So the number of working hours per week did not exceed 40 hours, work was carried out on weekdays, during the working day an hour break was provided, which does not apply to working time.

Since the activity of this study is to work on the PC, the use of the PC needs to be regulated. Among the main documents that regulate the working conditions and organization of the use of PCs are SanPiN 1.2.3685-21 «Hygienic standards and requirements for ensuring the safety and (or) harmlessness of environmental factors for humans». It contains many environmental hygiene requirements for the workplace 36. Also used GOST 12.2.032-78 SSBT «Work place when performing work while sitting. General ergonomic requirements» and GOST R50923-96.«Displays. Operator's workplace. General ergonomic and work environment requirements .Methods of measurement»3738, which regulates the working environment of PC users.

When working with a computer, it is necessary to create a comfortable and safe environment for it, and it is necessary to consider whether the correct location and layout of the engineer's work area is ergonomic. In the engineer's office space, the layout of the workplace takes into account that the distance between the desktop and the video monitor is not less than 2m, and the distance between the eyes and the monitor is 650mm. The monitor should also provide brightness and contrast adjustment to reduce eye strain and improve the comfort of working on a PC.

In addition, the shape of the desktop should be comfortable to maintain a rational posture of the user, so that he can change his body position to prevent fatigue. Tables 6.1 present the actual values of the parameters of the engineer's working area in accordance with the requirements of GOST 12.2.032-78 SSBT37 and GOST R 50923-96.38

Table 6.1 - Analysis of the workplace

Requirement	**Fact**
Windows should be oriented to the north and northeast	Windows oriented towards
Window openings must be equipped with adjustable devices such as: blinds, curtains, external visors	Windows with curtains and blinds
Area per workstation for PC users based on flat discrete screens (liquid crystal, plasma) -4.5 m2	The workplace is 20m^2

The premises where workplaces with PCs are located must be equipped with protective grounding (zeroing) in accordance with the technical requirements for operation.	Grounding available
You should not place workplaces with a PC near power cables and inputs, high-voltage transformers, technological equipment that interferes with the operation of the PC	There are no similar objects near

According to the comparison of all the requirements in Table 6.1, the analysis of the workplace shows that all but the first requirement are met.

6.2 Industrial safety

This section analyzes the harmful and dangerous factors that affect engineers working in PC-equipped workplaces.

To identify potential factors, according to GOST 12.0.003-2015 «Dangerous and harmful production factors. Classification» 39. All factors of production are divided into groups of elements: physical, chemical, biological and psychological. For this work, it is recommended to consider the physical and psychological detrimental and risk factors of production, which characterize the work area of a software engineer as a developer. Section 6.1 of this chapter describes working days and breaks, and providing a comfortable workplace, with further consideration of physical factors. Table 6.2 lists these factors.

Table 6.2 - Harmful and dangerous production factors when performing work on a computer

Factors (GOST 12.0.003-2015)	**Stages of work**		**Regulations**
	Development of a rollback	**Operation**	
1.Insufficient illumination of the working area	+	+	1. SP 52.13330.2016 Natural and artificial lighting.40 2. GOST 12.1.003-2014 SSBT. Noise. General safety requirements.41
2.Exceeding the noise level	+	+	

3Increased level of electromagnetic radiation	+	+	3. GOST 12.1.006-84 SSBT. Electromagnetic fields of radio frequencies. General safety requirements.42 4. GOST 12.1.038-82 SSBT. Electrical safety. Maximum allowable levels of touch voltages and currents.43 5. GOST 12.1.030-81 Occupational safety standards system (SSBT). Electrical safety. Protective ground. Zeroing.44
4.Increased voltage in the electrical circuit, the closure of which can occur through the human body	+	+	

In the case of software development, where the object is a workplace, including a personal computer and a room, such harmful effects include: improper lighting, noise, electromagnetic radiation, the danger of electric shock, and others. It is also important to take care of environmental safety and safety in emergency situations that may arise when working with a PC.

6.2.1 Illumination of the working area

Since the work of a software engineer involves a visual type of work, the organization of proper lighting has a significant place.

Neglect of this factor can lead to occupational eye diseases. The workspace combines natural lighting (through windows) and artificial lighting (using lamps when there is a lack of natural light).

The category of visual work of the programmer belongs to category III d (high accuracy), the parameters of artificial lighting are indicated in Table 6.3.

Table 6.3 - Normative values of illumination

Characteristics of visual work	The smallest or equivalent size of the object of distinction, mm	Category of visual work	Subcategory of visual work	The contrast of the object with the background	Characteristics of the background	Artificial lighting		
						Illumination, lx		
						With combined lighting system		With general lighting system
						Total	Including from	

							the general	
High precision	0.3 to 0.5	III	Г	Medium and large	Light and medium	400	200	200

Also, according to SanPiN1.2.3685-2145, lighting requirements are given in Table 6.4:

Table 6.4 -Lighting requirements for PC workstations

View	Requirement
Illumination on the desktop	200-400 lux
Illumination on the PC screen	Not more than 200 lux
Glare on the screen	Not higher than 40 cd/m2
Direct light source fading	200 cd/m2
Blindness index	No more than 20
Discomfort score	No more than 15
Brightness ratio	
Between work surfaces	3:1-5:1
Between wall and equipment surfaces	10:1
Ripple factor	No more than 10%

The artificial lighting of the engineer's office environment is calculated as follows. The dimensions of the office are as follows: length A = 6 m, width B = 4 m, height H = 3 m. The office uses lamps of the ODR type (diffuse general lighting with a shielding grid) and fluorescent lamps of the LB type (white light) with a power of 135W and a luminous flux of Φ = 4800 lm. The total number of office lamps is n=6, and the pulsation coefficient of such lamps does not exceed 5%, which is in line with the standard.

The illuminance of the room is calculated by the following formula:

$$E_{\phi} = \frac{n*\eta*\phi}{s*k*z} \quad (6.1)$$

Among them, n is the number of fixtures; η is the luminous flux utilization rate; Φ is the luminous flux of the lamp, lm; s is the room area, m^2; k is the illumination unevenness coefficient; z is the illumination unevenness coefficient correction value.

For rooms with computer technology k = 1.4. The correction factor for fluorescent lamps is z = 1.1. The area of the room is $S = A * B = 6 * 4 = 24\,m^2$.

The luminous flux utilization coefficient is determined using a table based on the room index and the reflection coefficients from the walls, ceiling and work surface. Therefore, we first find these indicators.

The room index is determined by the formula:

$$i = \frac{S}{h*(A+B)} \tag{6.2}$$

Where S is the area of the room, m^2; A is the length of the room, m; B is the width of the room, m; h is the height of the suspension of fixtures, m.

Meanwhile, the estimated height of the fixture suspension above the office work surface (h) is determined by the formula:

$$h = H - h_p - h_c \tag{6.3}$$

Where H is the height of the ceiling in the room, m; h_p is the distance from the floor to the working surface of the table, m; h_c is the distance from the ceiling to the luminaire, m.

Then the calculated height of the suspension of fixtures is equal to:

$$h = 3 - 0.8 - 0.01 = 2.19m$$

Let's substitute the obtained value into formula 6.2 to calculate the room index.

$$i = \frac{S}{h * (A + B)} = \frac{24}{2.19 * (4 + 6)} = 1.1$$

Because the ceiling in the room is pure concrete, the walls are concrete pasted over with light wallpaper with windows and the working surface contains a PC, then according to 46, we will take the reflection coefficients from the walls $\rho_c = 30\%$, the ceiling $\rho_n = 50\%$ and from the working surface $\rho_p = 10\%$.

According to the table of coefficients for using the luminous flux of luminaires with fluorescent lamps [46] for the corresponding values of i, ρ_c, ρ_n, we determine the utilization coefficient of the luminous flux.

It turns out for the ODR lamp at $i = 1.1$ $\rho_c = 30\%$ and $\rho_n = 50\%$ the luminous flux utilization factor is 43%.

Taking into account all the parameters discussed above, we find the illumination according to the formula 6.1:

$$E_{\phi} = \frac{6 * 0.43 * 4800}{24 * 1.4 * 1.1} = 335lux$$

In the room under consideration, the illumination should be at least 300 lux according to SanPiN1.2.3685-2145. In this room, the illumination is 335 lux and is within the normal range, therefore, additional light sources are not needed.

6.2.2Noise

Sources of noise are: running equipment, computer fans, copiers and air conditioners.

Noise has a negative effect on the human body: it reduces efficiency, increases fatigue, affects the hearing organs and the central nervous system, and reduces attention.

According to GOST 12.1.003-83 “System of labor safety standards (SSBT). Noise. General safety requirements”[47. The noise level of the programmer's workplace does not exceed 50dBA. Therefore, the noise level in the premises should be limited.

In the studio, due to the regular maintenance of the computer equipment in the room, the noise level does not exceed the value specified by the standard.

6.2.3 Electromagnetic radiation

Personal computers are sources of electromagnetic waves, that is, perturbations of the electromagnetic field (EMF) propagating in space. All electrical appliances emit such waves, but the monitor screen makes the biggest contribution. At certain levels, such fields have a harmful effect on a person: a violation of the functional state of the nervous and cardiovascular systems, this manifests itself in increased fatigue, a decrease in the quality of work operations, changes in blood pressure and pulse.

Since a liquid crystal monitor is used, soft X-ray control is not performed. Permissible radiation values are shown in Table 6.5, taking into account GOST 12.1.006-84 SSBT42.

Table 6.5 - Temporary allowable levels (VDU) of EMF created by PC

Name of parameters	VDU EMF	VDU EMF

Electric field strength	In the frequency range 5 Hz - 2 kHz	25 V/m	27 V/m
	In the frequency range 2 kHz -400 kHz	2,5 V/m	2,5 V/m
Electrostatic potential of the video monitor screen		500V	490V

The norms of permissible levels of electromagnetic fields depend on the time spent by a person in a controlled area. The presence of personnel at the workplace for 8 hours is allowed at a tension not exceeding 5 kV / m.

The main way to reduce the harmful effects is to increase the distance from the source (at least 50 cm from the user). Protection against the influence of the electromagnetic field of industrial frequency currents are stationary or portable grounded shielding devices. At the enterprise, electromagnetic radiation does not exceed 5 kV / m, therefore, when working at a computer, special screens and other personal protective equipment were not used.

6.2.4Electrical Hazard

Among the common hazards in the work area is electric shock. The danger of defeat is determined by the magnitude of the current passing through the human body or the voltage of contact.

When a person receives a discharge of electric current, electrical injuries, electric shocks and even death can be obtained. GOST 12.1.038-82 SSBT 43 defines the maximum allowable values of contact voltage and current at the workplace (see Table 6.6).

Table 6.6 - Permissible values of touch voltage and current

Type of current	**Touch voltage, V**	**Current, mA**
	No more	
Variable, 50 Hz	2.0	0.3
Constant	8.0	1.0

The main source of danger of electric shock is a personal computer. In order to avoid accidents, employees must undergo appropriate training without fail.

You should not work on a personal computer when:

High humidity (relative air humidity over 75%);

High temperature (more than 35 ° C);

The presence of conductive dust, conductive floors and the possibility of simultaneous contact with metal elements connected to earth and the metal case of electrical equipment.

The personal computer is powered by 220 V AC, 50 Hz. This voltage is life-threatening, so the following precautions are mandatory:

Before starting work, you need to make sure that the switches and the socket are fixed and do not have bare current-carrying parts;

If a malfunction of equipment and instruments is detected, it is necessary, without making any independent corrections, to inform the person responsible for the equipment.

To avoid electric shock, it is necessary to protect all current-carrying parts from possible contact, and metal cases must be grounded.

Thus, all the requirements when working with a PC were met, since all the necessary indicators of the norms are within acceptable limits.

6.3 Environmental Safety

At present, the problems of environmental safety and environmental protection are prominent and the top priority. This is why the normal disposal of computer equipment is subject to strict legal restrictions. To a greater extent, this is because many different materials are used in the production of such devices, which can cause irreparable damage to the environment and thus irreparable harm to human health. Its production includes toxic raw materials that require special handling and processing – without them, the material would gradually collapse, causing irreparable harm to the environment and human health. Many office equipment become hazardous wastes that endanger the atmosphere, hydrosphere and lithosphere after their useful life. For example, liquid crystal displays are a major source of greenhouse gases, and fluorescent lamps contain 10 to 70 milligrams of mercury.

According to the Code of Administrative Offenses of the Russian Federation 47, used equipment (including computers) is prohibited from being disposed of with ordinary waste and must be handled or disposed of by contacting special services. GOST 12.3.031-83 “Use of Mercury. Safety Requirements” requires that all mercury-containing waste and equipment be collected and returned only by certified personnel (electricians) 48.

In addition, to ensure environmental safety and harmless waste disposal, the office uses the practice of selective waste collection. As a prevention and prevention of the dangerous impact of electrical equipment on the environment, the sanitary standards SanPiN 1.2.3685-2145 recommend using it in an economical mode of operation, as well as paying attention to the compliance of the materials used in the computer with environmental safety norms and standards.

Failed fluorescent lamps are one of the most common sources of mercury pollution. In addition to glass and aluminum, each lamp contains approximately 60 mg of mercury, so spent fluorescent lamps are a dangerous source of toxic substances 49.

The disposal of such lamps consists in their transfer to processing enterprises that have special equipment for processing harmful lamps into harmless raw materials - a sorbent, which can be a material for other industries. According to GOST R 57740-2017 50 and GOST R 51768-200151, used fluorescent lamps are waste that is collected and sorted separately, so their disposal and storage must meet certain requirements.

Since a personal computer was used in the development of this master's thesis, it is necessary to describe the correct disposal of computer scrap after its failure. In accordance with a government decree, legal entities are prohibited from disposing of computer equipment on their own. To do this, you need to find a special company that deals with private recycling.

6.4 Emergency Safety

The most common emergency in the office is a fire. Such a workplace belongs to category “B” (fire hazardous), since this room contains dust, substances and materials that can burn when interacting with air.

A fire can start due to several factors:

- The occurrence of a short circuit in the electrical wiring due to a malfunction of the wiring itself or electrical connections and electrical distribution boards;
- Ignition of computing equipment devices due to insulation failure or malfunction of the equipment itself;
- Fire of furniture or floor due to violation of fire safety rules, as well as improper use of additional household electrical appliances and electrical installations;
- Ignition of artificial lighting devices.

Fire fighting methods include:

- Briefings, availability of evacuation plans, proper installation and operation of equipment, proper maintenance of buildings and territories, training in safety regulations, publication of special instructions and posters.
- Compliance with fire regulations, exclusion of the formation of a combustible environment, the use of hardly combustible materials.
- Provided means of signaling, fire extinguishers, automatic stationary fire extinguishing systems, timely evacuation.

To prevent a fire, a room with a PC must be equipped with primary fire extinguishing equipment: a carbon dioxide fire extinguisher of the OU-2 or OU-5 type.

A fire can cause not only harm to health, but also material damage. Applicable to the work performed, paper documents and/or electronic media may be destroyed in the event of a fire.

6.5 Summary of this chapter

After analyzing the working conditions at the workplace, we can conclude that the room in which software development is carried out meets the necessary standards and, if safety precautions and rules for using a computer are observed, work in this room will not lead to a deterioration in health.

The premises and workplace meet all regulatory requirements. In order to avoid a negative impact on health while working with a PC, it is necessary to take breaks and conduct specialized sets of physical exercises.

The effect of harmful and dangerous factors is minimized by appropriate measures.

The microclimate, lighting and electrical and fire safety comply with the requirements set forth in the relevant regulatory documents. Environmental safety at the enterprise satisfies the current regulatory documents.

Conclusion

In the research of this paper, a speech recognition model based on multi-scale residual convolutional neural network is mainly developed. The algorithm mainly uses the deep learning method to improve the accuracy of speech emotion recognition.

In this study, a VGG-like neural network is first constructed, aiming at the structural characteristics of the Mel Frequency Cepstral Coefficient (MFCC) speech emotion feature, which has performed well in the current speech emotion recognition research. It is used in the SAVEE sentiment database to test MFCCs with two different dimensions, 13-dimensional and 39-dimensional. Through the visualization of the mixed matrix, it is compared that the 39-dimensional MFCC has a stronger performance in the recognition of speech emotion. Therefore, the 39-dimensional MFCC is selected as the follow-up feature.

Then, according to two different research ideas in the field of neural network research, a multi-scale residual neural network model based on ResNet network is constructed. Use this model to conduct experiments on the EMO-DB and SAVEE datasets, and compare with the ResNet and GoogleNet networks. It can be concluded from the experimental results that the MS-ResNet model constructed in this paper has high accuracy in speech emotion recognition. Finally, it can be concluded that multi-scale can obtain the hidden emotional information in the speech signal from MFCC.

List of student publications

1. Chen J. Development of passenger lift control algorithm based on the controller with fuzzy logic // Modern Technologies, Economics and Education: Proceedings of the II All-Russian Scientific and Methodological Conference. – Tomsk, 2020. – P. 17-18.

2. Chen J. Using the Python API for human speech recognition // Science Today: Challenges, Perspectives and Opportunities. Materials of international scientific-practical conference. – Vologda, 2021. – P. 15-17.

3. Chen J. Speech emotion recognition based on deep residual convolutional neural network // Eurasian scientific journal. – 2022. – No. 3. – P. 20-24.

4. Chen J. Speech Emotion Recognition Based on Multiscale Residual Network // II International Scientific and Technical Conference "Actual problems of science and technology". – Sarapul, 2022 (in press).

5. Chen J. Speech Emotion Recognition Based on Deep Convolutional Neural Networks // IX International youth scientific conference "Mathematical and software of information, technological and economic systems". – Tomsk, 2022 (in press).

REFERENCE

1. Emotional process // Wikipedia. [2022]. URL: https://ru.wikipedia.org/?curid=2124606&oldid=121156470 (accessed on 03.04.2022).
2. Emotion // Wikipedia. [2022]. URL: https://ru.wikipedia.org/?curid=3428&oldid=120416982 (accessed on 03.04.2022).
3. Dellaert F, Polzin T, Waibel A. Recognizing emotion in speech[C]//Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96. IEEE, 1996, 3: 1970-1973.
4. Yu F, Chang E, Xu Y Q, et al. Emotion detection from speech to enrich multimedia content[C]//Pacific-Rim Conference on Multimedia. Springer, Berlin, Heidelberg, 2001: 550-557.
5. Slaney M, McRoberts G. Baby ears: a recognition system for affective vocalizations[C]//Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181). IEEE, 1998, 2: 985-988.
6. Schuller B, Rigoll G, Lang M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture[C]//2004 IEEE international conference on acoustics, speech, and signal processing. IEEE, 2004, 1: I-577.
7. Wang Y, Guan L. An investigation of speech-based human emotion recognition[C]//IEEE 6th Workshop on Multimedia Signal Processing, 2004. IEEE, 2004: 15-18.
8. Reynolds D A. Comparison of background normalization methods for text-independent speaker verification[C]//Fifth European Conference on Speech Communication and Technology. 1997.
9. Neiberg D, Elenius K, Laskowski K. Emotion recognition in spontaneous speech using GMMs[C]//Ninth international conference on spoken language processing. 2006.
10. Zhou Y, Sun Y, Zhang J, et al. Speech emotion recognition using both spectral and prosodic features[C]//2009 international conference on information engineering and computer science. IEEE, 2009: 1-4.
11. Luengo I, Navas E, Hernáez I, et al. Automatic emotion recognition using prosodic parameters[C]//Ninth European conference on speech communication and technology. 2005.
12. Satt, Aharon, Shai Rozenberg, and Ron Hoory. "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms." Interspeech. 2017.
13. Sabour, Sara, Nicholas Frosst, and Geoffrey E. Hinton. "Dynamic routing between capsules." Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017.
14. Wu, Xixin, et al. "Speech emotion recognition using capsule networks." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

15. Trigeorgis, George, et al. "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network." 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016.
16. Neumann, Michael, and Ngoc Thang Vu. "Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech." (2017).
17. S. Mirsamadi, E. Barsoum and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2227-2231.
18. Zhang T, Wu J. Speech emotion recognition with i-vector feature and RNN model[C]//2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP). IEEE, 2015: 524-528.
19. Ortony A, Turner T J. What's basic about basic emotions?[J]. Psychological review, 1990, 97(3): 315.
20. Ekman P, Friesen W V, O'sullivan M, et al. Universals and cultural differences in the judgments of facial expressions of emotion[J]. Journal of personality and social psychology, 1987, 53(4): 712.
21. Russell J A. A circumplex model of affect[J]. Journal of personality and social psychology, 1980, 39(6): 1161.
22. Lang P J, Bradley M M, Cuthbert B N. International affective picture system (IAPS): Technical manual and affective ratings[J]. NIMH Center for the Study of Emotion and Attention, 1997, 1(39-58): 3.
23. Livingstone S R, Peck K, Russo F A. Ravdess: The ryerson audio-visual database of emotional speech and song[C]//Annual meeting of the canadian society for brain, behaviour and cognitive science. 2012: 205-211.
24. Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech[C]//Interspeech. 2005, 5: 1517-1520.
25. Institute of Automation Chinese Academy of Sciences.The selected Speech Emotion Database of Institute of Automation Chinese Academy of Sciences(CASIA) [DB/OL].2012/5/17.
26. 韩文静, 李海峰, 阮华斌, 等. 语音情感识别研究进展综述[J]. 软件学报, 2014, 1.
27. Gu J, Wang Z, Kuen J, et al. Recent advances in convolutional neural networks[J]. Pattern Recognition, 2018, 77: 354-377.
28. Zhang W, Tanida J, Itoh K, et al. Shift-invariant pattern recognition neural network and its optical architecture[C]//Proceedings of annual conference of the Japan Society of Applied Physics. 1988: 2147-2151.

29. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
30. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.
31. Smolensky P. Information processing in dynamical systems: Foundations of harmony theory[R]. Colorado Univ at Boulder Dept of Computer Science, 1986.
32. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//International conference on machine learning. PMLR, 2015: 448-456.
33. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.
34. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.
35. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.
36. SanPiN 1.2.3685-21. Hygienic standards and requirements for ensuring the safety and (or) harmlessness to humans of environmental factors.
37. GOST 12.2.032-78. SSBT Workplace when performing work while sitting. General ergonomic requirements.
38. GOST R 50923-96. Displays. Operator's workplace. General ergonomic and work environment requirements. Measurement methods.
39. GOST 12.0.003-2015. Dangerous and harmful production factors. Classification.
40. SP 52.13330.2016 Natural and artificial lighting.
41. GOST 12.1.003-2014 SSBT. Noise. General safety requirements.
42. GOST 12.1.006-84 SSBT. Electromagnetic fields of radio frequencies. General safety requirements.
43. GOST 12.1.038-82 SSBT. Electrical safety. Maximum allowable levels of touch voltages and currents.
44. GOST 12.1.030-81 Occupational safety standards system (SSBT). Electrical safety. Protective ground. Zeroing.
45. SanPiN 1.2.3685-21 Hygienic standards and requirements for ensuring the safety and (or) harmlessness of environmental factors for humans.

46. Life safety: workshop / Yu.V. Borodin, M.V. Vasilevsky, A.G. Dashkovsky, O.B. Nazarenko, Yu.F. Sviridov, N.A. Chulkov, Yu.M. Fedorchuk. - Tomsk: Publishing House of the Tomsk Polytechnic University, 2009. - 101 p.
47. Code of the Russian Federation on Administrative Offenses of December 30, 2001 N 195-FZ (as amended on April 30, 2021, as amended on May 17, 2021).
48. GOST 12.3.031-83. System of labor safety standards. Mercury work.
49. 2011 RoHS Directive on Restriction of Hazardous Substances.
50. GOST R 57740-2017. Waste management. Requirements for the reception, sorting and packaging of hazardous municipal solid waste.
51. GOST R 51768-2001. Waste management. Methodology for the determination of mercury in mercury-containing waste.