Research On Emotion Management Based on Speech Analysis for Nursing Homes

Boyuan Liang *

Faculty of Information Science and Technology, Electronic Information Engineering, School of Nanjing Forestry University, Nanjing, China

* Corresponding Author Email: 1361156907@qq.com

Abstract. Speech emotion recognition is an important research area in artificial intelligence aimed at identifying the emotional states of speakers. This paper provides an overview of the current state and key algorithms of speech emotion recognition, focusing particularly on the emotional well-being of elderly residents in nursing homes. Firstly, the paper introduces the background and application areas of speech emotion recognition, emphasizing its significance in human-computer interaction, psychological health monitoring, and emotional intelligent systems. The paper extensively discusses methods of extracting emotional features, it adopts the K-Nearest Neighbors (KNN) classification algorithm, which serves as a simple yet effective classification method with potential in emotional speech recognition. This study aims to explore the design, implementation, and future directions of an emotion recognition system based on KNN. This algorithm exhibits a high fitting performance for emotional speech data, presenting advantages such as ease of implementation and alignment with the distribution characteristics of emotional speech data. Through multiple comparative experiments, the optimal K value for recognition has been determined, achieving a straightforward and convenient emotional recognition simulation. Finally, the paper summarizes the current state of speech emotion recognition, emphasizing its potential and significance in practical applications.

Keywords: Emotion Management; K Nearest Neighbor Algorithm; Speech Emotion Recognition; Mel- Frequency Cepstral Coefficient.

1. Introduction

1.1. Background and Significance of the Study

Speech emotion recognition is one of the research directions that have attracted much attention in the field of artificial intelligence, aiming to identify the emotional state of the speaker by analyzing the speech signal. The importance of speech emotion recognition is increasing as the demand for human-computer interaction, intelligent customer service, and mental health monitoring increases. With the increasing trend of global aging, the number of elderly people in nursing homes is rapidly increasing. Take Beijing as an example, according to the latest data released by the Beijing Municipal Bureau of Civil Affairs (like figure 1), as of March 2020, there were 560 various types of nursing institutions and 898 community-based elderly care institutions and facilities, totaling 1,458 in Beijing. From the point of view of the time of establishment of nursing homes in Beijing, the number of newly built nursing homes in the past ten years accounted for as much as 30% of the total number of nursing homes, with the number of new nursing homes reaching 266[1]. The following figure shows the trend of the number of nursing homes established in Beijing:

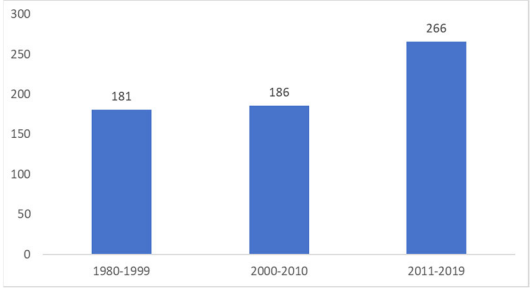

Fig. 1 Change trend of the number of nursing homes in Beijing

We can see that in the past decade, the number of nursing homes has been increasing, making it an important residence for the elderly. Managing the emotions of the elderly is a crucial matter.

The elderly often experience mental health issues such as depression, anxiety, and loneliness. Emotional management helps residents cope with these issues and improves their mental health and quality of life. Emotional management can assist residents in building positive social relationships and alleviating feelings of loneliness. Most importantly, emotional management helps reduce the risk of psychological stress and physical health problems, thereby enhancing the overall health of the elderly. In summary, the importance of emotional management in nursing homes lies in improving the mental health, happiness, and quality of life of the elderly, enabling them to better adapt to new living environments and social circles. To meet these needs, nursing homes should provide psychological support and social activities to help residents actively confront various challenges in life.

Therefore, developing a system capable of real-time monitoring and managing the emotional states of the elderly is crucial. Speech analysis technology can be an effective means to achieve this goal. Speech emotion analysis, as a potential technological solution, can monitor the emotional states of the elderly in real time, providing emotional support and intervention.

The importance of speech emotion recognition is mainly reflected in the following aspects:

1.Human-Computer Interaction:

With the widespread use of intelligent assistants and virtual AI systems, human-computer interaction has become an integral part of modern life. Being able to recognize users' emotional states helps intelligent systems better understand and respond to users' needs[2].

2.Psychological Health Monitoring:

Emotion recognition technology can be used to monitor individuals' psychological health status. By analyzing emotional features in speech signals, signs of mental health issues such as depression and anxiety can be detected. This aids in early intervention and treatment.

3.Emotion Analysis:

In the era of big data, particularly in social media and online comments, emotion analysis is crucial for understanding public sentiment and product feedback. Speech emotion recognition can be utilized to analyze phone recordings, customer service dialogues, and social media videos, providing a better understanding of users' emotional responses.

Taking into account speech emotion analysis and recognition technology, distributed architecture, and load balancing, this paper designs an intelligent speech emotion analysis system for elderly residents in nursing homes. The system boasts fast response times and high recognition rates[3]. This kind of system has broad application prospects and can provide valuable support in multiple fields.

1.2. Research Status

Speech emotion recognition aims to analyze a speaker's voice to identify their emotional state. This process mainly involves selecting emotional features, extracting and reducing emotional features, and performing classification. Acoustic features related to speech emotion include prosodic features, voice quality features, and spectrogram-based features. Some scholars have also proposed the use of nonlinear features. The accuracy of emotion recognition is closely related to the selected features and their extraction methods. Therefore, improvements in data dimensionality reduction and extraction algorithms can effectively enhance emotion recognition performance[4]. For example, the zero-crossing feature extraction method based on the maximum Teager energy operator can better reflect the features of different emotional states[5]. In addition to feature selection and extraction, selecting appropriate models for accurate emotion classification is crucial. Common emotion recognition models include Hidden Markov Models, Gaussian Mixture Models, Support Vector Machines, and Neural Networks. Sound propagates as waves, and processing speech information requires sampling, quantization, encoding, and storage[6]. To recognize emotions in speech, features of the sound need to be extracted. Therefore, in the development process of speech emotion recognition technology,

various algorithms have emerged, including traditional machine learning and deep learning algorithms.

Common emotion recognition models include Support Vector Machines (SVM), Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Long Short-Term Memory networks (LSTM), self-attention mechanisms, and others. Several studies have proposed methods based on Convolutional Neural Networks (CNN) for the recognition of images, videos, speech, audio, and music. These studies indicate that CNN-based analysis methods can be successfully applied to one-dimensional signals such as speech and audio[7].

Emotion management systems tailored for the elderly are continuously evolving in the market to meet the psychological health and emotional management needs of older adults. Here are some existing or developing emotion management systems in the market:

1. Intelligent Voice Assistants: Smart voice assistants like Amazon Alexa, Google Assistant, and Apple Siri can be used for emotional management among the elderly.

2. Mental Health Applications: There are numerous mental health applications in the market, such as Headspace, Calm, and Talk space.

3. Smartwatches and Health Monitoring Devices: Smartwatches and health monitoring devices can track physiological data such as heart rate, sleep quality, and physical activity, providing emotional management recommendations. Some devices also feature emergency assistance functions to address urgent situations.

2. Overview of Emotional Speech Features

2.1. Emotional Speech Features

Emotion is a crucial indicator of human intelligence and represents a subjective awareness in the human brain. Currently, methods for representing emotions can be categorized into two main types: discrete emotion theory and dimensional emotion theory. Dimensional emotion spaces commonly used today mainly include two-dimensional and three-dimensional emotion spaces. Russell and others employed a two-dimensional model, Valence-Arousal (VA), to represent emotions (like figure 2). In this model, Arousal signifies an individual's neural activation level, while Valence indicates the positivity or negativity of an individual's emotional state[8]. Taking Valence-Arousal emotional dimension space as an example, it is illustrated in the diagram below.

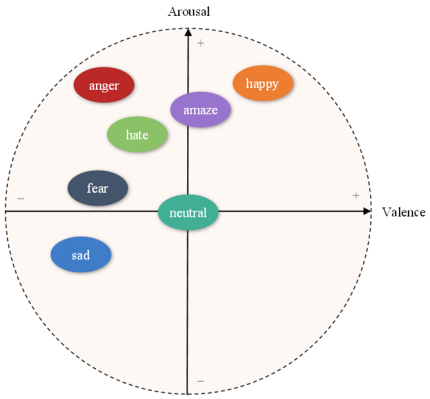

Fig. 2 Valence-Arousal dimension emotional space

Compared with discrete emotion models, dimensional emotion models can not only describe emotions qualitatively, but also describe emotions quantitatively, reflecting more subtle emotional changes. Therefore, in recent years, the research on the use of dimensional sentiment models for emotion recognition has been on the rise.

A variety of acoustic features can be extracted from affective speech, Compared to discrete emotion models, dimensional emotion models not only offer qualitative descriptions of emotions but also enable quantitative descriptions, capturing more subtle emotional changes. Therefore, in recent

years, there has been a rising trend in utilizing dimensional emotion models for emotion recognition. Various acoustic features can be extracted from emotional speech, reflecting the speaker's emotional behavior. The selection and extraction of these emotional features are crucial for the final outcome of emotion recognition. Consequently, determining which speech features can effectively capture emotional changes remains one of the key challenges in the field of speech emotion recognition. Some common speech features include:

1. Short-Time Energy: Distribution of energy in the sound signal, used to represent speech intensity.
2. Zero Crossing Rate: Frequency of the sound signal crossing the zero level, useful for detecting voiced and unvoiced speech sounds.
3. Voiced-to-Unvoiced Ratio: Ratio of voiced and unvoiced portions in the sound signal, helpful for phoneme classification in speech.
4. Duration of Pronunciation: Duration of each phoneme, used for speech analysis and emotion inference.
5. Speech Rate: Speed of speech, reflecting changes in emotional states.
6. Fundamental Frequency: Fundamental frequency of sound, related to pitch, which might be influenced by emotions.
7. Resonance Peak Frequency and Bandwidth: Reflecting the sound's resonance characteristics, potentially affected by emotions.
8. Mel-Frequency Cepstral Coefficients (MFCC): A parameter commonly used for sound feature extraction, capturing the sound's spectral characteristics. The relationship between MFCC and frequency is approximately as shown in equation (2-1):

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

where f is the frequency and the unit is Hz.

In constructing speech emotion features, we select the aforementioned eight categories of features and their derived parameters, forming a 140-dimensional set of speech emotion feature parameters for recognition (like figure 3).

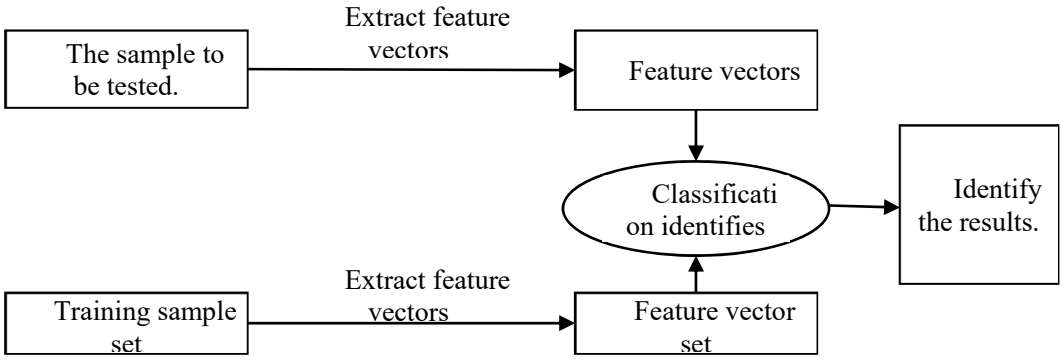

Fig. 3 The principle of speech emotion recognition

2.2. Establish a Database of Emotional Voices

In addition, we needed an emotional speech database.

In our experiment, we initially selected five emotional categories from the Berlin Emotional Speech Database: fear (A-fear), happiness (F-happiness), neutral (N-neutral), sadness (T-sadness), and anger (W-anger). Building upon the Berlin database, we conducted surveys and investigated the happiness index of elderly residents in nursing homes, as well as their commonly used emotional expressions in daily life. Through collaboration with nursing homes, we established an exclusive emotional speech feature database tailored for the elderly in these facilities. This database comprises numerous samples of elderly voices expressing various emotions, with 100 samples for each emotional category, totaling 500 samples. Among these, 250 sentences were used as training samples,

and the remaining 250 sentences were used as test samples. This emotional corpus played a significant role in our experiment.

3. Emotion Speech Recognition Algorithms

3.1. K Nearest Neighbor Classification Algorithm

The K-Nearest Neighbor (KNN) classification algorithm is a simple and intuitive machine learning method used for solving classification problems. Its core idea is based on distance measurement in the feature space of samples, classifying the sample to be classified as the majority category among its K nearest neighbors. The main steps of the KNN algorithm include:

Suppose the feature parameters of the samples to be classified are X, and the feature parameter set of the known labeled training sample set is ;$\{X_1, X_2, X_3, \dots, X_n\}$For the sample to be tested, Xwe calculate its Euclidean distance from each sample in the training sample set, i.e.,$D(X, Xi)l = 1,2, \dots, n$

$$D(X, X_l) = \sqrt{\sum_{t=1}^{n}(X(i) - X_l(i))^2}, l = 1,2, \dots, n \quad (2)$$

where represents the n dimension of the feature vector. We find the closest training samples, which are called the closest neighbors of the sample to be tested. Then, analyze which category of this nearest neighbor has the largest sample size, and classify the samples to be tested into that category.

Despite its relative simplicity, KNN performs remarkably well in many practical applications, especially in scenarios involving small datasets and the need for strong interpretability. It finds widespread applications in fields like image classification, speech emotion recognition, recommendation systems, and more.The simplicity and intuitiveness of the KNN algorithm make it widely used in speech emotion recognition. It does not require explicit model training but instead classifies samples by comparing them with neighboring samples. This approach can exhibit outstanding performance under certain circumstances.

3.2. Working Principle

In MATLAB, emotional feature extraction is typically based on audio signals. These features can be used to identify emotions in speech. The working principle of the emotional feature extraction process is as follows:

1. Audio Signal Acquisition and Preprocessing: First, audio signals are captured from the audio source. Then, preprocessing is performed, such as noise reduction, down-sampling, and normalization, to ensure audio quality and consistency.

2. Frame Segmentation: The audio signal is divided into short-time frames. Each frame contains 10 to 30 milliseconds of audio data.

3. Feature Extraction: Specific feature extraction functions are applied to each frame, including short-time zero crossing rate, fundamental frequency (pitch), energy ratio, Mel-frequency cepstral coefficients (MFCC).

4. Construction of Feature Vectors: The results of feature extraction for each frame are combined to form feature vectors.

5. Data Processing and Emotion Classification: The extracted feature vectors are input into an emotion classifier. The classifier is trained using labeled emotion data to map feature vectors to specific emotion categories during testing. In the classifier training process, feature vectors and their corresponding emotion labels are used to adjust the classifier's weights and parameters.

6. Emotion Recognition and Result Output: Feature extraction is performed on new audio signals, and the results are input into the trained classifier. The classifier outputs the emotion category to which the audio signal belongs.

Its working principle can be outlined as follows:

1. Data Preparation: Firstly, gather a labeled training dataset comprising features and their corresponding labels. Features describe the attributes of the data, while labels represent the categories or values to be predicted.

2. Feature Distance Measurement: For classification, KNN algorithm measures similarity based on the distance between features. Common distance metrics include Euclidean distance, Manhattan distance, and others. For regression, the reciprocal of distances between features is often used to measure similarity.

3. Selecting the Value of K: In KNN, K represents the number of nearest neighbors considered during prediction. Choosing an appropriate K greatly influences the algorithm's performance. A smaller K makes the model more sensitive to noise, while a larger K might lead to an overly smooth decision boundary.

4. Prediction Process: For a new data point, the algorithm calculates its distance from all data points in the training set. Then, it selects the K nearest neighbors to this data point. For classification, the category of the new data point is determined based on the most common label among these K nearest neighbors (majority voting). For regression, the average of labels from these K nearest neighbors is calculated and used as the prediction for the new data point.

The performance of the KNN algorithm heavily relies on the chosen features, distance metrics, and the appropriate value of K. Selecting relevant features and an optimal K value are crucial for the accuracy of the KNN algorithm.

3.3. System Simulation Process

The specific steps of the experiment are as follows:

1. Feature Extraction: Utilizing a pre-written feature extraction function, extract feature vectors from the audio files corresponding to emotional speech. Save these feature vectors as individual MAT files and place them in the same directory as the main program.

2. Main Program Implementation:Write the main program based on the principles of the K-Nearest Neighbors (KNN) algorithm. The main program's functions include constructing training and test sample sets, setting the value of K to implement the KNN classification algorithm. Ensure that the main program can read the extracted feature vectors and MAT files, construct training and testing data, and implement the KNN algorithm.

3. Experiment Execution:

Run the main program, use the KNN algorithm to recognize emotions in the test samples. Analyze the experimental results to see which emotion category each sample has been classified into. Repeat the tests, try different values of K, compare their classification results, and eventually determine the optimal value of K.

This experimental process allows you to classify emotional speech using extracted speech feature vectors and the KNN algorithm. By adjusting the value of K, you can find the most suitable K for your dataset to achieve the best classification results.

3.4. Experimental Results

By comparing different values of K, it was found that the most suitable K value is 24, at which the emotion recognition accuracy is the highest (like figure 4).

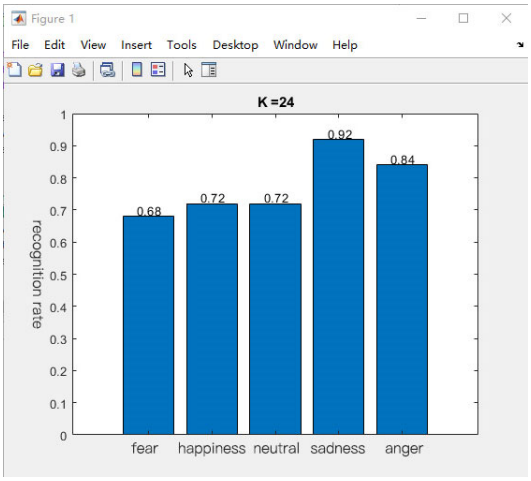

Fig. 4 Recognition rate comparison chart

4. Emotional speech recognition system

4.1. Emotion Management System

The emotion management system is an innovative application that can be carried around or placed at home. It monitors and provides feedback on emotional states by analyzing sound emitted by the human body, including variations in tone, pitch, as well as the strength and frequency of speech signals. The core idea of this system is to make emotion management more intelligent by using cloud-based technology to remind individuals to adjust their emotions in a timely manner. Additionally, this system can establish a vast information repository by continuously accumulating and refining data from small groups to large samples, becoming an intelligent assistant in the field of human emotions and health.

Designing an emotion management system for nursing homes is a complex task aimed at improving the mental health and quality of life of elderly residents. Here is an overview of the design of such a system, including its core functions, architecture, data processing, and privacy considerations (like figure 5). The emotion management system will include the following main components:

1. Data Collection and Processing:

Various sensors, such as heart rate monitors, temperature measurements, and cameras, are used to collect real-time physiological and behavioral data. Periodic emotional surveys are conducted to obtain subjective emotional feedback from elderly residents. The system needs to collect multiple types of data to monitor the emotional state of the elderly. This data needs to be preprocessed, cleaned, and standardized for emotional analysis.

2. Sentiment Analysis:

Establishing a comprehensive sentiment analysis system, utilizing both K-Nearest Neighbors (KNN) and Artificial Neural Network (ANN) algorithms to build models, is a core component of the system. Sentiment analysis employs machine learning and natural language processing techniques to understand the emotional states of the elderly. Advanced techniques, including Mel-Frequency Cepstral Coefficients (MFCC) and Convolutional Neural Networks (CNN), can be applied in the extraction of speech signals to identify and discern emotional information in voices. Sentiment analysis can be broken down into the following steps: first, feature extraction involves extracting relevant features from speech, text, and physiological data, such as intonation, vocabulary choice, and heart rate variations. Next is sentiment classification, Finally, analyzing emotional trends involves studying the trends in emotional data to identify long-term emotional changes and emotional fluctuations triggered by events.

3. User Interface and Feedback:

Design a user-friendly interface for the elderly residents and nursing home staff. The system can offer practical features: firstly, personalized emotional feedback provides customized emotional support and suggestions based on the results of sentiment analysis. Secondly, the system can provide data visualization, displaying emotional data and trends through charts and graphs to help the elderly understand and manage their emotional states better. There can also be social interaction features, facilitating interactions with family, friends, or other elderly residents to alleviate feelings of loneliness. The system can monitor the elderly residents' voice inputs in real-time and analyze their emotional states.

4. Privacy and Security:

Ensuring strict compliance with privacy regulations, sensitive data will be encrypted and protected. The system will provide the elderly residents with complete control over their data and clearly define data usage policies.

Designing an emotion management system for nursing home residents requires expertise from multiple fields, including machine learning, human-computer interaction, psychology, and healthcare. Therefore, interdisciplinary teamwork is crucial. In summary, the emotion management system is an innovative technological application poised to play a significant role in various domains, aiding people in managing emotions and improving mental health. However, in practical applications, attention must be given to privacy protection and timely intervention measures to ensure system effectiveness and acceptability. The success of the system will play a vital role in enhancing the emotional well-being and quality of life for the elderly.

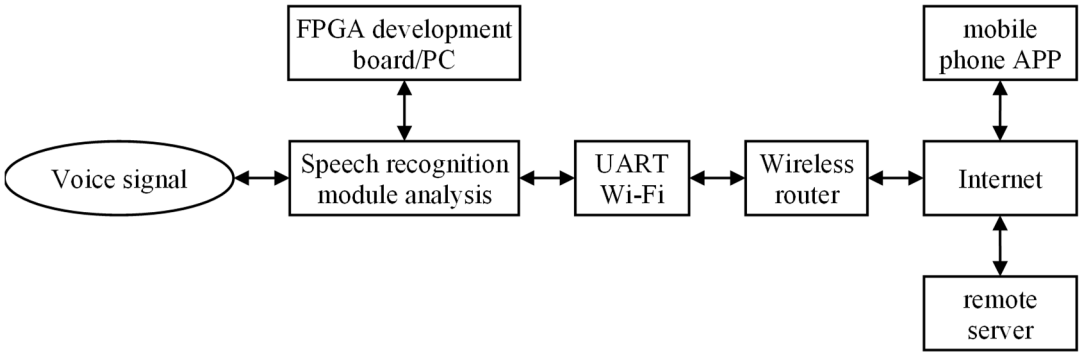

Fig. 5 Structure diagram of emotion management system

5. Summary and Outlook

5.1. Research Summary

This study, based on the K-Nearest Neighbors (KNN) classification algorithm, has developed an emotion speech recognition system implemented and tested on the MATLAB platform. The aim of this system is to provide emotional support and intervention for elderly individuals in their homes. By continuously monitoring emotional states through real-time speech analysis, this system holds the promise of enhancing emotional well-being among the elderly and improving their overall quality of life. Future research endeavors will focus on refining the system's functionality and exploring additional application areas to meet diverse needs of the elderly population. This paper has summarized the system's design, implementation steps, and performance evaluation results. Furthermore, it explores the prospects of emotion speech recognition systems in the future.

Experimental results demonstrated the system's strong performance in emotion speech recognition tasks.

5.2. Prospects

The field of emotion speech recognition holds vast future prospects, and here are some potential directions for development:

1.Multimodal Fusion: Integrating speech emotion analysis with other perceptual modalities such as images and physiological signals to enhance the accuracy and comprehensiveness of emotion recognition.

2.Application of Deep Learning: Deep learning algorithms, especially Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), hold potential in emotion speech recognition. Future systems can explore the application of deep learning techniques to improve performance.

6.Cross-language and Cross-cultural Emotion Recognition: Emotion speech recognition systems can expand to different languages and cultural backgrounds to meet the needs of users globally.

In conclusion, this study has demonstrated the application potential of emotion recognition in the field of speech through the design and implementation of an emotion speech recognition system based on the K-Nearest Neighbors (KNN) classification algorithm. In the future, with continuous technological advancements, emotion speech recognition systems will find applications in diverse fields, bringing forth innovations in human-computer interaction and psychological health monitoring.

References

[1] ZHOU Shangyi, LUO Mengting. Spatial difference analysis of effective demand for nursing home beds in Beijing [J]. Beijing Planning and Construction,2017(05):32-35.

[2] Feng Baoding, Zhang Guobao.Design and implementation of intelligent speech sentiment analysis system [J]. Industrial Control Computer,2020,33(02):31-32.

[3] Wu Zhuoheng, Zhao Jiaxi, Shi Xiaofang.Analysis and design of speech emotion recognition system based on BP neural network [J]. Computer Knowledge and Technology,2022,18(10):76-79.

[4] CHU Yu,LI Tiangang,YE Shuo,YE Guangming. Feature selection method in speech emotion recognition [J]. Applied Acoustics, 2019: 1-7

[5] SUN Ying, V. Werner, ZHANG Xue-ying. A Robust Feature Extraction Approach Based on an Auditory Model for Classification ofSpeech and Expressiveness [J]. Journal of Central South University, 2012,19(02):504-510.

[6] WANG Haikun,PAN Jia,LIU Cong. Research progress and prospect of speech recognition technology [J]. Telecommunications Science,2018,34(02):1-11

[7] HONG Zhaojin, WEI Chenyang, ZHUANG Yuan, et al. Speech emotion recognition and personality analysis based on deep neural network [J]. Information Research,2020,46(01):48-53.

[8] Tao Jianhua, Chen Junjie, Li Yongwei.Review of speech emotion recognition [J]. Signal Processing,2023,39(04):571-587.