International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622
International Conference on Industrial Automation and Computing (ICIAC- 12-13th April 2014)

RESEARCH ARTICLE OPEN ACCESS

Emotion recognition using Speech Processing Using k-nearest neighbor algorithm

Anuja Bombatkar, Gayatri Bhoyar, Khushbu Morjani, Shalaka Gautam,Vikas Gupta
Department of Electronics and Telecommunication Engineering,
Yeshwantrao Chavan college of Engineering, Nagpur

Abstract—
Speech Emotion Recognition (SER) is a current research topic in the field of Human Computer Interaction (HCI) with wide range of applications. The purpose of speech emotion recognition system is to automatically classify speaker's utterances into four emotional states such as anger, sadness, neutral, and happiness. The speech samples are from Berlin emotional database and the features extracted from these utterances are energy, pitch, ZCC, entropy, Mel Frequency cepstrum coefficients (MFCC). The K Nearest Neighbor (KNN) is used as a classifier to classify different emotional states. The system gives 86.02% classification accuracy for using energy, entropy, MFCC, ZCC, pitch Features.
Keywords- Speech Emotion; Automatic Emotion Recognition; KNN; Energy; Pitch; MFCC; ZCC.

I. INTRODUCTION

Speech emotion recognition aims to automatically identify the current emotional state of a human being from his or her voice. It is based on in-depth analysis of the generation mechanism of speech signal, extracting some characteristic parameters which contain emotional information from the speaker's voice, using these parameters and taking appropriate pattern recognition methods to identify emotional states from speech.

Now, Automatic Speech Emotion Recognition is a very active research topic in the Human Computer Interaction (HCI) field and has a wide range of applications. For distance learning, Indentifying students' emotion timely and making appropriate treatment can enhance the quality of teaching. In automatic remote call center, it is used to timely detect customers' dissatisfaction. It is also used to aid clinical diagnosis or to play video games. The research of automatic speech emotion recognition, not only can promote the further development of computer technology, it will also greatly enhance the efficiency of people's work and study, and help people solve their problems more efficiently. It will also further enrich our lives and improve the quality of life.

In recent years, a great deal of research has been done to recognize human emotion using speech information. Many Speech databases were built for speech emotion research, such as BDES (Berlin Database of Emotional Speech) that is German Corpus and established by Department of acoustic technology of Berlin Technical University [1], There are also some Mandarin Affective Speech Databases, Many researchers have proposed important speech features which contain emotion information, such as energy, pitch frequency formant frequency [3], Mel-Frequency Cepstrum Coefficients (MFCC) and its first derivative [4].

Furthermore, many researchers explored several Classification methods, such as Neural Networks (NN) [5], Gaussian Mixture Model (GMM), Hidden Markov model (HMM), Maximum Likelihood Bayesian classifier (MLC), Kernel Regression and K-nearest Neighbors (KNN) and Support vector machines (SVM).

In this paper, we use Berlin as well as Hindi Emotional database to train and test our automatic speech emotion recognition system. Prosody and Spectral features have been widely used in speech emotion recognition.

The paper is organized as follows. Section II describes the database used in the experiments. Section III introduces the automatic speech emotion recognition system. The speech features are presented in this section. Section IV introduces the Support Vector Machine algorithm. Experiments to assess the proposed system are performed in section V. Section VI concludes this paper.

II. SPEECH DATABASE

A. Berlin Database of Emotional Speech

One of the database used in this paper is Berlin emotional speech database, which is a simulated speech database. It is an open source speech database and easy to access, and it is frequently used in fields of speech emotion recognition. This database contains four basic

emotions: anger, happiness, sadness and neutral. All of the speech samples are simulated by ten professional native German actors (5 actors and 5 actresses). There are totally about 500 speech samples in this database, in which 286 speech samples are of female voice and 207 samples are of male voice. The length of the speech samples varies from 2 Seconds to 8 seconds. These samples are divided into about 1100 segments of 2 seconds in our analyzing [1].

B. *Hindi Database of Emotional Speech*

The database used in this paper is Hindi emotional Speech database, which is a simulated speech database. This database contains four basic emotions: anger, happiness, sadness and neutral. All of the speech samples are simulated by ten professional native Hindi actors. There are totally about 100 speech samples. The length of speech samples varies from 2 Seconds to 5 Seconds.

III. FEATURE EXTRACTION

The Speech signal contains a large number of parameters that reflect the emotional characteristics, and the different parameters result in changes in emotion. Thus, the most important step in speech emotion recognition is how to extract the feature parameters, which can express mostly the emotion of speech.

In recent research of speech emotion recognition, some common features are speech rate, energy, pitch, formant, and some spectrum features, Mel-Frequency Cepstrum Coefficients (MFCC) and its first derivative and so on. There have been a handful of research on these features in the past years, and these features have been used in many speech emotion recognition research. In the use of these features parameters while some other features of a good discrimination have been proposed: Mel Energy spectrum Dynamic coefficients (MEDC).

In our research, we calculate the statistics of energy, pitch, formant frequency, ZCC, MFCC and make use of them to classify the speech emotion. In the research of speech emotion recognition, seven basic emotions: anger, happy, sadness, fear, boredom, disgust, and neutral. When people are in different emotional state, their speeches have different changes in speak rate, pitch, energy, ZCC, and spectrum.

Usually, anger has a highest mean value and variance of pitch, and mean value of energy. Disgust and Boredom have a low mean value of pitch and energy. Happy has an improvement of mean value, variation range and variance of pitch, and the mean value of energy. On the contrary, the mean value, variation range and variance of pitch of sadness is decrease, the energy is weak, the speak rate is slow and the decrease of the spectrum in high frequency components. As a result, we can extract the statistics of pitch, energy and some spectrum features of speech to recognize the emotion in speech.

A. Energy and related features

The Energy is an important feature of speech, and the analysis of energy is focused on short-term energy and short-term average amplitude. In order to obtain the statistics of energy feature, we use short-term function to extract the value of energy in each speech frame. Then we can obtain the statistics of energy in the whole speech sample by calculating the energy, such as mean value, max, value, variance, variation range, contour of energy.

B. Pitch

The pitch signal is another important feature in speech emotion recognition. The vibration rate of vocal is called the fundamental frequency FO or pitch frequency [10]. The pitch signal is also called the glottal wave-form; it has information about emotion, because it depends on the tension of the vocal folds and the sub glottal air pressure, so the mean value of pitch, variance, variation range and the contour is different in seven basic emotional statuses. The method widely used to extract the pitch is based on the short-term autocorrelation function. We calculate the value of pitch frequency in each speech frame.

C. Mel-Frequency Cepstrum Coefficients (MFCC)

There are many researches on MFCC feature parameters, it is widely used in speech recognition and speech emotion recognition studies, and it obtained a good recognition rate. MFCC is based on the characteristics of the human ear's hearing, which uses a nonlinear frequency unit to simulate the human auditory system. Mel frequency scale is the most widely used feature of the speech, with a simple calculation, good ability of the distinction, anti-noise and other advantages [11]. Usually the process of calculating MFCC is shown in Figure.

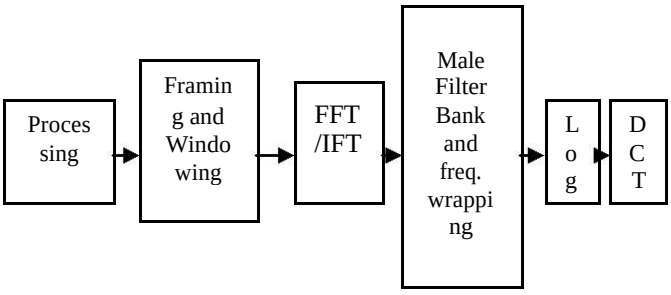

Fig. Speech Emotion Recognition System

MFCC in the low frequency region has a good frequency resolution, and the robustness to noise is also very good, but the high frequency coefficient of accuracy is not satisfactory. So we give up the high-level order of the MFCC and use

only low-level order as audio feature parameters. In our research, we extract the first 13-order of the MFCC coefficients. For each order coefficients, we compute all the frames of the mean, variance, maximum and minimum in entire speech. Each MFCC feature vector is 52-dimensional. Mean, Variance, Maximum, Minimum of each coefficient across all the frames.

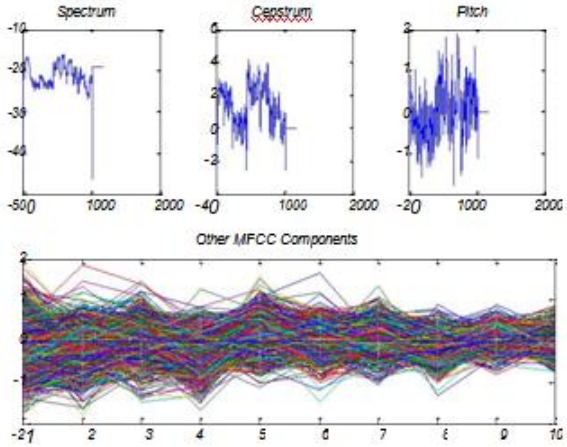

Fig.2 Plot of various features extracted.

IV. K-NN CLASSIFATION ALGORITHM

In recent years in speech emotion recognition, researchers proposed many classification algorithms, such as Neural Networks (NN), Gaussian Mixture Model (GMM),Hidden Markov Model (HMM), Maximum Likelihood Bayesian classifier (MLC), Kernel Regression and K-nearest Neighbors (KNN) [1] and Support vector machines (SVM).

In pattern recognition, the ***k*-Nearest Neighbors algorithm**

(or ***k*-NN** for short) is a non-parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the *future space*. The output depends on whether k-NN is used for classification or regression.

In the classification phase, *k* is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the *k* training samples nearest to that query point. A commonly used distance metric for continuous variables is Euclidean distance. In our research, we use the system to extract the speech's feature. After the feature extraction, we give each speech sample with the corresponding emotion class label. After that we input them to the LIB SVM classifier and gain a model file by training the data set. When an unclassified speech sample come into this system, the system extract the feature coefficients and use the model file to classify the speech emotion.

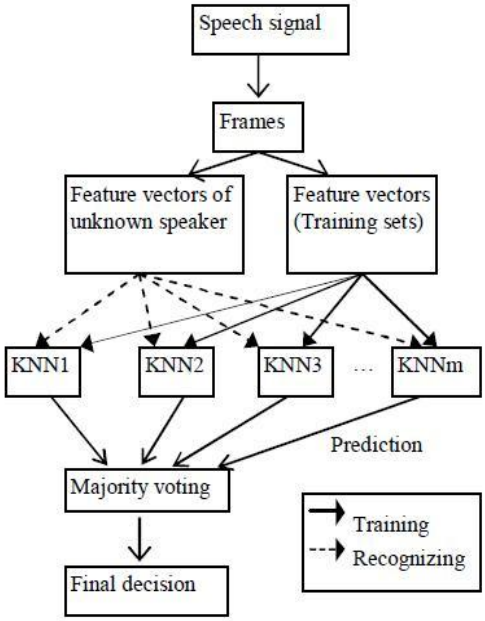

Fig.4.1 KNN Structure and Algorithm

The diagram of the speaker identification system is shown in Fig4.1.After pre-processing, we segment the raw signal into a sequence of 20ms frames. Then extract 16 dimensions Mel Frequency Cepstrum Coefficient (MFCC) of each frame as feature vectors. Ensemble learning algorithms are adopted in the system. During the training phase, component learners (KNN) are trained from different samples, which are generated from the original data set via different algorithms, such as bagging and proposed Bag WithProb. In the identification phase, features will be recognized by all component learners. Then, predictions are combined via majority voting to make the final decision.

V. OUTPUT

Sr No.	Emotion	Accuracy (Hindi)	Accuracy (Berlin)
1.	Angry	90%	90%
2.	Happy	80%	90%
3.	Neutral	70%	80%
4.	Sad	75%	80%

For Hindi database the accuracy level of Anger and Happy emotion is good since there is large difference in features obtained for them than the other ones.

The neutral and the sad emotion have many overlapping feature. Their classification mainly occurs on pitch frequency other features are almost same. Hence accuracy level of classifying Sad and Neutral emotion is quite low.

For Berlin database the accuracy level of Anger and Happy emotion is good also for sad and neutral since German ascent is different.

VI. CONCLUSION

In this paper, we presented a novel method speech recognition, KNN algorithm It has many advantages over other conversational methods such as implicity and good generalization ability. At the same time, the generalization ability of an ensemble could be significantly better than that of LPCC method

In order to accelerate the application of speaker identification system, it will be encouraging for us to explore more effective algorithm. Future work will be done to test the algorithm under different signal to noise ratios (SNR) with a greater population of speakers and

improve the accuracy in different environment we can use the Visual features.

REFRANCES

[1] Yan Zhang, Ensemble Learning and Optimizing KNN Method for Speaker Recognition
[2] Lawrence R. Rabiner. Ronald W. Schafer. Introduction to Digital Speech Processing. Foundations and Trends Signal Processing 1:1- (2007).
[3] Moataz El Ayadi , Mohamed S. Kamel, Fakhri Karray, Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 2010.
[4] Database of German Emotional Speech, http://pascal.kgw.tu-berlin.de/emodb/