IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p- ISSN: 2278-8735.Volume 10, Issue 2, Ver.1 (Mar - Apr.2015), PP 55-61
www.iosrjournals.org

Analysis of Emotion Recognition System through Speech Signal Using KNN & GMM Classifier

[1]Chandra Prakash, [2]Prof. V.B Gaikwad, [3]Dr. Ravish R. Singh [4]Dr. Om Prakash
Shri L.R Tiwari College Of Engineering, Mira Road, Mumbai Mumbai University

Abstract: *In machine interaction with human being is yet challenging task that machine should be able to identify and react to human non-verbal communication such as emotions which makes the human computer interaction become more natural. In present research area automatic emotion recognition using speech is an essential task which paid close attention. Speech signal is a rich source of information and it is an attractive and efficient medium due to its numerous features of expressing approach & extracting emotions through speech is possible. In this paper emotions is recognized through speech using spectral features such as Mel frequency cepstrum coefficient prosodic features like pitch , energy and were utilized & study is carried out using K- Nearest Neighbor classifiers and Gaussian mixture model classifier which is used for detection of six basic emotional states of speaker's such as anger ,happiness , sadness , fear , disgust and neutral using Berlin emotional speech database.*

Keywords: *Classifier , Emotion recognition, features generation , spectral features, prosodic features***.**

I. Introduction

When In past decade, the researchers highly attracted towards emotion recognition using speech signal in the field of speech signal processing, pattern recognition. There is an enormously important role of emotions in human life. As per human's perspective or feelings emotions are essential medium of expressing his or her psychological state to others. Humans have the natural ability to recognize the emotions of their communication partner by using all their available senses. They hear the sound, they read lips, they interpret gestures and facial expression Humans has normal ability to recognize an emotion through spoken words but since machine does not have capability to analyze emotions from speech signal for machine emotion recognition using speech signal is very difficult task. Automatic emotion recognition paid close attention in identifying emotional state of speaker from voice signal [1]. An emotions plays a key role for better decision making and there is a desirable requirement for intelligent machine human interfaces [1][3]. Speech emotion Recognition is a complicated and complex task because for a given speech sample there are number of tentative answer found as recognized emotion .The vocal emotions may be acted or elicited from "real", life situation [2].The identification and detection of the human emotional state through his or her voice signal or extracted feature from speech signal means emotion recognition through speech. it is principally useful for applications which require natural machine human interaction such as E-tutoring , electronic machine pet , storytelling, intelligent sensing toys , also in the car board system application where the detected emotion of users which makes it more practical [1].

Emotion recognition from speech signal is Useful for enhancing the naturalness in speech based human machine interaction To improve machine human interface automatic emotion recognition through speech provides some other applications such as speech emotion recognition system used in aircraft cockpits to provide analysis of Psychological state of pilot to avoid accidents. speech emotion recognition systems also utilizes to recognize stress in speech for better performance lie detection , in Call center conversation to analyze behavioral study of the customers which helps to improve quality of service of a call attendant also in medical field for Psychiatric diagnosis, emotion analysis conversation between criminals would help crime investigation department. if machine will able to understand humans like emotions conversation with robotic toys would be more realistic and enjoyable, Interactive movie, remote teach school would be more practical [2][3].

There are various difficulties occurs in emotion recognition from the speaker's voice due to certain reasons such as, existence of the differ in speaking styles, speakers, sentences ,languages, speaking rates introduces accosting variability affected different voice features this a particular features of speech are not capable to distinguish between various emotions also each emotion may correspond to the different portions of the spoken utterance. The same utterance may show different emotions & hence recognized emotional states which are not clear [4]. To recognizing emotional state of human being from speakers voice or speech signals

several system are proposed in last several years in the field of emotion recognition there are a variety of intellectual systems researchers have been developed using some universal emotions which includes anger, happiness, sadness, surprise, neutral, disgust, fearful, stressed etc. This different system also differs by different features extracted and classifiers used for classification. There are different features utilizes for recognizing emotion from speech signal such as spectral features and Prosodic features can be used. Because both of these features contain large amount of emotional information. Some of the spectral features are Mel-frequency cepstrum coefficients (MFCC) and Linear predictive cepstrum coefficients (LPCC). Some prosodic features formants , Fundamental frequency, loudness , Pitch ,energy and speech intensity and glottal parameters are the prosodic features also for detecting emotions through speech some of the semantic labels, linguistic and phonetic features also used[3][5].

To made human machine interaction becomes more powerful there are various types of classifiers which are used for emotion recognition such as Gaussian Mixtures Model (GMM) ,k-nearest neighbors (KNN), Hidden Markov Model (HMM), Artificial Neural Network (ANN) , GMM super vector based SVM classifier ,and Support Vector Machine (SVM). A. Bombatkar, et.al studied K Nearest Neighbour classifier which give recognition performance for emotions upto 86.02% classification accuracy for using energy, entropy, MFCC, ZCC, pitch Features. Xianglin et al**.** has been performed emotion classification using GMM and obtained the recognition rate of 79% for best features . Also emotion recognition in speaker independent recognition system typical performance obtained of 75%, and that of 89.12% for speaker dependent recognition using GMM if this study was limited only on pitch and MFCC features. M. Khan et.al. performed emotion classification using K-NN classifier average accuracy 91.71% forward feature selection while SVM classifier has accuracy of 76.57%.Table 3 and 4 show SVM classification for neutral and fear emotion are much better than K-NN [1] [2]-[4] - [6]-[7].

In this paper , K nearest Neighbor classifier and Gaussian mixture model (GMM) are two different classifiers are utilized for classification of the basic six emotional states such as anger, happiness , sad, fear, disgust and neutral state and no distinct emotion is observed. The pitch features , energy related features, formants, intensity, speaker rate are some prosodic feature and Mel-frequency cepstrum coefficients (MFCC),fundamental frequency are some spectral features which were used for the emotion recognition system. The classification rates of both of these classifiers were observed [2][3].

II. Emotional Speech Database

The efficiency of recognition is highly depends upon the naturalness of database used in the speech emotion recognition system. The collection of suitable database is most key task concerning to an emotion recognition system using speech signal. An important parameter is to consider for detecting emotions is the degree of naturalness of the database used to evaluate the performance of emotional speech recognizer. If a quality of database used is poor then inaccurate recognition occurs. Besides, for better classification the design of the database is significantly important that need to consider. Speech samples if collected from real life situations then it will be more realistic & natural. emotional speech samples are difficult to collect due to some natural parameters such as noise included at the times recordings. On the basis of different emotional states of human being and there differing cultural and linguistic environment different databases are implied by different researchers. R. Cowie et.al constructed their own English language emotional speech database for 6 emotional states such as anger, happiness, disgust, neutral, fear, sadness etc. In One of the Research M. Liberman, et al utilizes the database consists of 9 hours of speech data. It contains speech in 15 emotional categories, such as hot anger, cold anger, panic-anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, disgust and contempt , Constructed at the University of Pennsylvania. Most of the researcher used Berlin emotional speech database is a simulated speech database contains is totally about 500 acted emotional speech samples. Which are simulated by professional actors for emotion recognition through speech [7] [8]. In this study we utilizes the same database in which each speech sample corresponds to one emotion and by using this database the classification based on KNN & GMM is carried out as

III. Automatic Emotion Recognition Using Speech System

Automatic Emotion recognition system through speech is similar to the typical pattern recognition system. In this study, the block diagram of the emotion recognition system using speech considered illustrated in Figure 1 Emotion recognition system through speech is similar to the typical pattern recognition system. This sequence is also called the pattern recognition cycle, it implies various stages will present in the speech emotion recognition system through speech. In assessment of Emotion recognition system using voice signal there is vital role of speech database used. Proposed system is based on some prosodic features and some spectral features of voice signal. It consists of the emotional state speech sample preprocessing , train & test

sets of inputs, generation & selection of features , classification of Emotional state using different classifiers such as K Nearest Neighbor and Gaussian Mixture Model and recognition of emotion al state as the output. The emotional speech input to the system may contains the collection of the real world speech data as well as the acted speech data. After collection of the database containing short Utterances of emotional speech sample which was considered as the training samples and testing samples, appropriate and essential features such as spectral features and prosodic features were generated from the speech signal. These feature values were provided to the K Nearest Neighbor and Gaussian mixture Model for training of the classifiers then recorded emotional speech samples presented to the classifier as a test input. Then classifier classifies the test sample into one of the emotion & gives output as recognized emotion from the above mentioned six emotional state [2]-[6][7].

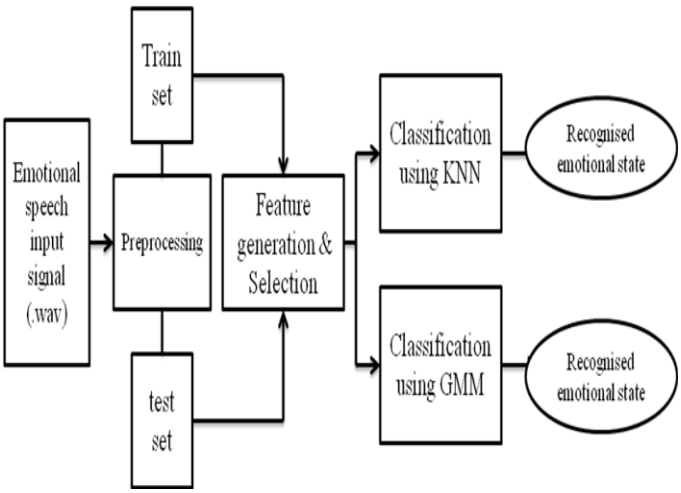

Fig 1. Emotion recognition system using speech

IV. Generation and Selection of Features from Speech Signal

In emotion recognition System using speech signal has significant features having a large amount of emotional information about speech signal which is most essential medium for recognition .Any emotion from the spoken voice is represented by various types of parameters contained in the speech and change in these parameters will provides outcome as corresponding variation in emotions. Therefore in speech emotion recognition system extraction of proper features of speech signal which represents emotions is an essential factor [2] [10].

There are different categories of speech features in which two main categories that are long term and short term features. Speech features are an effective constraint of speech signal To perform analysis of parameter of speech signal an important aspect of the feature extraction must be taken into consideration. Several researches have shown various features are use to parameterized speech signal in that prosodic features such as formant frequency, speech energy, speech rate ,fundamental frequency & spectral features such as Mel frequency cepstrum coefficients (MFCC) which are primary indicator of the speakers emotional states & provide potentially high efficiency to identify particular emotional states of speakers .Speech features fundamentally extracted from excitation source, vocal tract or prosodic points of view to perform different speech tasks. Speech signal is beaked into pieces of small intervals of 20 ms or 30 ms respectively for feature extraction through that signal that smaller partitioning of signal called frames. [6]In this work generation of some prosodic and spectral feature from speech signal has been done for emotion recognition. An important feature for identifying emotional state is pitch features which conveys considerable information about emotional status of the speaker from his speech.The vibration of the vocal folds, tension of the vocal folds and the sub glottal air pressure while speaking by speaker produced the pitch signal. Vibration rate of vocal cords is also called as fundamental frequency. The pitch signal is also called the glottal wave-form. We considered these fundamental frequencies as feature for emotion recognition to extract the pitch feature commonly used method is based on the short-term autocorrelation function [3] [4].

Another important feature is energy of speech signal. Speech energy is having more information about emotion in speech. The speech signal energy provides a representation in terms of amplitude variations. The analysis of energy is paying attention on short-term average amplitude and short-term energy. In this short time energy features estimated energy of emotional state by using variation in the energy of speech signal. To

obtain the statistics of energy feature we implied short-term function to extract the value of energy in each speech frame [6].

Mel-Frequency Cepstrum coefficients is the most important feature of speech it widely used spectral feature for speech recognition and speech emotion recognition which provides ease of calculation , reduction in noise , having better capability to distinguish. MFCC having high recognition rate and having low frequency region has a good frequency resolution. MFCC is based on the characteristics of the human ear's hearing & perception, which uses a nonlinear frequency unit to simulate the human auditory system [6]. Mel frequency scale is the most widely used feature of the speech, Mel-frequency cepstrum feature provide improved rate of recognition. The cepstral analysis in the speech processing applied to extract vocal tract information. MFCC is an illustration of the short-term power spectrum of sound. The Fourier transform representation of the log magnitude spectrum called as the cepstrum coefficients. This coefficient are most robust and more reliable and useful set of feature for speech emotion Recognition and speech recognition [8]-[10] [12]. Therefore the equation below shows by using Fourier transform defined cepstrum of the signal y (n) is

$$CC(n) = FT^{-1}\{\log|FT\{y(n)\}|\} \quad (1)$$

Frequency components of voice signal containing pure tones never follow a linear scale. Therefore the actual frequency for each tone, F measured in Hz, a subjective pitch is measured on a scale which is referred as the 'Mel' scale [6] [10]. The following equation shows the relation between real frequency and the Mel frequency is

$$Mel(f) = 2595 * log_{10}\left(1 + \frac{f}{700}\right) \quad (2)$$

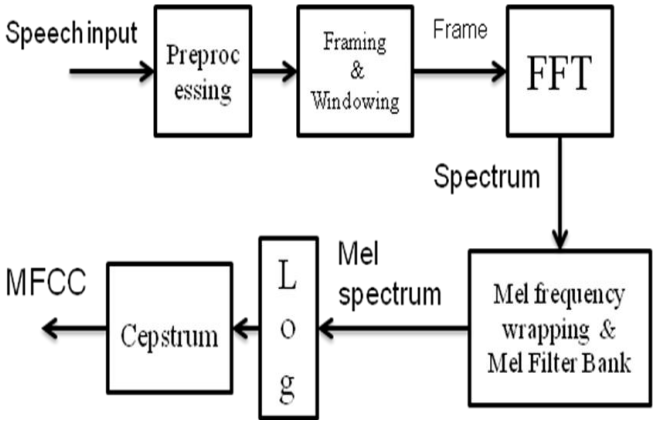

Figure 2: Mel Frequency Cepstum Coefficient Generation

while calculating MFCC firstly performing the pre processing input speech to remove discontinuities next framing and windowing in this process performed windowing over pre-emphasize signal to make frames of 20 sec of speech signal and constructed emotional database & after this the fast Fourier transform is calculated & speech signal spectrum is obtained and then performed Mel frequency wrapping on this spectrum in which spectrum is filtered by a filter bank in the Mel domain. Then taking the logs of the powers at each of the Mel frequencies to obtain Mel spectrum . Then this Mel spectrum is converted by taking log to cepstrum & obtains the Mel frequency cepstrum coefficients. Here we extract the first 12-order of the MFCC coefficients [2][6] [10].

V. Emotional state classification

An emotional state classification has a vital role in emotion recognition system using speech. The accuracy of classification, on the basis of different features extracted from the speech samples of different emotional state. The performance of the system influenced. The classifier is provided by proper features values to classify emotions. In introduction section describes much type of classifiers, out of which K Nearest Neighbor (KNN) and Gaussian mixture model (GMM) classifiers were used for emotion recognition.

5.1 K nearest Neighbor (KNN) classifier

KNN is simplest & influential method of classification of an emotional state, similar observations belong to similar classes is the key idea behind KNN classification.The Nearest Neighbor is the most traditional methods in diverse supervised statistical pattern recognition methods. If costs of error are equal for each class, the estimated class of an unknown sample is selected to be the class that is most commonly represented in the collection of its K nearest neighbors. The nearest neighbor technique based on rather than the classification of only single nearest neighbor, considering the classification of an unknown sample on the "votes" of K nearest neighbor. In this, k is a user-defined constant, and an unlabeled vector is used for classification assigning the label which is most frequent among the k training samples nearest to that query point , in which the input consists of the k closest training examples in the future space. It includes Euclidean distance as the continuous variable as distance [7]. In the training data set the effects of noisy points reduce by Larger K values , cross validation performs the choice of K . The classification of the samples of speech signal in which the nearest training distance is calculated. It involves a training set of all cases. KNN finds the k neighbors nearest to the unlabeled data from the training space based on the selected distance measure. Here we have considered six emotional states namely anger ,happiness , sadness , fear , disgust and Neutral **[9]**

5.2 Gaussian mixture model classifier

GMM is extensively used classifier for the task of speech emotion recognition & speaker identification. It is a probabilistic model for density assessment using a convex arrangement of multivariate normal densities. It is parametric probability density function characterize as a weighted sum of Gaussian component densities. GMM is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. GMMs are broadly used as probability distribution features, such as fundamental prosodic features and vocal-tract related spectral features in an emotion recognition system as well as in speaker recognition systems. GMMs estimated from training data using the iterative Expectation-Maximization (EM) algorithm and using a convex combination of multivariate normal Densities. They model the probability density function of observed data points using a multivariate Gaussian mixture density. After set of inputs given to GMM , by using expectation-maximization algorithm refines the weights of each distribution. Computation of conditional probabilities can be calculated for given test input patterns only when a model is once generated. In these six different emotional states such as anger, happiness, sadness, fear, disgust and Neutral are considered [2][4][11].

VI. Implementation & Experimental Results for Emotion Recognizer

6.1 Experimental Results using KNN Classifier

K nearest neighbor (KNN) based classification of speech emotion recognition is implemented in this experiment on the basis of six different modes for six emotional states, in this first database is sort out according to the different emotional states of speech signal then this sorted database is preprocessed to obtain different training and testing sets are made for emotion recognition then the features were generated from input speech signal. These generated features were added to the database. According to the modes the emission matrix and transition matrix has been made, from this the value of k must be determined on the basis of nearest neighbor value and Euclidian distance has been calculated & classification has been done, the matching of calculated output of KNN with different modes of emotional states is stored in database & obtained the result comparing mode that is most match for emotion recognition.

Table1: K Nearest Neighbor Classifier based Recognition Rate for Emotional states

	Recognized emotions (%)					
Emotion	Anger	Happy	Sad	Fear	Disgust	Neutral
Anger	87	0	0	0	0	13
Happy	0	76.35	5.40	0	0	17.25
Sad	0	0	81.50	9.50	10	0
Fear	15.25	0		74.50		10.25
Disgust	10	0	14.50	0	75.50	0
Neutral	0	15.5	0	0	0	84.5

The emotion recognition rate by using KNN classifier is calculated by passing test input to classifier, which is as shown in table 1 the classification results for Speech emotion recognition with respect to particular mode is obtained. In which for Anger state classifier correctly classified testing speech sample with the recognition rate of 87% as neutral whereas misclassified 13%. Test samples for Happy state were classified as Happy at 76.35% and misclassified 17.25% as neutral state whereas 5.40% as sad state. The test sample for sad

state is correctly classified as 81.50% and also classified as fear state as 9.50%. KNN classifies the fear state with recognition rate of 74.50% & misclassified anger 15.25% and Neutral State 10.25.The disgust state was classified as disgust at 75.50% and also classified angry and sad state as 0% and 14.50%.The neutral state were correctly classified at 84.50% and misclassified 15.5 % as happy state.

6.2 Experimental Results using GMM classifier

While performing emotion recognition using Gaussian Mixture Model (GMM), first the database is sort out according to the mode of classification. In this study, six different emotional states are considered and from different emotional state speech input signal features were extracted. The database is created using extracted features. Then transition matrix and the emission matrix have been made according to modes of emotional states. which generates the random sequence of states and iterative Expectation-Maximization (EM) algorithm is utilized to estimates the probability of state sequence with multivariate normal densities, from this probability of GMM describes matching of mode with the database from the outcome of GMM result obtained as the mode which is most match with the specified mode.

Table2. Gaussian Mixture model Classifier based Recognition Rate for Emotional states

	Recognized emotions (%)					
Emotion	Anger	Happy	Sad	Fear	Disgust	Neutral
Anger	90.6	0	0	0	13.4	0
Happy	0	86	0	0	0	14
Sad	0	15.50	69.60	0	0	14.90
Fear	13.50	0	0	76.50		0
Disgust	0	0	18.50	0	71.50	0
neutral	12	0	0	0	0	78

As shown in the table 2 , the classification results for emotion recognition using GMM classifier with respect to particular mode of different emotional state is calculated in which for Anger state classifier correctly classified testing speech sample with the recognition rate of 90.6% as anger & misclassified 13% as disgust state. For emotional state happy classifier correctly classified at the recognition rate of 86% as happy whereas they were misclassified 14% as neutral state. Test samples for sad state were classified as sad state at 69.60% and misclassified 15.50% and 14.90% as happy state and neutral state respectively. The fear state was classified as fear at 76.50% and also classified 13.50% as anger. The disgust state was classified as fear at 71.50% and also classified 18.50% as sad state. The neutral state were correctly classified at 78.00% and misclassified 12% as anger state. Therefore from this results which were calculated using Gaussian mixture model one can observe that there was confusion between two or three emotional state.

VII. Conclusion

In this paper, utilization of spectral and prosodic feature were extracted from speech signal with different emotional state for Emotion recognition system through speech signal using two classification methods viz. K nearest Neighbor and Gaussian mixture model were studied. Speech Emotion Recognition has a promising future and its accuracy depends upon the combination of features extracted, to increase the performance of system combined features utilized. The feature were extracted from emotional speech samples such as pitch, energy, speech rate, Mel frequency cepstrum coefficient (MFCC) feature and combined to provide better classification and efficiency. Both the classifiers provide relatively similar accuracy for classification. The efficiency of system is extremely depends on proper database of emotional speech sample. Emotional speech database with affective states, improving the performance and engagement of the current interfaces with machine. Therefore it is necessary to create a proper and correct emotional speech database. Emotion recognition system can provide more efficiency with combination of different classifier or implementing hybrid classifiers for better recognition rate.

References

[1]. Ayadi M. E., Kamel M. S. and Karray F., 'Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases', Pattern Recognition, 44 (16), 572-587, 2011.
[2]. A. S. Utane, Dr. S. L. Nalbalwar , "Emotion Recognition through Speech Using Gaussian Mixture Model & Support Vector Machine" International Journal of Scientific & Engineering Research, Volume 4, Issue 5, May -2013
[3]. Chiriacescu I., 'Automatic Emotion Analysis Based On Speech', M.Sc.Thesis, Department of Electrical Engineering, Delft University of Technology, 2009.
[4]. N. Thapliyal, G. Amoli "Speech based Emotion Recognition with Gaussian Mixture Model" international Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 5, July 2012
[5]. Zhou y., Sun Y., Zhang J, Yan Y., 'Speech Emotion Recognition using Both Spectral and Prosodic Features', IEEE, 23(5), 545-549, 2009.

[6]. Chung-Hsien Wu, and Wei-Bin Liang "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels "Ieee Transactions On Affective Computing, Vol. 2, No. 1, January-March 2011.
[7]. Anuja Bombatkar et.al. , "Emotion recognition using Speech Processing Using k-nearest neighbor algorithm" International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, International Conference on Industrial Automation and Computing (ICIAC- 12-13th April 2014
[8]. Dimitrios Ververidis and Constantine Kotropoulo, " A Review of Emotional Speech Databases"
[9]. M. Khan, T. Goskula, M Nasiruddin "Comparison between k-nn and svm method for speech emotion recognition", International Journal on Computer Science and Engineering (IJCSE).
[10]. Rabiner L. R. and Juang, B., 'Fundamentals of Speech Recognition', Pearson Education Press, Singapore, 2nd edition, 2005.
[11]. Xianglin Cheng, Qiong Duan, "Speech Emotion Recognition Using Gaussian Mixture Model" The 2nd International Conference on Computer Application and System Modeling (2012).
[12]. Albornoz E. M., Crolla M. B. and Milone D. H. "Recognition of Emotions in Speech". Proceedings of 17th European Signal Processing Conference, 2009.