Emotional state analysis based on speech recognition signals

Ji Qin

Tianjin University Tianjin, China

3235322359@qq.com

Abstract. At present, as intelligent speech recognition is widely used, the importance of speech emotion recognition technology is also gradually gaining recognition. Considering the current social pressure, the generation of negative emotions has become particularly common. Therefore, it is necessary to increase the recognition of different emotions by using the speech recognition system, which has been perfected, so as to play a role in anticipation and prevention. In this paper, we focus on the negative effects of negative emotions on people's lives, and use K-nearest classifier to identify different emotions, and accurately identify them, and prevent and monitor them. Finally, the feasibility of this method is confirmed by simulating the process.

Keywords. emotion recognition, speech recognition, negative emotion, K-nearest neighbor classifier.

1. Introduction

Mood and emotion, mainly refers to an inner emotional reaction that arises from whether objective things meet people's psychological needs, and it integrates a series of subjective psychological experiences formed by people's feelings, perceptions and behaviors [1]. Among them, emotional speech includes joy, anger, surprise, sadness, etc. At present, emotions can be divided into basic emotions and compound emotions, where basic emotions can be included as social emotions, physical emotions, cognitive emotions, and mental emotions, while compound emotions are the complex of two or more emotions that contain basic emotions [2], and are often heard in life as compound emotions [3]. In addition to this there is another way of classifying emotions as negative and positive emotions [4].

Through a study by Wan Lingling et al. on undergraduate college students enrolled in Hubei Academy of Fine Arts from September 2010 to September 2012, it was found that the incidence of mental illness in the study subjects in the negative emotion group was significantly higher than that in the non-negative emotion group, and the difference between the groups in the trial was statistically significant. Thus, for the college population, the generation of negative emotions significantly increases the incidence of psychological disorders, so there is an urgent need to provide active interventions and preventive measures for college students with negative emotions through external factors [5]. Therefore, researchers have chosen to deepen the already mature speech recognition technology so that it can accurately identify various emotions, mainly negative emotions, and detect and prevent them.

Essentially, recognizing the emotional state of a speaker from speech is a pattern recognition problem that requires classifying different emotions and extracting the features of emotional speech as the basis for technical implementation [6]. In fact, this concept of emotion computing was proposed by the Massachusetts Institute of Technology (MIT) in 1985 [7], and the idea has been receiving a lot of attention [8].

In this paper, we extract and save the feature vectors of different emotions in the training sample set, and then extract the corresponding feature vectors from the sample to be tested, and finally match the two sets of feature vectors to obtain the recognition results. In this experiment, negative emotions are extracted and reminded or managed when they reach a certain level, thus playing a role in the prevention of mental illness to a certain extent. Therefore, the significance of speech emotion recognition is based on the function of speech recognition, to identify the speaker's emotion and provide timely feedback and guidance. It is mainly used in two situations.

The first is when the speaker does not fully understand his or her emotional state at the moment and makes some regrettable decisions because of his or her "impulsiveness", if the speech detection is used, the speaker can understand his or her emotional state in time to avoid wrong decisions. Especially for college students, their state of mind is not very mature, so they often cannot get some negative emotions in time, which will not only lead to some heart diseases but also reduce the ability to fight some diseases [9], and depression is frequently seen in college students.

The second is that in the present time, with the increase in the number of only children, when they encounter some negative emotions, they do not have the opportunity to talk to their siblings or do not want to spread their negative energy, so they will choose to digest it internally, and then this way will undoubtedly lead to a deadly cycle, which means that thinking negatively alone will only increase the negative emotions. However, due to the emergence of voice recognition emotion management system, it can detect the owner's emotional state in time for timely emotional relief, which is more conducive to people's physical health. At the same time, as opposed to only children, their parents also need this kind of robot care with emotion recognition, facing the huge pressure of life, women have to put more time on work, so the parents can only be less and less company, and for the age of parents need more is to accompany, so the emotion robot can take up this responsibility, accompany the elderly chat, more This will also play a good role in the prevention of heart disease in the elderly.

2. System Block Diagram

The emergence of voice emotion recognition can not only play a good role in the prevention and monitoring of most people's heart diseases, but also be a kind of psychological guidance and troubleshooting for most empty nesters. The emotion recognition system can detect and relieve emotional changes in time, and secondly, it can be combined with heart disease treatment. Firstly, it detects and predicts the speaker's emotions based on the voice signal emotion management system, and then provides timely advice. Thus avoiding one's lack of understanding of one's emotions and making mistakes again and again. However, for the emotion detection system, it is necessary to record the speaker's speech in different emotions beforehand in order to provide a comparison for subsequent comparison, thus increasing the workload to some extent. There are also various treatment options for the intervention of negative emotions. For example, by combining a voice emotion recognition system with a cell phone client, the speaker's words can be used to determine the current emotional changes, and the phone can automatically alert the speaker when the accumulated time of negative emotions reaches a certain amount, and give appropriate solutions according to the accumulated time. "When the accumulated time is up to one week, the system can give more urgent response measures, such as "seek help from a psychiatrist or medication to mediate" and it has been found that the positive medical response

It has been found that a positive medical response helps people to mitigate the effects of negative emotions and improve their quality of life [10]. Therefore, it can largely solve the problem of extreme behavior due to the person's lack of understanding of the current negative emotional changes. The specific procedure is shown in Figure 1-1

In this process, there are 5 main steps. The first step is the extraction of the speech to be measured, which can be extracted and saved by computer or collection microphone for the desired speech. The second step is audio pre-processing, as there are more or less errors and other noises added in the collection process, so it is necessary to carry out preliminary processing of the speech to be measured to facilitate the subsequent work of extracting features. The third step is the extraction of feature parameters. Since the feature parameters of the speech with different emotions will be different, they need to be extracted for the subsequent classification work. The fourth step is to compare the database with the feature parameters to be measured using KNN algorithm, where the database plays a key role in the whole process, it is the collection and summarization of thousands of different emotions in the audio, so it also provides a great convenience for the detection work. Finally, the results are obtained

as well as the analysis and processing of the different results. When too many negative emotions are detected, the system should give certain alerts as well as corresponding countermeasures.

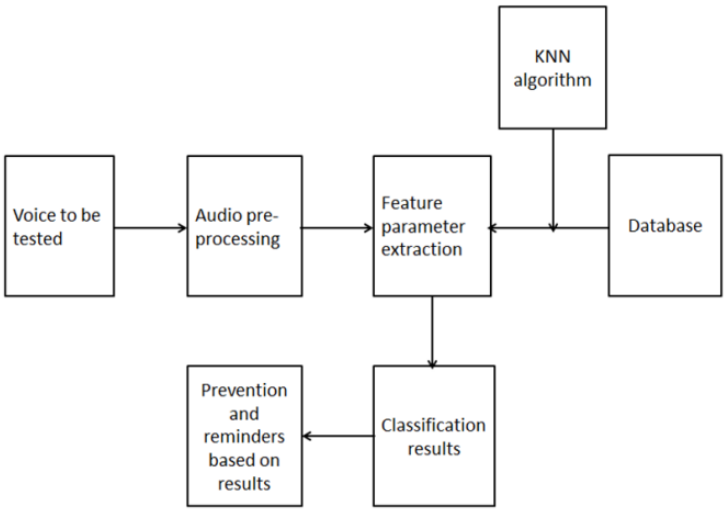

Fig. 1 Flow chart of negative emotion detection

3. System Design

By acquiring speech in different emotional states to generate emotional speech datasets or already existing speech databases, and extracting the feature parameters used to clearly distinguish emotions in these voices, and training the models to finally realize the recognition and detection of emotional states of college students or elderly people by voice.

The whole process of speech emotion recognition monitoring includes audio data collection, audio pre-processing, feature extraction, model training and recognition monitoring under different emotion states.

3.1. Audio Capture

Before the speech recognition, the work needs to be done is to collect the audio, and the accuracy and precision of the audio collection also plays a vital role in the speech recognition result. With the rapid development of speech recognition technology, speech emotion recognition technology has also gained more and more attention, so with the increase of demand, the establishment of speech emotion database has become particularly important, and also provides a great convenience for the subsequent research of speech emotion recognition.

At present, most of the audio acquisition devices are single microphone, dual microphone or microphone matrix. In the subsequent experiments, it is also necessary to choose different audio acquisition modes according to the specific algorithm requirements.

3.2. Voice Emotion Database

The speech emotion dataset is an important foundation for studying speech emotion recognition. The database is a collection of audio under different emotions in order to build a solid foundation for subsequent speech emotion recognition techniques. Among them, emotion recognition refers to the use of speech analysis and emotion monitoring system to analyze and process the signal dataset collected from the speech acquisition equipment to derive the emotional state that the other party is in, and also requires timely intervention and regulation if pessimistic emotions are detected.

The internationally popular speech emotion databases are AIBO (Artificial Intelligence Robot) corpus, VAM (The Vera am Mittag) database, Danish database, Berlin database, SUSAS (Speech under Simulated and Actual Stress) database, etc.

3.3. Audio data pre-processing

Since the collected audio data is not properly used in the subsequent speech recognition system, the database needs to be optimized to some extent, including cleaning, segmentation, reconstruction, normalization and denoising. Among these methods, the denoising of data is the most common, because it can largely avoid the interference of noise on the detection accuracy. The current denoising methods are mostly direct filters, including high-pass, low-pass and band-pass filtering, empirical modal decomposition, etc. The subsequent selection of the type of filter needs to be based on the database obtained and the thought of achieving a certain desired effect of recognition.

3.4. Feature Extraction

Feature extraction is a key step in the process of speech emotion recognition. Feature extraction, in short, is the extraction of speech parameters from the inherent database or from the audio information that will be collected again for use in later model training. There are many common speech parameters that can be used for sentiment analysis, among which the key feature parameters include energy [11], short-time energy and its derived parameters, fundamental and its derived parameters, resonance peaks and its derived parameters, Meier cepstral coefficient (MFCC), correlation dimension, maximum Lyapunov exponent, and Kolmogorov entropy.

According to the characteristics of speech signal with short-time smoothness, the speech signal can be processed to extract the required speech recognition feature parameters. Firstly, windowing and framing of speech signal can effectively utilize the short-time smoothness of speech recognition signal for feature extraction and analysis. Windowing is to multiply the original speech signal with a specific window function $\omega(n)$ to get the windowed speech signal $X\omega(n) = X(n)* \omega(n)$, so that the short-time sound can be considered to become a time-invariant smooth signal. The common features are as follows:

3.4.1 Short-time zero rate

The principle of short-time over-zero rate is to first assume that there are a total of N data in a frame of sound, and to compare the data symbols from the beginning with the adjacent ones, and if they are not the same, then calculate them as 1. Finally, the over-zero rate can be obtained by dividing the obtained value with the frame length. Although this indicator can effectively detect the fluctuation of the sound signal, the noise immunity is poor.

3.4.2 Short-time energy

Short-time energy is used to calculate the square of the sound data frame to get the energy level of the current frame. Therefore, when there is sound, the short-time energy is large, and when there is no sound, the short-time energy is small.

3.4.3 Linear prediction coefficient

The linear prediction coefficient treats the human vocal system as a linear time-invariant system, so that the current data can be calculated from the past data. The linear prediction accuracy is usually measured by the sum of squares of the predicted and actual values, so the smaller the sum of squares, the higher the accuracy of the prediction coefficient.

3.4.4 Merle spectrum coefficient

Although linear prediction coefficients have good applications in speech recognition, they do not work very well in other kinds of speech recognition. This is due to the fact that the auditory characteristics of the human ear are not linear, but a nonlinear system, i.e., the human ear responds well to certain frequencies of sound waves, but not to other frequencies of audio, so the Mel non-linear spectrum is mentioned, which means that it still exists linearly in that spectral range. Therefore, the Mayer spectral coefficients can be more consistent with the auditory characteristics of the human ear.

3.4.5 Mel inverse spectral coefficients

The Mel inversion coefficients are obtained by discrete cosine transformation of the Mel spectral coefficients, i.e., by reconverting the Mel spectral features to the time domain, where the Mel inversion coefficients are extracted by preprocessing, fast Fourier transform, mode squared, Mel filter set, logarithmic energy, and discrete cosine transformation.

Since the merit of emotion features plays a very important role in the final recognition effect of emotion, how to extract and select the speech feature results that can effectively reflect the change of emotion is one of the most important problems in the field of speech emotion recognition at present.

4. Model training with speech analysis and sentiment prediction

The following algorithms are commonly used for speech emotion recognition.

1) algorithms based on template matching, such as dynamic time regularization (DTW) and vector quantization (VQ), etc.

2) algorithms based on statistical probability models, such as Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM), etc.

3) Traditional machine learning based classifier algorithms, such as support vector machine (SVM), K-nearest neighbor classifier, etc.

4) Artificial neural network-based algorithms, such as convolutional neural networks (CNN) and recurrent neural networks (RNN), etc.

In this paper, we will use K nearest neighbor classifier as the main speech emotion recognition algorithm. kNN is one of the simpler algorithms among many speech emotion recognition algorithms, and its principle is to find the points similar to the sample within a certain k-value range, and if most of the points are similar to the sample points within this range, then the sample to be tested will have a clear emotional attribution. In fact, in simple terms, it is similar to saying that if two people look alike, it is because most of the features in the face are similar. So the KNN principle in it is the same.

Suppose the feature parameters of the classified samples are X, and the feature parameters of the training sample set of known categories are {X1, X2, X3, ..., Xn}; for the sample X to be tested, calculate its Euclidean distance D(X, XL) with each sample in { X1, X2, X3, ..., Xn }, L=1,2,3, ...,n, i.e.

$D(X,X_L)=(i=1N(X(i)-X_L(i))(X(i)-X_L(i)))$

where L=1,2,3,...,n

where N represents the dimensionality of the feature vector. min{D(X, XL)} is called the nearest neighbor of X, and the first K values of D(X, XL) arranged from smallest to largest are called the K nearest neighbors of X. The analysis of which category the largest number of K nearest neighbors belongs to is used to assign X to that category.

The KNN algorithm can be roughly divided into 4 steps. The first one is to extract the feature functions of the training samples by the feature extraction function and form the set of training sample feature vectors. The second step is to set the value of K in the algorithm, where K does not have a fixed value, so the appropriate K value can be selected through continuous debugging, etc. The K value selected in this paper is 9, which will give a more accurate classification result. The third step is to use the feature vector extraction function to extract the feature parameters from the samples to be tested, and to calculate the Euclidean distance between the vectors of the samples to be tested and each sample. Finally, the Euclidean distance needs to be counted, and finally the classification is obtained based on the statistical results.

Although it is said that the principle of KNN classifier utilizes the limit theorem to a certain extent, in the practical application only a small number of neighboring samples are considered , and it does not rely on the feature space region to which the calculated category belongs. Therefore, the KNN classifier is more advantageous in dealing with classification problems with a large number of overlapping category domains.

5. Test results

There is no clear standard for the classification of emotions, although in most cases, people prefer to classify emotions into 6 categories, which are happiness, interest, disgust, fear, pain and anger [12], while a few research institutes also classify them into 5, 6 or even 8 categories according to scientific research needs [13]. However, as of now for the research on emotion recognition of speech, it is preferred to classify emotions into 4 categories, including neutral, happy, sad and angry emotions. Therefore, in this test, the four emotions of frightened, happy, neutral, and angry in the speech database stock are mainly used to debug KNN technology respectively, and the following are the MATLAB recognition results of different emotion speech, as shown in Fig.2.

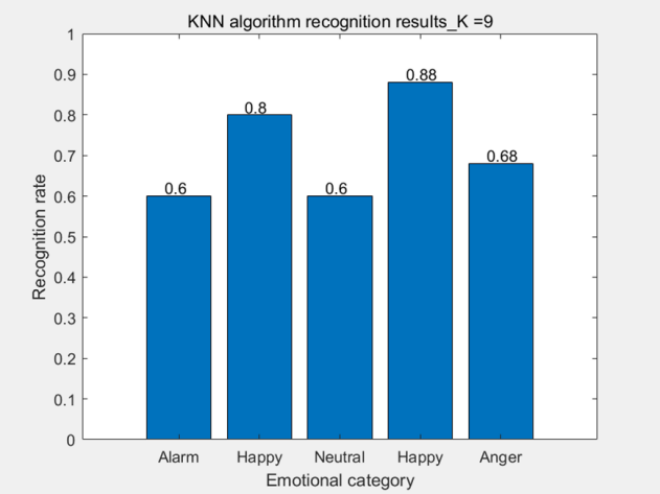

Fig. 2 MATLAB recognition results

From the simulation results, it can be seen that the recognition rate of machine learning speech emotion recognition using K-nearest neighbor classification reaches up to 88% when the amount of data is relatively small. At present, the KNN technique cannot reach 100% accuracy because the principle of K-nearest neighbor classification calculation takes the idea of mathematical limit, so it is not completely accurate. Secondly, there is also an emotional collection bias in the database collection, because the current various emotional speech libraries generally only include typical emotions and do not take into account the number of speech fragments included in various variations is not very neat, so it is not possible to obtain reliable emotional recognition results [14]. Therefore, in order to improve the recognition rate, a comprehensive emotional speech database needs to be established in the subsequent work, and the database should be updated continuously in order to keep up with the times.

In addition to the existing databases that can be used, one can also build one's own database according to the needs of scientific research, mainly by recording one's own or others' speech audio in different emotions through the microphone that comes with the computer or cell phone and saving and classifying them, and in order to make the experimental data more general, it is necessary to record a large number of and need to record different emotions without the knowledge of people [15]. However, if one wants to understand one's mood changes more accurately, one's own database is more convincing than a standard one.

6. Conclusion

As negative emotions increase, the role of identifying them becomes particularly important.

In this paper, the KNN algorithm is used to identify different emotions, mainly negative emotions. This algorithm can also be combined with cell phone clients, so that people can better understand their own mood fluctuations, and when faced with too much negative emotion and they do not know it, the system can be based on the cumulative time of negative emotion to judge and remind people, which can largely avoid the harm caused by negative emotion to everyone.

In the future, we hope to apply the voice emotion recognition technology in the future treatment of heart disease, at present, due to excessive stress, heart disease is more and more, but there are still some people are not willing to go to the hospital to receive treatment, so we can combine the voice recognition emotion management system with the cell phone app, so that we can carry out emotional prediction and management at home, in order to avoid whether to go to the hospital and produce psychological stress. In addition, it is possible to combine the voice emotion recognition system with a mobile app. In addition, the combination of voice emotion recognition system and robotics will also be a technological breakthrough. A robot that understands each other's emotions is undoubtedly a good playmate or an attentive companion, both for children's intellectual development games and for the increasing number of widows and orphans who come as companions.

At present, there are still many shortcomings in speech emotion recognition systems, so we need to continue to strengthen the systematic research of emotion recognition, but thankfully, we are using the already existing speech recognition technology as a guide to improve the reality of life and improve our lives.

References

[1] Li, H. S.. Psychology of Consciousness [M]. Beijing: China Legal Publishing House. 2017: 183-190.

[2] Buck R. Biological affects: a typology [J]. Psychological Review, 1999, 106 (2) : 301-336.

[3] Ye Yinghua. Emotion management in modern people [D]. Hangzhou: Zhejiang University. 2008: 4.

[4] Tang Song Xiao. A Review of Research on Emotion Recognition[J]. Journal of Jilin Institute of Chemical Technology,2015,32(10):109-114.DOI:10.16039/j.cnki.cn22-1249.2015.10.031.

[5] Wan L.L.,Chen J. A study on the effect of negative emotions on the incidence of mental illness among college students[J]. Practical preventive medicine,2015,22(04):444-445.

[6] Lin Han,Zhang Kan. Applied and fundamental research on speech emotion recognition[J]. Human Ergonomics,2009,15(02):64-66+73.DOI:10.13837/j.issn.1006-8309.2009.02.012.

[7] Minsky M. The Society of Mind [M]. Simon &Schuster, New York, NY,1985.

[8] Zhao Li, Kobayashi, Y, Niimi, Y. Tone recognition of Chinese continuous speech using continuous HMMs[J]. Journal of the Acoustical Society of Japan, 1997.53(12):933-940

[9] Zhang Chunzhi. Social Hazards of Negative Emotions among College Students and Case Analysis[J]. Journal of Shenyang Agricultural University (Social Science Edition) 2006(02):301-303.

[10] Shao Jin, Fan FuMin. Current status and Prospect of domestic group counseling research in China from 1996 to 2013. Development trends Chinese Journal of Mental Health, 2015, 29(4): 258-263

[11] Gao, Hui, Su, Guangchuan, Chen, Shanguang. Research Progress of Emotional Speech Feature Analysis and Recognition[J]. Aerospace Medicine and Medical Engineering ,2004(05):386-390.DOI:10.16289/j.cnki.1002-0837.2004.05.018

[12] Fox S. Spector P E. A Model of Work Frustration –aggression[J]. Journal of Organizational Behavior 1999 20 (6):915-931

[13] Judge TA Scott B A I lies R. Hostility. Job Atti-tudes. and Workplace Deviance. Test of a Multi-level Model[J]. Journal of Applied Psychology. 2006 91 (1):126-138

[14] Ververidis D Kotropoulos C. Emotionals peechRec-ognition:resources features and methods [J]. Speech Communication 2006 48 (9):1162-1181

[15] Gao Hui,Chen Shanguang,An Ping,Wang Jun,Qin Haibo,Wang Bo. Study on speech recognition of a stressful emotion in simulated spaceflight environment[J]. Aerospace Medicine and Medical Engineering,2010,23(04):248-252.DOI:10.16289/j.cnki.1002-0837.2010.04.015.