ISSN 2616-5775 Vol. 4, Issue 3: 60-64, DOI: 10.25236/AJCIS.2021.040309

The Emotion Recognition System Based on Support Vector Machines

Yuwei Wei[1, *], Linghao Kong[2], Xinhang Li[3], Mingxuan Pan[4], Chengke Tang[5], Shicong Sun[6]

[1]Xidian University, Xi'an, Shaanxi, China
[2]Southwest Jiaotong University, Chengdu, Sichuan, China
[3]Shanghai University of Engineering Science, Shanghai, China
[4]The Chinese University of Hong Kong, Shenzhen, Shenzhen, Guangdong, China
[5]University of Nottingham Ningbo China, Ningbo, Zhejiang, China
[6]Southwest University, Chongqing, China
**Corresponding author: 444564940@qq.com*
**These authors contributed equally to this work*

Abstract: *Now speech recognition plays an important role in the filed of human-computer interaction. As the main way to communicate with each other in daily life, voice carries a wealth of emotional information. It is necessary for artificial intelligence to process these information to reach better interaction. In this article, we aim to study a speech emotion recognition system based on support vector machines(SVM), which can recognize different emotions through people's voice to help people manage their emotions by analysing changes of their emotions. To achieve this goal, we have built a small Chinese voice database including four emotions, each from six different people. We wrote the Matlab program to complete speech feature parameters extraction, model training and emotion recognition, thus realizing the emotion classification of the speech signal.*

Keywords: *emotion recognition, support vector machines(SVM), extract characteristic parameters*

1. Introduction

As one of the most rapidly developing core technologies in the field of information science at present, sound signal acquisition is an emerging subject that studies the processing of sound signals by using phonetics technology and digital signal processing.

The development of speech recognition technology dates back to 1952, when researchers at Bell LABS built a system called Audrey for single-person numerical speech recognition. But at the time it was more of a conceptual thing, and speech recognition was inefficient. While today, with the development of science and technology, artificial intelligent furniture with speech recognition function appears in the market. Also with the development of this kind of communication technologies such as sound acquisition and processing, more and more categories of smart home appear in people's lives, and people are willing to do some things with emerging furniture instead of labour. Therefore, under the help of professional technology and equipment, people's voice collection, identification and analysis came into being.

Now, people are not satisfied that machines can only recognize what they say, they want more and better human-computer interaction, they want machines can "feel" their emotions. Of course, machines do not have emotions, but we can analyze a speaker's emotions by speed, stress, intonation, etc., so machines can "sense" human emotions.

Actually, related research has already been launched by many researchers and many models were raised.

Modern general speech recognition system is based on hidden Markov model, which is a statistical model to output a series of symbols or quantitative sequences. The hidden Markov model is used in speech recognition because speech signals can be regarded as segmented stationary signals or short-time stationary signals. Over short time scales (such as 10 milliseconds), speech can be approximated as a stationary process. Speech can be thought of as a random Markov model. Professor Zhang Jing, Yang Jian and Su Peng have done research in this field and described in detail the research results of syllable

ISSN 2616-5775 Vol. 4, Issue 3: 60-64, DOI: 10.25236/AJCIS.2021.040309

recognition combining hidden Markov model with other algorithms. [1]

Another method of speech recognition is based on Dynamic Time Warping (DTW). DTW has historically been used for speech recognition, but it has been largely replaced by more successful methods based on hidden Markov models. DTW is an algorithm used to measure the similarity between two sequences that may vary in time or speed, but it needs large calculation and storage capacity. Wen Han, Huang Guoshun did some research in this area and proposed a strategy to improve the calculation of cumulative distance and ease the limit of endpoint alignment. [2]

Since 2014, there has been a great deal of research interest in "end-to-end" ASR. Traditional speech-based approaches (i.e., HMM based models) require separate components and training for speech, acoustic, and language models. The end-to-end model works together to learn all the components of a speech recognizer. This is important because it simplifies the training process and the deployment process. [3]

2. Design considerations

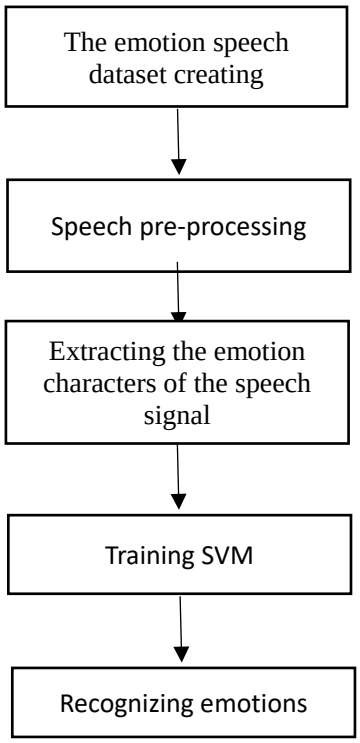

Figure 1: Flow chart of the emotional recognition system

Our team did some experiments concerning speech recognition based on MATLAB and we want to design a device applied to the concept of emotional management, realize people's individual voiceprint entry and according to the speed of sound tone to identify family members then to judge the speaker's emotional state (peace, happiness, anger and fear, etc.). Ultimately the device will conclude the monthly times of the mood changes, then give hints or help individuals to have a emotion management, hoping to provide a certain help to the core technology of signal processing. We used a method of support vector machine (SVM) for the purpose. The process as Figure 1 shows.

2.1 The emotion speech dataset creating

The first step is creating a dataset of emotion speech. In this paper, the data are from my team members' speech recordings, and also providing four different emotion, respective speaking with placid, happy, sad and anger, and finally the dataset totally has 28 speech samples.

There are two very significant aspects that may influence emotion recognition result and cause some deviations, so we need to intentionally avoid unnecessary deviation that created by human, the first aspect about the speech transcript, which sentence we choose must without any emotion tendency. The second aspect is we need to pay attention on sentence component such as consonant, auxiliary word and the

ISSN 2616-5775 Vol. 4, Issue 3: 60-64, DOI: 10.25236/AJCIS.2021.040309

difference of speakers' gender.

2.2 Speech pre-processing

In this paper, we use FIR filter for the speech signal pre-emphasis, in this way low frequency interference at 50Hz or 60Hz will be filtered out, the purpose of pre-emphasis is to lift the high-frequency portion of the signal to flatten the spectrum, maintain in the entire frequency band from low to high, and being able to use the same SNR for spectral analysis or channel parameter analysis, increasing the high frequency resolution of speech.

Also processing speech signal with windowing, reducing the truncation effect of voice frames, facilitating application of time processing technology.

2.3 Extracting the emotion characters of the speech signal

Analyzing several important features: speech endpoint, maximum value of fundamental period, maximum value of vibration peak, average short-time zero crossing rate and average short-time energy.

In the study of speech emotion recognition, extracting speech feature parameter is one of the most important steps in whole study. In this paper, we choose several parameters that have the highest weight for emotional feature recognition are the dominant features.

2.3.1 Resonance

Resonance peaks are areas of relatively concentrated energy in the sound spectrum domain, mainly reflecting information about changes in the sound channel. The resonance peak is a sonic feature, with the emotion changing major in the first, second, and third resonance peaks.

2.3.2 Estimating Pitch Period

Generally, we have several methods to do pitch detection, such as: Autocorrelation Function method, peak extraction algorithm, Average Difference Function method, Parallel processing technology etc.

In this paper, using method of Autocorrelation Function to detect pitch period. The speech signal s(m) is intercepted by a window of length N as a section of the windowed speech signal S_n (m), after which the autocorrelation function (ACF) R_n (k) of S_n (m) is defined (i.e., the short-time autocorrelation function of the speech signal s(m)) as follows:

$$R_n(k) = \sum_{m=0}^{N-k-1} S_n(m) S_n(m+k) \tag{1}$$

The range that R_n(k) not equal to zero is k = (-N+1) to (N-1), and it is an even function. The autocorrelation function of the turbid tone signal has a peak in the position of an integer multiple of the fundamental period, while the autocorrelation function of the clear tone has no obvious peak. Therefore, the detection of whether the peak can be judged to be a clear or turbid tone, the detection of the location of the peak can be extracted from the value of the fundamental tone period.

2.3.3 Endpoint Detection

Speech endpoint detection essentially distinguishes between speech and noise based on the different characteristics exhibited by the same parameters. The traditional speech endpoint detection algorithm combines short-time energy and over-zero rate to detect unvoiced sound and short-time energy to detect voiced sound, which together enable endpoint detection in the case of large signal-to-noise ratio. In this paper, we use bipartite limiting method. In this method short-time energy detection distinguishes between voiced and muted sounds. But for unvoiced sounds, due to their energy is relatively low, they may be misjudged as silence in short-term energy testing because they are below the energy threshold. Also, the short over-zero rate can be distinguished from silence and unvoiced sounds in speech. So, we combine the two methods, detecting speech segments (unvoiced and voiced sounds) and silence segments.

2.3.4 Mel Frequency Cepstrum Coefficient

Mel Frequency Cepstrum Coefficient (MFCC) parameters are based on the hearing characteristics of the human ear, and finally convert the frequency spectrum into coefficients in the cepstrum domain. It effectively combines the human ear's auditory perception characteristics with the voice signal generation mechanism, has good recognition performance and anti-noise ability. So it is widely used in speech

ISSN 2616-5775 Vol. 4, Issue 3: 60-64, DOI: 10.25236/AJCIS.2021.040309

recognition.

The MFCC parameter is to first convert the ordinary frequency into the Mel frequency in the frequency domain, and then transform to the cepstrum domain, and obtain the cepstrum coefficient through calculation.

2.4 Support Vector Machines

This paper chooses support vector machine (SVM) algorithm to realize speech emotion recognition. SVM can be applied to linear or nonlinear sample classification problems. To solve the non-linear sample classification problem, the core is to establish an optimal classification hyperplane, and map the linearly inseparable data samples to the high dimensional feature space through the kernel function to achieve linear separability. SVM has obvious advantages in dealing with small samples and non-linear pattern classification problems.

The core of SVM speech emotion recognition is the determination of the kernel function. The radial basis kernel function (RBF) is widely used and can accurately describe the distribution structure of the data. Therefore, this paper chooses the radial basis function as the SVM kernel function, as shown in formula (2).

$$K\left(x_i, x_j\right) = \exp\left(-\frac{1}{2\sigma^2}\left\|x_i - x_j\right\|\right)^2 \tag{2}$$

σ is the width parameter of the function. The parameter has significant effect on the performance of the SVM generalization. Through experiments, the optimal values of the kernel parameters and penalty factors can be determined.

Emotion recognition based on SVM is realized by training modules and testing modules. Model training mainly uses the extracted global statistical emotional features to train an SVM model with emotional classification capabilities. In this paper, 12 emotional materials are selected as training materials, and a total of 6 classification models are established. Through one-to-one comparison of different emotions in the training sample set, different sub-SVM models are built respectively. Finally, input the test sample data into the SVM model to output the recognition result.

3. Conclusion

After the project completed the preprocessing of the emotional speech data set, the emotional feature parameters of the speech data in the data set were extracted by estimating the interval period, endpoint detection, and extracting the Mel frequency cepstrum coefficients, and selected support vector machines. The algorithm uses the principle of constructing an optimal classification hyperplane to achieve linear separation. After the machine learning training is carried out through this algorithm, the actual emotion test is carried out on the speech data. The test results show that the machine can better realize the correct recognition of four different emotions.

As an indispensable part of human-computer interaction, speech emotion recognition has become a current research hotspot in the field of artificial intelligence and human-computer interaction. It has a wide range of applications and extremely high application value. By judging the emotional state of the speaker, this project can count the emotional changes over a period of time, thereby guiding people to control their emotions and have the effect of self-cultivation. In addition, voice emotion recognition has also been applied in many other fields. For example, when the car is driving for a long time, the driver’s speech emotion can be used to judge whether the driver has entered fatigue driving, so as to remind the driver to reduce the possibility of traffic accidents. When the doctor is performing an operation, the emotion recognition device can be used to monitor the doctor's mental condition in real time, so as to avoid accidents caused by the doctor's mood fluctuations. These are the different degrees of application of speech emotion recognition in different fields.

4. Related works

In this part, we will talk about the related work done by other researchers.

In our research, we used SVM (support vector machine) to do sentiment classification and try to

ISSN 2616-5775 Vol. 4, Issue 3: 60-64, DOI: 10.25236/AJCIS.2021.040309

improve the correctness by modifying the algorithm. We did sentiment classification with peoples' voice and wanted to divide them into 4 categories, happy, sad, angry and harmony. In our research, people spoke in Chinese and they spoke when they had different moods. There was also a study did emotion recognition in Chinese speech. They were trying to extract 5 features from the sample (2017): Mel Frequency Cepstrum Coefficient (MFCC), pitch, formant, short-term zero-crossing rate and short-term energy. In order to extract those 5 features, they chose Deep Belief Network (DBN) combined with SVM to finish this research. To speed up, they used conjugate gradient method to train the model. Their new method achieved the accuracy of 95.8%, which is higher than either DBN or SVM. Furthermore, the dataset can be diverse. Researchers also did sentiment analysis on other social media. There are still a lot of other studies about sentiment classification with well-known techniques. Zhang et al (2019) combined SVM and LSA (latent sentiment analysis) to classify online views into 4 categories - happiness, hope, disgust, and anxiety. [4] The accuracy and the efficiency for their model to do sentiment classification improved a lot (Zhang et al., 2019). In the research conducted by Zainuddin et al. (2018), they did aspect-based sentiment analysis on social media -Twitter. [5] In order to improve the accuracy, they embedded feature selection method into their model. They compared the accuracy when using principal component analysis (PCA), latent semantic analysis (LSA), or random projection (RP) feature selection. The result is good. The method improves the accuracy by more than 70%. Further research can be done with combining different kinds of methods and using different ways to train the model to increase the efficiency and accuracy for sentiment classification.

References

[1] Zhang Jing, Yang Jian, Su Peng (2020), "A review of monosyllabic recognition in Speech recognition", Computer Science, S2, 2020, P172-174+203.

[2] Wen Han, Huang Guoshun (2010), "Research on improvement of DTW algorithm in Speech Recognition", Microcomputer Information, 19, 2010, P195-197.

[3] Li Minghao (2018), "Research on continuous Speech Recognition based on deep Neural network", Jilin University, 2018.

[4] Zhang, W., Kong, Sx., Zhu, Yc. et al. Sentiment classification and computing for online reviews by a hybrid SVM and LSA based approach. Cluster Comput 22, 12619–12632 (2019).

[5] Zainuddin, N., Selamat, A., & Ibrahim, R. (2018). Hybrid sentiment classification on twitter aspect-based sentiment analysis. Applied Intelligence, 48(5), 1218-1232.

[6] Zhu, Lianzhang, Chen, Leiming, Zhao, Dehai, Zhou, Jiehan, & Zhang, Weishan. (2017). Emotion Recognition from Chinese Speech for Smart Affective Services Using a Combination of SVM and DBN. Sensors (Basel, Switzerland), 17(7), 1694.