ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0011

Researchon Emotion Management based on Speech Signal Analysis Technology

Xiuyuan Zhao[1, *], Hanxi Wu[2, a], Zhanbao Xu[3, b], Zhenyu Zhang[4, c]

[1] University of Electronic Science and Technology of China, Chengdu, Sichuan, China

[2] Beijing University of Posts and Telecommunications, Beijing, China

[3] Southwest Jiaotong University, Chengdu, Sichuan, China

[4] Liaocheng NO.1 Senior High School, Liaocheng, Shandong, China

[a]2900829400@qq.com, [b]939485129@qq.com, [c]1548150514@qq.com

[*]Corresponding author: 1625334680@qq.com

These authors contributed equally to this work

Abstract

As we all know, Emotionalization will not only have a negative impact on people's body and mind, but also make people make mistakes in judgment. If there is a portable "management system", which based on human speech. And then, through the mathematical parameter extraction of its speech signal, the formant, MFCC and other parameters are obtained. Finally, it is compared with the sample signal of speech database and classified by KNN algorithm, and the emotion classification of the test speech signal is obtained. At the same time, we also designed a system framework for carrying voice input, emotion detection, cloud storage and other functions to realize a complete speech recognition and detection process.

Keywords

Emotion Management System; MFCC; KNN Algorithm; Cloud Database.

1. Background:

In today's environment, people generally live under high pressure and high intensity. Most of us face loneliness, depression, self-doubt and denial, low self-esteem, violent tendencies, and many other mental problems accumulated from experiences or bad emotions. Emotions, which can be produced at any time as long as a person is alive, seriously fill and control people's life and behavior. We want change So, we wanted to create a cloud that could recognize emotions and be with us all the time. It can be through the way of our emotions, such as subtle psychological changes and activities, such as something that leads to intense emotions, to collect, identify, try to control their own emotions, to relieve, dredge the severity of harmful emotions that individuals cannot think about or choose to escape. Therefore, in order to complete the expected results, we should first make a preliminary design of the speech recognition system. The following is the course of our experiment.

2. Establishment of Database

The emotional voices in this database are based on five basic emotional states: happy, anger, neutrality, sadness, and fear.

First, search for natural emotional voice from the Internet. This experiment uses CASIA library, a Chinese emotional voice library, to classify the emotional voice after listening to voice artificially.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0011

After listening to several emotional voices, select 300 clearly expressed and emotionally distinct voices as the database of this experiment, and extract and analyze their features.

3. Overall Design Process:

As we can see from the previous article, many different acoustic features can be extracted from emotional speech, and these characteristics can be used to characterize the emotional characteristics of the speaker. Therefore, how to select specific acoustic characteristics and how to extract acoustic features efficiently has become the factor to judge the merits of speech emotion recognition. In the past ten years, researchers in different fields have done a lot of research on what parameters in speech characterize emotion, from the perspective of psychology and linguistics, combined with the physical extraction of speech parameters. Therefore, this experiment will refer to the previous study, select the more suitable parameters for speech emotion recognition and screening.

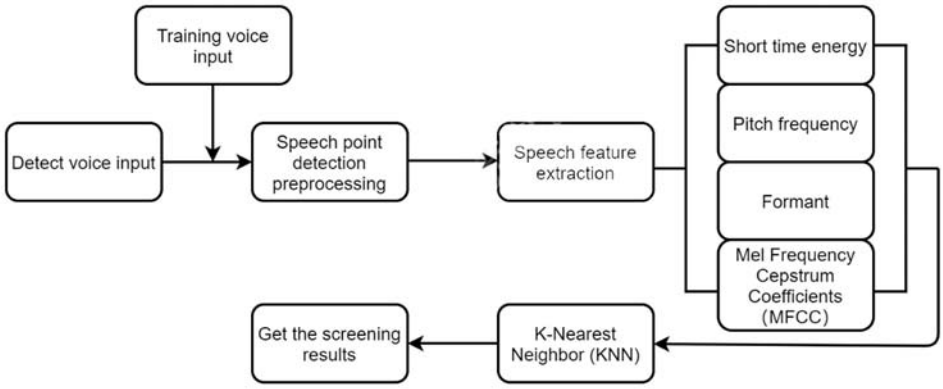

Figure 1. Emotional Recognition Flow Block diagram

In this experiment, the input phonetic signal is first pre-detected at the endpoint to extract the part of the voice signal that contains content for analysis. Secondly, the input voice signal and the standard emotional voice signal for training are extracted, and the characteristic parameters including short-term energy, resonance peak, gene frequency, Mel inverted spectrum coefficient (MFCC) are obtained, and finally k near-neighbor analysis (K) is used NN),calculates the Euclidean distance between the test signal and the training signal, and obtains the first K nearest neighbor parameter between the test signal and the training signal by setting the maximum number of nearest neighbor samples K. and filter the categories, and finally get the emotional category of the signal to be measured.

4. Principles of Speech Emotion Analysis

4.1 Classification of Speech Emotion Characteristics:

Traditional speech emotion characteristics can be divided into rhythmic features, sound quality features, and spectral-based correlation analysis characteristics. In recent years, while new speech features have emerged, several common types of features have become widely accepted and widely used in related experiments. Among them, rhyme characteristics mainly include base frequency-related characteristics, energy-related characteristics, resonance peak-related characteristics, etc. , rhythmic characteristics can generally convey more speech emotion information base audio rate characteristics in statements, and The related characteristics of spectrum are mainly considered to be the embodiment of the correlation between channel shape change and vocal motion, including linear prediction inverted spectral coefficient, Mel frequency inverted spectral coefficient, Mel spectrum energy dynamic system and so on.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0011

Table 1. Constructed phonological emotional features

Feature type	Specific features
Time structure	Short Time Average Cross zero ratio, Silent part time ratio
Amplitude structure	Short time average energy,Short time energy change rate Short time average amplitude,Average change rate of amplitude Short time maximum amplitude
Fundamental frequency structure	Maximum value of fundamental frequency trajectory curve Fundamental frequency average Average rate of change,Mean square deviation
Formant structure	Maximum, mean and average change rate of first, second and third formant frequencies
MFCC coefficient	12 order MFCC coefficient, first-order difference MFCC, second-order difference MFCC

4.2 Voice Emotion Feature Selection:

Depending on the various classifications of language emotional characteristics, we can know, such as base audio rate, resonance peak, Mer inverted spectrum coefficient (MFCC), etc. These are important speech features, which are widely and important in speech enhancement, speech synthesis, speech recognition, emotion recognition, sound source positioning and other fields. In this experiment, several speech characteristics are used to analyze:

A based audio rate: It is one of the important parameters to describe the excitation source in processing speech signals, the human sound signal is divided into clear and turbid sound, where turbid sound needs periodic vibration of the vocal cords, so it has obvious periodicity, solid can be used for detection. It mainly contains the envelope of the characteristic information of the base audio rate and the linear prediction coefficient of the base audio rate.

Table 2. The bass frequencies between different emotional voices

Emo	Angry	Fear	Happy	Neutral	Sad	Surprise
Mean Base audio frequency	229.6	158.3	260.5	138.5	165.6	183.1

B Short-term energy: It is a function that measures changes in the amplitude value of a voice signal. Because the energy of the voice signal changes over time, the energy difference between clear and turbid sounds is significant. Therefore, for short-term energy machine jitter analysis and linear regression analysis, can be very good for the detection of speech emotion.

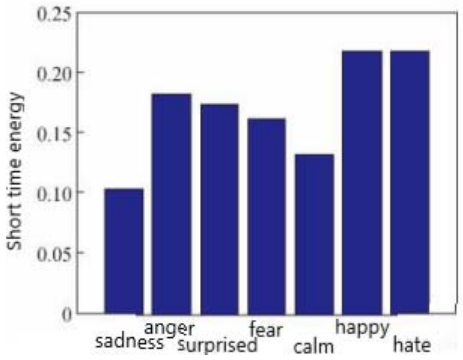

Figure 2. Comparison diagram of short-time energy average value

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0011

C. Short-term average zeroing rate: By analyzing the speech signal, it is found that, although the channel has some resonant peaks, its speech energy is generally concentrated below 3KHz due to the high matching drop of the spectrum caused by the sound gate wave, while most of the energy appears on higher frequencies when the sound is cleared, because high frequency means high short-term Average zero-crossing rate, low frequency means low average zero-crossing rate, so it can be considered that turbid sound has a lower zero-crossing rate, and clear sound is the opposite. This relationship is only relative, i.e. there is no precise value to quantify. In this experiment, we can use the short-term average zero rate not only to judge the clear tone, turbid tone to assist in the detection of speech emotion, but also to use it from the background noise to find out the voice signal, to judge the silence of the beginning and end of the segment, this part of the content will be detailed in the later VAD endpoint detection.

D Resonance peak: Sound is filtered by the cavity, so that the energy of different frequencies in the frequency domain is redistributed, partly because the resonance of the resonance cavity is strengthened, the other part is attenuated, the solid part is called resonance peak. It is a detailed description of the rhythmic characteristics, an important feature that reflects the resonant characteristics of the channel, and represents the most direct source of pronunciation information.

Table 3. The resonance peaks of emotional speech

Emo	Angry	Fear	Happy	Neutral	Sad	Surprise
Mean Resonance Peak	2780	3715	3166	1737	1054	2162

E Mel Frequency Inverted Spectrum Coefficient (MFCC): MFCC is an inverted spectral parameter extracted from the Mel frequency scale domain that can be analyzed by the auditory principle of the human ear, which is comparable to the normal converse inverted spectrum. The frequency band of the linear interval in is more similar to the human auditory system. Such nonlinear representations can make sound signals better represented in many fields, that is, they represent the main spectral characteristics of sound signals. The MFCC first-order differential parameters associated with it also have good characterization capabilities.

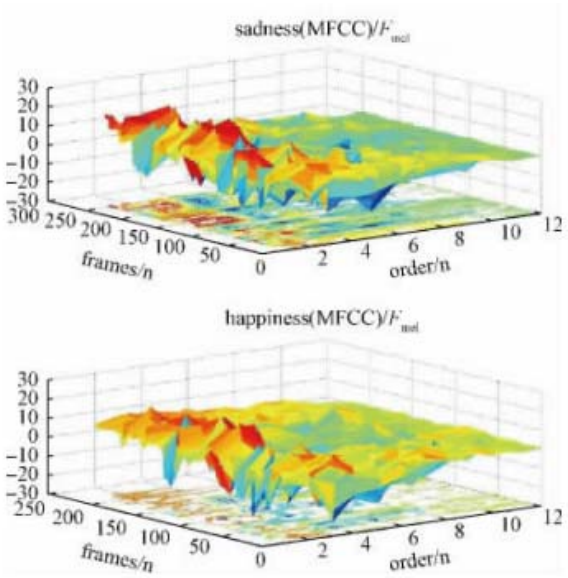

Figure 3. The signal features of speech emotion feature extraction (MFCC)

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0011

The above figure is the characteristic signal diagram of the speech waveform when you are sad and happy. After extracting the speech emotion features of MFCC, we can obviously find that there are great differences in the distribution of speech features of different emotions in the dimensions of order and frame number. If you analyze a large number of corresponding sentences according to different emotions, you can get the characteristic parameters of different emotions, So as to establish a more accurate classification model.

4.3 Choice of Speech Recognition Algorithm

There are many mainstream classification recognition algorithms, including: decision trees, Bayesians, artificial neural networks, K-near neighbors, support vector machines, and classification based on association rules, as well as integrated learning algorithms for combining single classification methods, such as Bagging and Boosting and, secondly, intelligent optimization algorithms such as mixed frog jumping algorithms are also emerging. Various classification methods, it should be said that each has its own different characteristics and advantages and disadvantages. Which algorithm to use, need to refer to and compare the criteria of classification recognition algorithm. Such as :(1) the accuracy of the prediction. The ability of the model to correctly predict the class label of the new sample; Includes the time it takes to construct a model and classify it using it; The model's ability to correctly predict noise data or vacancy data; The ability to construct models effectively for data sets with large amounts of data; The simpler and easier the model description is to understand, the more popular it becomes.

Table 4. Comparison of various recognition algorithms

Recognition Algorithm	Fitting effect of speech emotion	Recognition rate	advantage	shortcoming
GMM	High	performs well in Aibo database and text library	Strong data fitting ability	Strong dependence on training data
SVM	Higher	The Berlin library performed well	Suitable for small sample training	A little deficiencies in multi class classification problems
KNN	Higher	The Berlin library was average	Easy to implement and consistent with emotional data distribution	Large amount of calculation
HMM	Average	The Berlin library performed well	Suitable for the recognition of time series	Greatly affected by phonemic information
Decision Tree	Average	General performance on Aibo database	Suitable for discrete emotion category implementation	Recognition rate needs to be improved
ANN	Higher	Generally expressed in Japanese emotional pronunciation	Approaching complex nonlinear relations	Easy to fall into local minima
Hybrid Leapfrog Algorithm	Higher	good performance in Chinese language emotion	Strong optimization ability, helpful to find potential patterns in emotional data	Easy to fall into local optimization in the later stage of iteration

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0011

K near-neighbor classification algorithm is one of the simplest methods in data mining classification technology, based on KNN algorithm has a good choice of classification of speech recognition, and easy to implement, concise description, solid use of this algorithm to achieve emotional recognition.

5. Speech Emotional Feature Extraction

5.1 Short-Term Energy-Related Parameter Extraction:

The extraction of short-term energy for voice signals generally involves the sharding and windowing. First, set the voice waveform time domain signal as $x(n)$, because the voice signal is a quasi-steady state signal, divide it into shorter frames, can be considered a steady state signal in each frame, and use steady state signal processing methods to process them. Therefore, we gather N sampling points into an observation unit, which is called a frame, and N is called the frame length. At the same time, to avoid the transition caused by too large a parameter difference between one frame and another, the transition is usually caused by overlapping parts between the two frames, and the displacement of the latter frame to the previous frame is called frame shift. Typically, the frame length is set to 256and the frame moves to 128. Therefore, for $x(n)$, the i-frame voice signal obtained by windowing function $\omega(n)$ and framing is $y_i(n)$, $y_i(n)$ satisfied.

$$y_i(n) = \omega(n) * x((i-1) * inc + n) \qquad 1 \le n \le L, 1 \le i \le fn$$

'L' is the frame length, 'inc' is the frame shift length, 'fn' is the total number of frames after the split frame, and the window function $\omega(n)$ is generally rectangular window or Hemming window, because the rectangular window main valve is narrow, the frequency resolution is higher , but the side flaps are also high, interference is serious, and Hanming window base Ben in contrast, solid-based experiment on the voice signal of the pre-added windows are using Hanming window. Therefore, the short-term energy formula for calculating the voice signal $y_i(n)$ in frame 'i' can be expressed as:

$$E(i) = \sum_{n=0}^{L-1} y_i^2(n) \qquad 1 \le i \le fn$$

Therefore, after obtaining the energy of each frame of the speech signal, the maximum value, mean, variance, etc. of the short-term energy of the voice signal can be filtered out for subsequent feature comparison, and the jitter rate of the short-term energy and the linear regression coefficient of the short-term energy can also be calculated to characterize the parameters of the speech signal.

5.2 Short-Term Average Zero-Crossing Rate Extraction:

In the case of discrete time voice signals, zero crossing occurs if adjacent samples have different algebraic symbols. The number of times a unit of time has passed zero is called a zero-crossing rate. The short-term average zero-crossing rate represents the number of times a voice signal waveform in a frame of speech passes through the horizontal axis (zero level), which is the number of times a sample value can be thought of as changing the symbol.

In this experiment, after we have defined the speech signal $y_i(n)$ after the framing, we still set the frame length to L, and the short-term average zero-crossing rate can be expressed as:

$$Z(i) = \frac{1}{2} \sum_{n=0}^{L-1} |\mathrm{sgn}[y_i(n)] - \mathrm{sgn}[y_i(n-1)]| \qquad 1 \le i \le fn$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0011

Where 'sgn' is a symbolic function, that is

$$\operatorname{sgn}[x] = \begin{cases} 1(x \geq 0) \\ -1(x < 0) \end{cases}$$

5.3 Base Audio Rate and Related Parameter Extraction:

The detection and extraction of base audio rate has always been an important research topic, for which scientists have proposed a variety of genetic detection algorithms, for pure speech signal genetic detection methods, can be broadly divided into: auto-correlation function method (ACF), average amplitude difference function method (AMDF),inverted spectrometry, linear prediction method, wavelet method and so on. In this experiment, the gene frequency of speech signal is extracted by inverted spectrometry.

For previously defined voice signal sequences $x(n)$, its Fourier transforms to:

$$X(\omega) = FT[x(n)]$$

the sequence:

$$\hat{x}(n) = FT^{-1}[\ln | X(\omega) |]$$

Among them, $\hat{x}(n)$ is called cepstrum, that is, the cepstrum sequence of $x(n)$ is the inverse Fourier transform of the logarithm of $x(n)$ amplitude spectrum. The dimension of $\hat{x}(n)$ is Quefrency, which is called inverse frequency, its actual unit or time unit.

For speech signals, it can be thought of as being filtered by channel response $u(n)$ by sound gate pulse $v(n)$ excitation,it expressed as:

$$x(n) = u(n) * v(n)$$

In the frequency domain, the relationship between these three quantities can be written as:

$$\hat{x}(n) = \hat{u}(n) + \hat{v}(n)$$

It can be obtained that in the frequency domain, $u(n)$ and $v(n)$ are relatively separated, shows that the sound pulse cepstrum containing the pitch information can be separated from the vocal tract response cepstrum, separated from it and recovered $u(n)$.The specific approach of this experiment is to find the maximum value of the inverted spectral function between the inverted spectral $P_{\min} \sim P_{\max}$, and the maximum sample point of the inverted spectral function is the base tone period of the i frame speech signal $T_0(i)$. And the base sound period can be easily extracted by using the most valuable function in MATLAB. Finally, the gene frequency can be obtained by dividing the sampling frequency. In addition, calculating the maximum, mean, variance, and first-order jitter of a gene frequency can be used more precisely as parameters for analysis reference.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0011

5.4 Resonance Peak-Related Parameter Extraction:

Resonance peak information is contained in the envelope of the language audio spectrum, the key to extract its parameters is to estimate the envelope of the natural language audio spectrum, and consider the maximum value in the spectrum envelope to be the resonant peak. Similar to extracting the base tone, the extraction of resonance peaks is difficult to be precise, and the traditional and commonly used methods of estimating resonance peak parameters can be broadly divided into: LPC method, Hilbert-Huang transformation method (HHT), inverted spectrometry, etc. This experiment uses the LPC rooting method to estimate the resonance peak parameters.

Based on the principle of linear prediction and analysis, we can think that the sound gate pulse excitation $u(n)$ can be restored to a speech signal by a transfer function $H(z)$, and the channel transfer function $H(z)$ can be expressed by using the full pole model:

$$H(z) = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}}$$

Where' a_i 'p represents the order and G is the gain factor (that is, the parameters of the model), that is, Linear Prediction Coeffic (LPC). In addition, because the signal is actually objective, the use of models to fit is bound to have errors, so, here $\hat{x}(n)$ is generally introduced as the estimated value of signal $x(n)$, and the difference between $x(n)$ and $\hat{x}(n)$ is said to be the linear prediction error. The filter of the prediction error system $A(z)$ can be expressed as:

$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i}$$

And the polynomial coefficient decomposition of $A(z)$ precisely the central frequency and bandwidth of resonance peaks. The root of the polynomial is generally obtained using the function roots in MATLAB. (Mostly The roots in $A(z)$ are usually complex conjugate pairs:

That is, let $z_i = r_i e^{j\theta_i}$ be a certain kind of complex root obtained by roots, and the formant frequency corresponding to z_i is denoted as F_i, the 3dB bandwidth is set to B_i , and T is the sampling period, the relationship can be expressed as:

$$F_i = \theta_i / 2\pi T$$

$$B_i = -\ln r_i / \pi T$$

In the actual program, we require the frequency of resonant peaks to be within 150Hz-3,400Hz to exclude conjugate complex roots of some non-resonant peaks, and set up a maximum output of 4 roots with an output of NaN when in sufficient.

5.5 Mel Frequency Cepstrum Coefficient and its Related Coefficient Extraction:

Mel inverted spectral coefficient uses the non-linear relationship between the human ear's perception of sound and frequency, and obtains a set of spectral characteristics, the corresponding relationship of frequency is:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0011

$$Mel(f) = 2595\log_{10}(1+\frac{f}{700})$$

For the extraction of MFCC coefficients, the following block diagram can be used for the Mel filter in the speech inverted spectrum analysis:

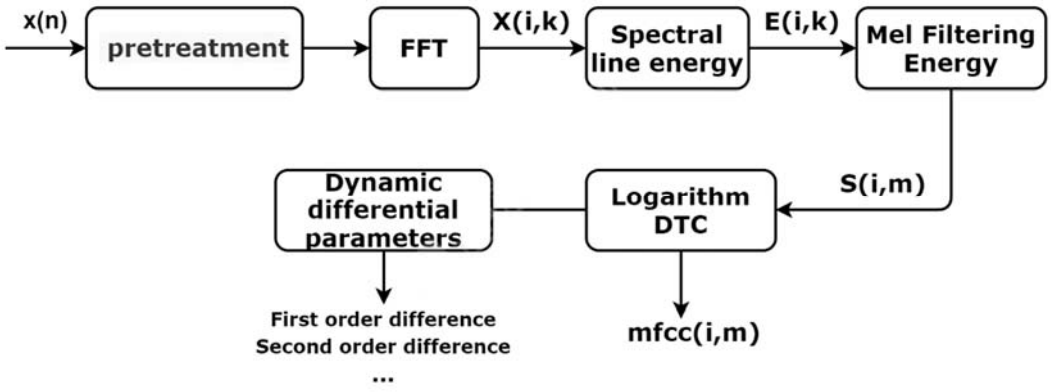

Figure 4. MFCC and correlation coefficient extraction process

(1) Pre-processing:

a. Pre-emphasis: The purpose of pre-emphasis is to raise the high-frequency portion, flatten the spectrum of the signal, and maintain the spectrum in the entire band from low to high frequency, using the same signal-to-noise ratio. At the same time, it is also to eliminate the effect of vocal cords and lips in the process of occurrence, to compensate for the high frequency part of the speech signal inhibited by the pronunciation system, therefore, pre-emphasis is commonly used a high-pass filter to achieve, this experiment set it as:

$$H(z) = 1 - az^{-1}$$

Where 'a' is a constant, usually between 0 .9-1.0, set to 0.94 in this experiment.

b. Framed and windowed:

According to the previous mentioned, the voice signal is framed, and in order to reduce the leakage of the signal in the frequency domain, it is necessary to add windows to each frame of speech, in this experiment, the Use of Heming window as a window function.

The voice signal sequence $x(n)$ is preprocessed to $x_i(m)$, where 'i' represents the i-th frame after the split frame.

(2) Fast Fourier Transformation (FFT).

A rapid Fourier transformation is carried out on each frame signal, the purpose of which is to transform the time domain data into the frequency domain data, and to obtain the spectrum of each frame signal, expressed as:

$$X(i,k) = FFT[x_i(m)]$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0011

(3) Calculate spectral energy:

The spectral energy is calculated for each frame of F FT-transformed data to obtain the energy distribution:

$$E(i,k) = |X(i,k)|^2$$

Where 'i' represents frame i and 'k' represents the k curve in the frequency domain.

(4) Calculate the energy through the Mel filter:

The calculated energy spectrum for each frame spectrum is calculated by the Mel filter through the Mel filter, whereas in the frequency domain, the energy spectrum $E(i,k)$ equivalent to multiplying each frame by the frequency domain response of the Mel filter $H_m(k)$ and adding up, i.e.

$$S(i,m) = \sum_{k=0}^{N-1} E(i,k) H_m(k) \qquad 0 \le m < M$$

Where M represents the number of Mel filters, in MATLAB, we can use the melbankm function to quickly get a matrix of Melfilter groups, where t represents the use of triangular window functions.

(5) Calculate the DCT inversion:

DCT, or Diskrete Cosine ransform, has the characteristics of rich signal spectrum component, energy concentration, and so on, with a larger voice enhancement effect. We know that the FFT inversion of the sequence $x(n)$ can be written as:

$$\hat{x}(n) = FT^{-1}[\ln\{FT[x(n)]\}] = FT^{-1}[\ln\{X(k)\}]$$

The DCT of the sequence $x(n)$ can be written as:

$$X(k) = \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} C(k) x(n) \cos[\frac{\pi(2n+1)k}{2N}] \qquad k = 0,1,\cdots,N-1$$

Where 'N' is the length of the sequence $x(n)$, and C(k) is the orthogonal factor, it can be expressed as:

$$C(k) = \begin{cases} \sqrt{2}/2 & (k=0) \\ 1 & (k=1,2,\cdots,N-1) \end{cases}$$

Like the cepstrum of FFT, the cepstrum of DCT can be calculated by taking the logarithm of the energy of the Mel filter similarly:

$$mfcc(i,n) = \sqrt{\frac{2}{M}} \sum_{m=0}^{M-1} \log[S(i,m)] \cos[\frac{\pi n(2m-1)}{2M}]$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0011

Where the energy of the Mel filter $S(i,m)$ mentioned above is referred to above, M refers to the m-th Mel filter, 'i' refers to the i-th frame, and N refers to the number of spectral lines after DCT.

(6) Dynamic differential parameter:

The standard inverted spectrum parameter MFCC only reflects the static characteristics of speech parameters, and the dynamic characteristics of speech can be described by the differential spectrum of these static features. Experiments show that combining dynamic and static features can effectively improve the recognition performance of the system, so the MFCC parameters extracted above are generally calculated as differential parameters:

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

In, d_t represents the t-th first-order difference, C_t is the t-th inverted spectrum coefficient, and Q represents the order of the inverted spectral coefficient, K Represents the time difference of the first derivative, which is taken as 1 in this experiment. Substitute the results of the upper class to obtain the parameters of the second-order difference.

6. Speech Emotion Recognition System Design

6.1 Endpoint Detection

VAD, or voice endpoint detection technology, is primarily tasked with accurately locating the start and end points of speech from noisy speech. In general, speech contains a long mute, by separateing mute from actual speech, making it easier to extract and detect other speech data later, so VAD is one of the key technologies in speech signal processing, and its quality will directly affect success or failure.

In this experiment, the two-gate method is used to detect the speech endpoint, that is: using the zero-crossing rate to detect the clear tone, with short-term energy to detect the turbid tone, the two cooperate with each other. First of all, the two are set up two thresholds, E_low, E_high that is, short-term energy threshold, Z_low, Z_high that is, zero-crossing threshold, lower threshold threshold is small, sensitive to change signals, high threshold threshold is high, not easy to achieve. We believe that when E (x)> E_high and zcr > Z_high, And then several consecutive frames meet this condition, is considered to be the starting point of speech, the same can also get the speech end of the algorithm.

6.2 K NN-based Recognition Algorithm:

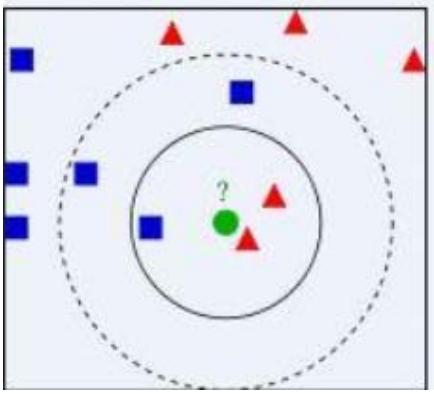

Figure 5. K-Nearest Neighbor algorithm

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0011

Proximity algorithm, or k-Nearest Neighbor classification algorithm, is one of the simplest methods of data mining classification technology. The so-called K nearest neighbor, is the meaning of k nearest neighbor, that is, each sample can be represented by its closest k neighbor. If most of the k closest samples in the feature space belong to a category, the sample is also classified into this category. Therefore, in the KNN algorithm, the neighbors selected are objects that have been classified correctly. The method determines the category to be subdivided by the category of one or more samples closest to it in the class decision.

A specific description of the K-neighbor algorithm can be broadly divided into the following steps:

(1) Calculate the distance between the test data and the individual training data, where the distance generally refers to the Euclidean distance, and under the multidimensional array, the distance formula is expressed as:(when multiple dimensions are different from each other).

$$E(x,y)=\sqrt{\sum_{i=0}^{n}(x_i-y_i)^2}$$

In MATLAB, you can use the norm function to get the Euclidean distance between two multidimensional arrays directly.

(2) Sort by the increment of distance.

(3) Select the smallest distance of K points. Here K represents the number of proximity, that is, taking a few near points to predict when predicting the target point. If the value of K is too small, the presence of noise component will have a relatively large impact on the prediction, if the value of K is taken too large, it is equivalent to using training examples in a larger neighborhood to predict, the approximate error of learning will increase. Instances farther away from the input target point also play a role in the prediction, causing the prediction to go wrong. Therefore, according to the experience of other researchers, the average k value is not more than 20, the upper limit is the open side of n, as the data set increases, the value of K also increases.

(4) Determine how often the first K points appear in the category.

(5) Returns the most frequently occurring category of the previous K points as a predictive classification of the test data.

In the actual program, the characteristic extraction of the training sample set is generally combined with the feature of the sample to be tested, the total feature vector set is obtained, and then the training sample set and the sample set to be tested are divided, which can improve the efficiency of the test.

7. Experimental Test Results:

After the overall system is built, the system can be tested. In this experiment, we use the speech database mentioned above to select part of the speech and divide it into two categories. One is for testing and the other is for training. The speech of the test part is used to simulate the input speech signal. After entering the system, the speech parameters are extracted together with the training signal, and the measured signal is classified by KNN classification algorithm, Finally, the accuracy of classification is verified.

At the same time, we know that the K value of k-nearest neighbor algorithm has a great impact on the accuracy of prediction results, and the selection of K value is generally set artificially. Therefore, in order to select the best K value, we use cyclic algorithm to classify and detect different K values in turn, showing the recognition accuracy of various emotion classifications under different K values, The best K value is selected as the fixed value of speech emotion recognition system. The

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0011

approximate selection of K value depends on the number of samples. The range of K value selected in this validation experiment is 7 < K < 15.

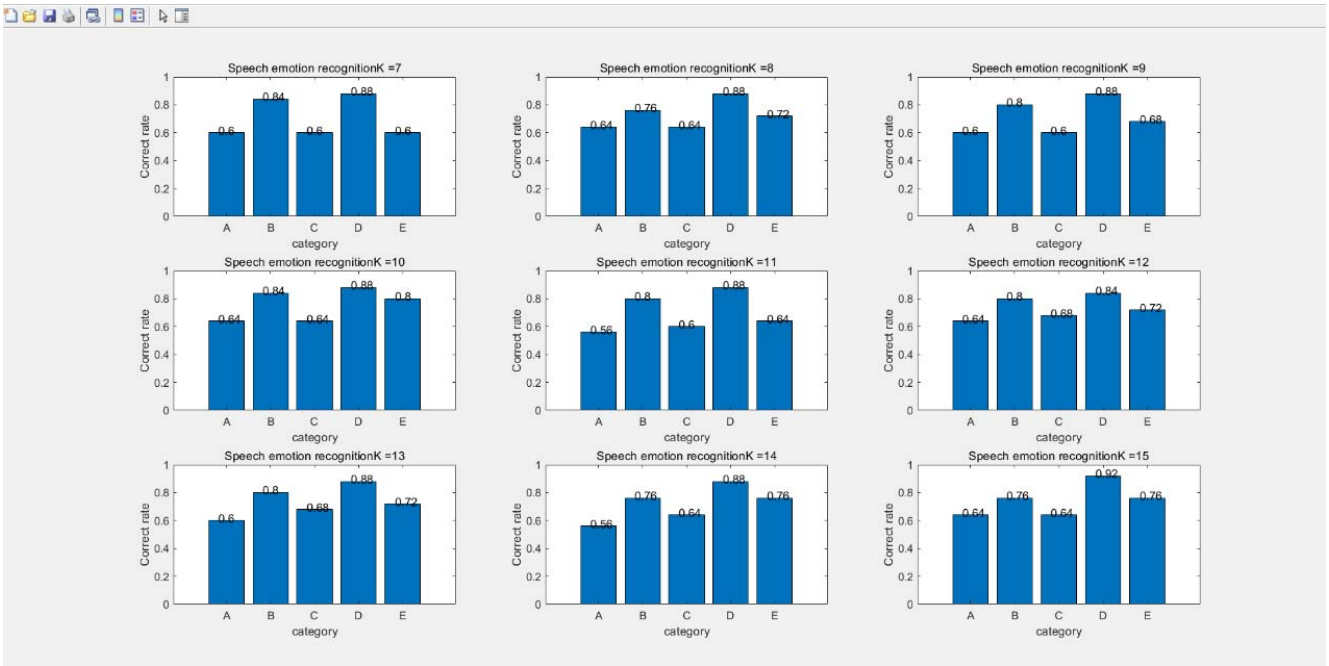

Figure 6. Emotion recognition accuracy test results

According to the detection results, the K value shows good selection correctness and stability at 10, and the detection rate of various speech emotions is basically as high as 70%. In addition to verifying the contribution of various speech emotion parameters and KNN classification algorithm to emotion classification, it also shows the success of this experimental system, that is, it proves that this algorithm can basically meet the requirements of emotion analysis.

8. Overall Demand

As mentioned earlier, it is necessary to design and implement a portable cloud emotion recognition management system based on acoustic feature analysis. By distinguishing the sound signals emitted in different environments of the human body, the system can detect, remind, and guide people to manage their emotions in time. At the same time, it detects emotional changes and transfers the data to the cloud database, allowing the system to provide customized suggestions to different people.

The system functions are as follows:

1) Analyze the entire recording file and extract meaningful statistical features.

2) According to the statistical parameters, the effective recording is identified from the recording file.

3) Analyze and classify the speaker's emotional attitude (joy, anger, surprise, sadness, fear and neutral).

4) Write other identified complex statistical parameters into the cloud database, so as to further extract information from the voice signal and provide data reserves for other people's sentimental analysis.

9. System Overall Architecture Design

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0011

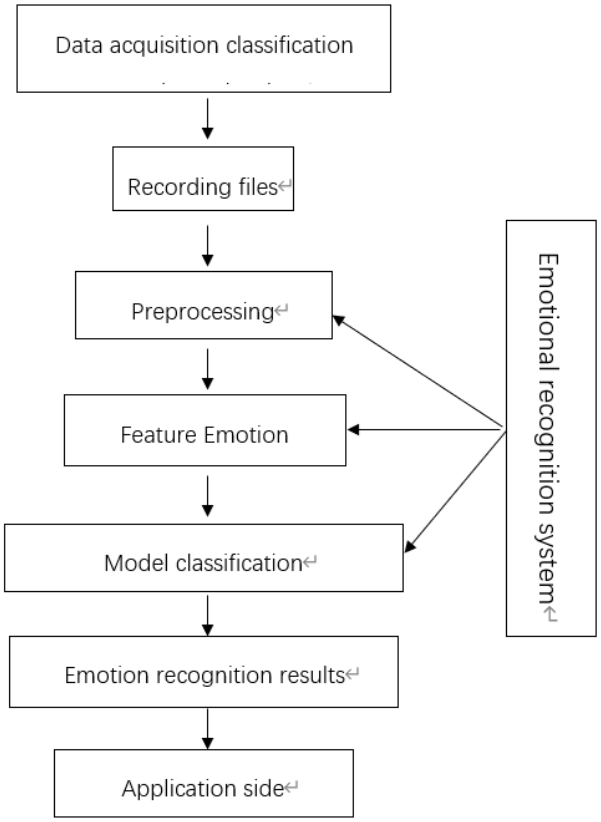

Figure 7. System design block diagram

The data collection equipment converts the collected voice signals into recording files and inputs them into the emotion recognition system.

The feature parameter extraction module is mainly responsible for signal processing of recording files, extracting feature parameters that can reflect the characteristics of emotions (speech speed, intelligibility, average pitch, pitch change, pitch range, intensity, sound quality), training of the emotion recognition system and predictions depend on the work of this module.

The training and classification module uses traditional classification methods or convolutional neural networks to automatically learn data feature classification methods to perform emotion recognition on speech. Traditional classification methods require a lot of energy to extract representative features, and require high professional knowledge and selection of features. Convolutional neural networks do not require manual intervention, and can automatically extract features that are helpful for classification and directly output them to the classification device.

The results output by the emotion recognition system will be displayed on the application side, supplemented by personalized emotion management suggestions. At the same time, the application transfers the data to the cloud to improve the voice emotion database.

10. System Use Process Design

1) Establish an emotional corpus. Use a corpus suitable for Chinese speech recognition and integrate the data generated during the use of the system.

2) Voice enhancement. In the actual environment, there are many sources of noise. Use speech enhancement technology to eliminate the noise in the signal as much as possible, thereby improving system performance.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0011

3) Feature extraction. Before feature extraction, signal preprocessing is required. Perform emotional feature extraction on the preprocessed speech signal to make it meet the training and learning requirements.

4) Emotion recognition system can be divided into emotion recognition training stage and emotion recognition prediction stage. The main function of the emotion recognition training stage is to train the emotion recognition classifier. The input is a sample set of audio files that have been manually classified, and the output is a trained emotional classifier. The emotion recognition prediction stage generates emotion recognition results and statistical data through the analysis and recognition of the recording files. The input is a recording file that has not been manually classified. For each individual recording file, the module generates emotion recognition results and related statistical information.

In general, the system uses the extracted features, selects a suitable model, trains and learns the speech samples, and obtains a training model. After the speech samples are output from the training model, they are classified using a classifier, and finally the result of sentiment analysis is obtained.

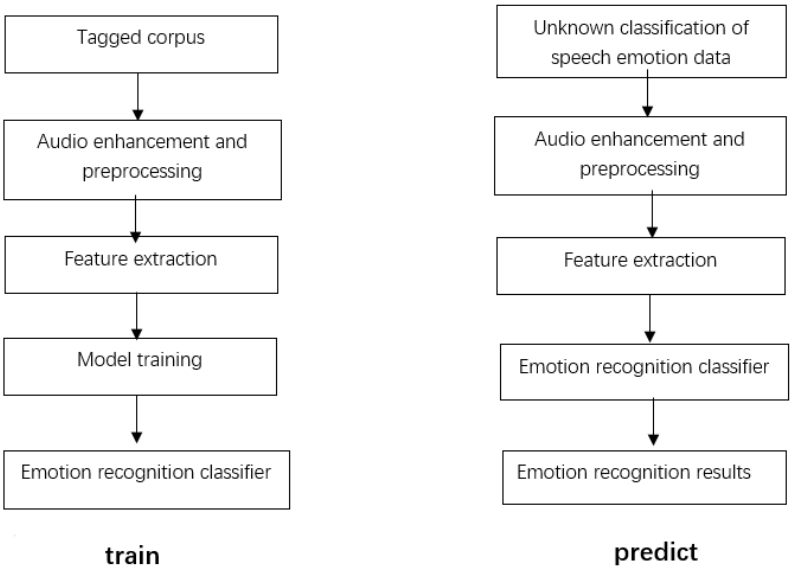

Figure 8. System use process design

11. Conclusion

This is the research result that we have learned to the best of our ability, which can effectively solve people's emotional problems. However, we are well aware of the limitations and certain negative effects of such technology products -- people's emotions should be solved through human behavior as much as possible, and we should not rely on some technology to suppress and control our humanity and emotions. We still have to try our best to use their own strength to fight, I hope you can pay attention to this problem, try to be familiar with their own emotions, learn to control their own emotions.

References

[1] Xingjun Yang, Huisheng Chi, etc., Digital Processing of Speech Signals, Electronic Industry Press, 1995.

[2] Zhiping Wang, Analysis and Recognition of Emotional Features in Speech Signals, Southeast University, Radio Engineering.

[3] Daning Jiang, Lianhong Cai, Chinese Emotional Speech Classification Based on Rhyme Features, Department of Computer Science, Tsinghua University.

[4] Jianxia Chen, A Review of Speech Emotion Recognition, Department of Computer Science, Xiamen University.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0011

[5] Bo Xie et al, Statistical analysis of Mandarin emotional speech database and its rhythmic features, Zhejiang University, Computer Science and Technology School.

[6] Li H, Xu S-L, Wu G-X, Ding CH-Y, Zhao X-M. Research on speech emotion feature extraction based on MFCC[J]. Journal of Electronic Measurement and Instrumentation, 2017, 31(03):448-453.

[7] Liu Z. To, Xu J. P., Wu M., Cao W. H., Chen L. F., Ding X. W., Hao M., Xie Q. Q., A review of speech emotion feature extraction and its dimensionality reduction methods ,2017, Vol. 40.

[8] Wang W, Yang L, Wei L, Liu Y. Extraction and analysis of speech emotion features[J]. Laboratory Research and Exploration,2013,32(07):91-94+191.

[9] Yongao Zhang, Qingyu Ma, Qing Sun. Research on speech emotion analysis based on MFCC and CHMM techniques and its application in education[J]. Journal of Nanjing Normal University (Engineering Technology Edition),2009,9(02):89-92.

[10] Buchen Tang, Ruiyu Liang, Jie Wang. Experimental design and implementation of speech emotion recognition for undergraduate education[J]. University Education,2018(09):105-107.

[11] Praseetha, V. M.,Joby, P. P.. Speech emotion recognition using data augmentation[J]. International Journal of Speech Technology,2021(prepublish).

[12] Yuanlu Kuang. Speech emotion recognition based on HMM and RBF hybrid models [D]. Hunan University,2013.

[13] Pengjuan Guo. Speech emotion feature extraction method and emotion recognition research[D]. Northwestern Polytechnic University, 2007.

[14] Zhiyong Song. Application of MATLAB in speech signal analysis and synthesis [M]. Beijing: Beijing University of Aeronautics and Astronautics Press, 2013.

[15] LAN S K, SHI Y B. An improved algorithm for mfcc parameters in speaker recognition system[J].Journal of Luoyang Institute of Science and Technology: Natural Science Edition, 2013, 23(4) : 23-24.

[16] Yi Yang, Zewei Li, Beixing Deng, Xiaohong Ma. Reform and practice of speech signal processing experiments [J]. Laboratory Research and Exploration, 2014 (4): 123-126.

[17] Alan V. Oppenheim, Ronald W. Schafer Digital Signal Processing[M]. Prentice-Hall Inc. ,1975.

[18] Kai Du. LPC analysis of resonant peaks of speech signals [J]. Journal of Harbin Normal University, 1998, 14(2): 49-52.

[19] Song P et al. "Speech emotion recognition method based on LDA+kernel-KNNFLC." (2015). [20] Müller V C. Fundamental issues of artificial intelligence. Cham, Switzerland: Springer, 2016.

[20] Song R. Acoustics feature based Chinese speech mood recognition system, 2014.

[21] Zhang R.F. The Design and Implementation of Old People Speech Emotion Recognition, 2018.

[22] Hao X.L. Design and Implementation of Speech Emotion Recognition System Based on Android, 2018.

[23] Zhao L. Speech Signal Processing, Mechanical Industry Press, 2016.