Implementation of A New Speech Negative Emotion Recognition System

Yuqiong Wang¹*, Zehui Zhao², Zhiwei Huang³

¹Nari Group Corporation/State Grid Electric Power Research Institute, Nanjing 211106, China
²The University of Edinburgh, Old College, South Bridge, Edinburgh, Scotland, United Kingdom
³Xi'an Jiaotong University, No.28, Xianning West Road, Xi'an, China

*fsenous@163.com

Abstract. A new emotion recognition system based on speech is constructed to improve the ability of recognizing negative emotions. Multi-dimensional acoustic characteristics were tested and among them, short-term energy and Mel-frequency cepstral coefficients (MFCC) were selected to be used as parameters for recognition. The system consists of two modes: single recognition and group recognition. Single recognition adopts BP neural network model based on MFCC, while group recognition adds support vector machine model based on short-term energy on the basis of single recognition which the group recognition rate of 20 speech can reach 97%. With the increase of the number of speech in each group, the recognition accuracy of negative emotion tends to 100%.

Keywords: Emotion recognition, Characteristic parameters, Group recognition, Back Propagation Neural Network.

1. Introduction

Voice is the most basic medium of communication in people's daily life[1]. Emotional information is a kind of speech information. For the same sentence, people's happiness, anger and sadness can be expressed through different features such as intonation, speed, accent, etc. Researches show that strong emotional fluctuations can have a significant impact on people's daily lives[2]. For example, the mood of drivers is related to whether they can drive safely, the mood of patients is connected with whether they can face the illness positively and optimistically, the mood of astronauts is associated with whether they can complete a space mission efficiently.

In 1995, Professor Picard R of MIT Media Laboratory proposed the concept of "emotional computing"[3]. The objective characteristic parameters and subjective emotion are linked to construct the "emotion model". At present, the application of "emotion model" has been carried out in many aspects: facial micro expression, posture analysis, speech emotion recognition and so on constitute multimodal computing[4], so as to make human-computer interaction achieve personalized service. The
combination of emotion recognition and human-computer interaction can be applied to many aspects, such as lie detection of criminal cases, emergency monitoring of inpatients and drunk drivers[5]. Therefore, it is of great value to study the emotion of speech signal, especially the negative emotion.

Since the 21st century, Tawari Ashish and Trivedi Mohan scholars have specially designed a voice based auxiliary emotion classification framework for drivers to monitor drivers' emotional fluctuations in real time with an unobtrusive way[6]. The laboratory of Teheran University combines facial expression with speech, and simulates human cognition of emotion according to its correlation with emotion recognition. The experimental results show that the use of mixed features and decision level fusion can improve the results of unimodal system[7], which also reflects the trend of multimodal emotion recognition. Turkey Tokat University Computer Engineering Department has developed a new tool for speech feature extraction and classification[8], which reduces the number of features and increases the classification success rate. In conclusion, the optimization of speech emotion recognition technology has gradually become a key issue in the field of emotion recognition.

According to the research results in the past two years, the research based on acoustic features is still the mainstream trend. This paper uses the Chinese emotional corpus of the Institute of automation, Chinese Academy of Sciences, and selects the best Mel frequency cepstrum coefficient and short-term energy characteristics through repeated experiments, including the maximum, minimum, mean value, variance, jitter, linear regression coefficient, mean square error and 36 dimensions frequency cepstrum coefficient.

2. Method

The emotion recognition system can be used for single recognition and group recognition: single recognition is suitable for accurate detection of a single speech; group recognition is suitable for detecting all the voices spoken by the recognized person within a period of time, so as to identify the emotion of the identified person in the current situation, which eliminates certain contingency compared with single recognition.

2.1. single recognition

Single recognition uses MFCC as the recognition basis, processes the voices in CASIA emotional voice database of the Academy of Sciences, extracts MFCC parameters of each emotion and normalizes them with the maximum-minimum method to eliminate the magnitude difference between input and output data to minimize the recognition error. A 37th dimension marker bit is added to the end of each normalized frame to mark the mood represented by the frame data, and then it is input into the neural network for supervised training. To reduce the chance, the initial values of weights and thresholds of the neural network are obtained by generating random numbers, as well as the method of the network order of each voice input in the training database used. After training, the detected speech is recognized.

The initial weights and thresholds of neural network are random values, and the input sequence of training data is also random, so the final network parameters will have certain differences after each training. Moreover, the MFCC parameters of some emotions have little difference, and the output values of the same frame calculated according to different weights and thresholds may be different, resulting in recognition errors. Therefore, the recognition results obtained by building neural network only once are not very reliable.

The solution is to build a neural network for many times, and recognize the same detected speech in the network with different weights and thresholds. In the statistical results of each recognition, different scores are added to the corresponding emotions according to the proportion of frames. The emotion with the highest total score is considered as the final recognition result of the speech.

2.2. group recognition
Group recognition uses MFCC and short-term energy as the recognition basis. The two features of the detected speech are input into BP neural network and support vector machine respectively for recognition. Finally, the final recognition result is obtained by adding the two machine learning results according to a certain weight. The steps of training and recognizing MFCC features using BP neural network are exactly the same as single recognition. Here, we only explain the method of training and recognizing short-term energy features using support vector machine.

Processing the speech in CASIA emotional voice database of Chinese Academy of Sciences, extracting the short-term energy parameters of each emotion, are normalized by maximum minimum method. The same method is used to process a group of detected speech, and the short-term energy parameters are extracted and normalized. After that, a new column vector is created as a flag to represent the emotion corresponding to each parameter, so that the SVM can conduct supervised learning. Similarly, in order to reduce the chance of recognition results, we still choose the method of "building machine learning model for many times, random training data input sequence, getting scores on the basic of single recognition results, obtaining final results with statistical scores ". Finally, the emotion with the highest total score is considered as the result of this group of speech recognition.

BP neural network is used to train MFCC features, and support vector machine is used to train short-term energy features. In the recognition accuracy results of the two for different emotions, select several emotions which have a large ratio of neural network accuracy and support vector machine accuracy, and calculate the average value of the ratio of accuracy rate. Finally, each weighted coefficient is assigned according to the average value, and the final recognition result is obtained after the weighted combination.

3. Testing system

In order to ensure the accuracy of the test voice emotion, and to show that the system has universality for the detected voice, we choose another emotional voice database EMO-DB German emotional voice database to test, instead of using the Chinese Academy of Sciences emotional voice database.

3.1. Test of single recognition function

The EMO-DB German emotional voice database used in the test contains 50 speech sounds for each emotion. A total of 200 voices of four emotions are tested in turn. The single recognition accuracy of the system is shown in the following figure.
Fig.1 Single recognition accuracy of each emotion

The four columns of each emotion from left to right, represent the frequency of the first, second, third, and fourth ranking of the emotion scores in the test. It can be seen that the recognition accuracy of the system for anger and sadness is better, while the accuracy rate for happiness and peace is lower. Although the accuracy of peace is not high, but the ranking of scores is in line with the trend from high to low.

3.2. Test of group recognition function

In the group recognition test, 10 of the 50 speech in EMO-DB German emotional voice database were randomly selected as a group of inputs to the system for recognition. Each emotion is tested 50 times, and the group recognition accuracy is shown in the following figure.
Due to the short-term energy based support vector machine (SVM) has better recognition effect on anger and sadness than on happiness and peace, it will increase and decrease the recognition accuracy of corresponding emotions when it is combined with the recognition results of MFCC neural network in proportion. If the number of each group of speech is increased to 20, the result is shown in Fig. 3.

It can be seen that when the number of speech in each group increases, the recognition accuracy of each emotion will increase. This is because the more the number of data, the higher accuracy of data
normalization, the smaller the error in the recognition process. Another reason is that the proportion of some speech sounds which are not accurate in emotion is reduced, and the influence on the overall output is also smaller. In practical application, we can obtain enough number of speech by cutting speech appropriately to improve the overall recognition rate.

4. Conclusion

Through the test of single recognition and group recognition, we can see that the recognition accuracy of the system for anger and sadness is very high, which meets the purpose of the system construction: the single recognition rate is 72%, the group recognition rate of 10 speech is 89%, and the group recognition rate of 20 speech is 97%. Therefore, in the practical application, the system can effectively monitor the abnormality of the monitored object and avoid the occurrence of dangerous situation. In addition, two different databases are used in training and testing, which also proves that the system has enough universal applicability and can achieve good results in the actual use process.

References

[1] Rui S 2015 Chinese speech emotion recognition system based on acoustic feature analysis D. Beijing University of Posts and telecommunications
[2] Zijun H, Liang Z 2010 Overview of speech recognition technology and application J. Journal of Nanchang Normal University 31 (3) pp44-46
[3] Picard R W ,Papert S , Bender W and et al 2004 Affective Learning — A Manifesto J. BT Technology Journal 22(4) pp253-269
[4] Jianghe M, Ying S and Xueying Z 2019 Multimodal emotion recognition based on speech signal and EEG signal J. Journal of Xi'an University of Electronic Science and technology
[5] Zhentao L, Jianping X, Min W and et al 2018 An overview of speech emotion feature extraction and dimension reduction methods J. ActaSinicaSinica 41 (12) pp199-217
[6] Tawari A, Trivedi M 2010 Speech based emotion classification framework for driver assistance system C. Intelligent Vehicles Symposium (IEEE)
[7] Bejani M, DavoodGharavia 2014 Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks J. Neural Computing and Applications 24(2) pp399-412
[8] TurgutÖzseven, MuharremDüğenci 2018 SPeechACoustic (SPAC): A novel tool for speech feature extraction and classification J. Applied Acoustics 136 pp1-8