Volume 3, Issue 4, April 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Emotion Recognition Through Speech Using Gaussian Mixture Model And Hidden Markov Model

Akshay S. Utane
Departmentof E &TC
Dr. Babasaheb Ambedkar
Technological University, lonere
Raigad, MH, India -402103

Dr. S.L.Nalbalwar
Departmentof E &TC
Dr. Babasaheb Ambedkar
Technological University, lonere
Raigad, MH, India -402103

Abstract— In human machine interaction automatic speech emotion recognition is yet challenging but important task which paid close attention in current research area. As the role of speech is an increase in human computer interface. Speech is attractive and effective medium due to its several features expressing attitude and emotions trough speech is possible. Here study is carried out using Gaussian mixture model and Hidden Markov model classifiers used for identification of five basic emotional states of speaker"s as anger, happiness, sad, surprise and neutral. in this paper to recognize emotions through speech various features such as prosodic features like pitch , energy and spectral features such as Mel frequency cepstrum coefficient were extracted and based on this features emotional classification and performance of classification using Gaussian mixture model and Hidden Markov Model is discussed.

Keywords— Emotion recognition, Feature extraction, Gaussian mixture model, Hidden markov model, MFCC, spectral features and prosodic features

I. INTRODUCTION

Emotion recognition through speech is an area which increasingly attracting attention within the engineers in the field of pattern recognition and speech signal processing in recent years. Automatic emotion recognition paid close attention in identifying emotional state of speaker from voice signal. Emotions play an extremely important role in human life. It is important medium of expressing humans perspective or fillings and his or hers mental state to others. Humans have natural ability to recognize emotions through speech information but the task of emotion recognition for machine using speech signal is very difficult since machine does not have sufficient intelligence to analyze emotions from speech [1]. Recognition of emotions in speech is a complex task that is furthermore complicated because there is no unambiguous answer to what the "correct" emotion is for a given speech sample. The vocal emotions explored may have been induced or acted or they may be have been elicited from more "real", life contexts. Machine can detect who is said and what is said by using speaker identification and speech recognition techniques but if we implied emotion recognition system through speech then machine can also detect how it said [2].as emotions plays an important role in rational actions of human being there is a desirable requirement for intelligent machine human interfaces for better human machine communication and decision making [4]. Emotion recognition through speech means detection of the emotional state of human through feature extracted from his or her voice signal. Emotion recognition through Speech is particularly useful for applications in the field of human machine interaction to make better human machine interface. Some other applications which require natural man machine interaction such as Interactive movie, storytelling, and electronic machine pet, remote teach school & E-tutoring application. Where response of system depends on the detected emotion of users which makes it more practical [4]. Other applications of the emotion recognition system are lie detection ,in the psychiatric diagnosis, intelligent toys, , In aircraft cockpits ,in call center and in the car board system[3]. In the field of emotion recognition through speech several system are proposed for recognizing emotional state of human being from speakers voice or speech signal. On the basis of some universal emotions which includes anger, happiness, sadness, surprise, neutral, disgust, fearful, stressed etc. for this different intelligent systems have been developed by researchers in last two decades. This different system also differs by different features extracted and classifiers used for classification. Prosodic features and spectral features can be used for emotion recognition from speech signal. Because both of these features contain large amount of emotional information. Pitch ,energy, Fundamental frequency, loudness, and speech intensity and glottal parameters are the prosodic features . some of the spectral features are Mel-frequency cepstrum coefficients (MFCC) and Linear predictive cepstral coefficients (LPCC)[5]. Also some of the linguistic and phonetic features also used for detecting emotions through speech. There are several types of classifiers are used for emotion recognition such as Hidden Markov Model (HMM), k-nearest neighbors (KNN), Artificial Neural Network (ANN) , GMM super vector based SVM classifier , Gaussian Mixtures Model (GMM) and Support Vector Machine (SVM). Xianglin Cheng et al. has been performed emotion classification using GMM and obtained the recognition rate of 81%. But this study was limited only on pitch and MFCC features [3].

schuller et al. emotion classification has been performed using HMM and obtained the recognition rate of 84%. as an experiment performed on the Berlin emotional database [2]-[4] -[6]. In this paper , the basic five emotional states such as happy, sad, surprise, angry and neutral state are classified using two different classifiers such as Gaussian mixture model (GMM) and Hidden Markov Model (HMM) classifier 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. The remaining paper is organized as follows: Section two describes about database for emotion recognition system through speech. The section three describes emotion recognition system through speech. The section four describe various extracted features which were used in the emotion classification. The detailed information about the emotion classification by using Gaussian mixture Model and Hidden Markov Model is provided in the Section five. Experimental results obtained during this study were discussed in section six. The section seven is provided Conclusion of this paper.

II. DATABASE SELECTION

In emotion recognition system through speech selection of proper database is a critical task. The efficiency of the speech emotion recognition system is highly depends upon the naturalness of database used in the system . Good recordings of spontaneously produced emotional speech samples are difficult to collect. Different databases are implied by different researchers based on different emotional states of human being. 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. Some of the researchers used Danish emotional corpus database for emotional speech recognition. R. Cowie and E. Cowie constructed their own English language emotional speech database for 5 emotional states such as happiness ,neutral ,fear ,sadness ,anger etc[7]-[8] . In this study we constructed our own database contains short Utterances of emotional speech of speaker"s covering five primary emotional states namely neutral, angry, happy, surprise and sad. Each utterance corresponds to one emotion and by using this database the classification based on GMM and SVM is carried out.

III. EMOTION RECOGNITION SYSTEM THROUGH SPEECH

The block diagram of the emotion recognition system through speech considered in this study is illustrated in Figure 1. Emotion recognition system through speech is similar to the typical pattern recognition system. An important issue in evaluation of Emotion recognition system through speech is the degree of naturalness of the database used. Proposed system is based on prosodic and spectral features of speech. It consists of the emotional speech as input, feature extraction, classification of Emotional state using GMM or HMM classifier and detection of emotion as the output. The emotional speech input to the system may contains the collection of the acted speech data the real world speech data. After collection of the database containing short Utterances of emotional speech sample which was considered as the training samples, proper and necessary features such as prosodic and spectral features were extracted from the speech signal. These feature values were provided to the Gaussian mixture Model and Hidden Markov 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 from the above mentioned five emotions and gives output as recognized emotion[2]-[8].

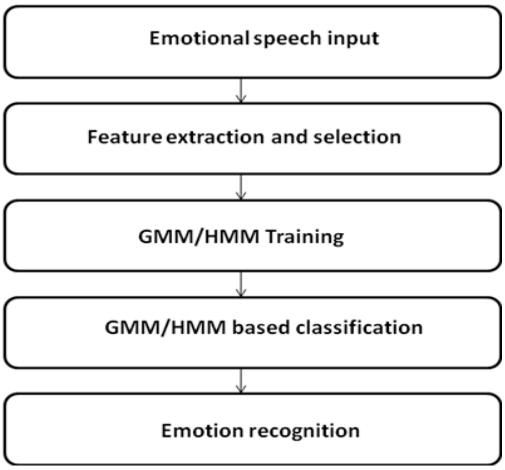
Emotional speech input
Feature extraction and selection
GMM/HMM Training
GMM/HMM based classification
Emotion recognition

Fig 1. Block diagram of Emotion Recognition System through speech

IV. EXTRACTION AND SELECTION OF FEATURES

An important step in emotion recognition System through speech is to select a significant feature which carries large emotional information about the speech signal. Several researches have shown that effective parameters to distinguish a particular emotional states with potentially high efficiency are spectral features such as Mel frequency cepstrum coefficients (MFCC) and prosodic features such as pitch ,speech energy, speech rate ,fundamental frequency. Speech Feature extraction is based on smaller partitioning of speech signal into small intervals of 20 ms or 30 ms respectively known as frames[6]. Speech features basically extracted from vocal tract , excitation source or prosodic points of view to perform different speech tasks. In this work some prosodic and spectral feature has been extracted for emotion recognition. Speech energy is having more information about emotion in speech. The energy of the speech signal provides a representation that reflects these amplitude variations here short time energy features estimated energy of emotional state by using variation in the energy of speech signal. The analysis of energy is focused on short-term average amplitude and short-term energy. We implied short-term function to extract the value of energy in each speech frame to obtain the statistics of energy feature. Another important feature carries information about emotion in speech is pitch. The pitch signal is also called the glottal wave-form. The pitch signal produced due to the vibration of the vocal folds , tension of the vocal folds and the sub glottal air pressure. Vibration rate of vocal cords is also called as fundamental frequency [6]. Another features considering is a simple measure of the frequency content of a signal which is the rate at which zero crossings occur. Zero-crossing rate is a measure of number of times in a given time interval/frame such that the amplitude of the speech signals passes through a value of zero .it is one of the important spectral feature [4].

The next important type of spectral speech features are Mel-frequency cepstrum coefficients (MFCC). It is widely used in speech recognition and speech emotion recognition studies. 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 , Mel-frequency cepstrum feature provide better rate of recognition for speech recognition as well as emotion recognition system through speech [6]. MFCC is a representation of the short-term power spectrum of sound. It is the cepstral analysis is applied in the speech processing to take out the vocal tract information. The Fourier transform representation of the log magnitude spectrum called as the cepstrum coefficients. This high frequency coefficient with high efficiency, are most robust and more reliable and useful set of feature for speech emotion Recognition and speech recognition [8]-[9]. Therefore the equation below shows by using Fourier transform defined cepstrum of the signal y(n) .

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

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 [9]. The following equation shows the relation between real frequency and the Mel frequency is

$$F_{mel} = 3233\log_{10}(\,1 + \frac{F_{HZ}}{1000}\,) \qquad (2)$$

The MFCC coefficients can be obtained as shown in fig 2,

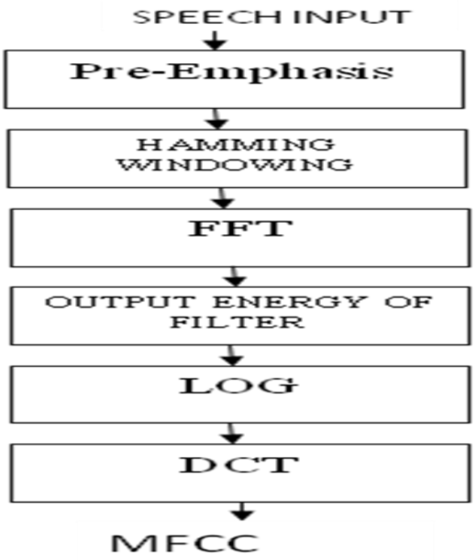

Fig 2. Block diagram of MFCC (Mel frequency cepstrum coefficient)

while calculating MFCC firstly pre-emphasize of speech signal from constructed emotional database has been done .after this performed windowing over pre-emphasize signal to make frames of 20 sec then the Fourier transform is calculated to obtain spectrum of speech signal and this spectrum is filtered by a filter bank in the Mel domain. Then taking the logs of the powers at each of the Mel frequencies . Then the inverse Fourier transform is replaced by the cosine transform in order to simplify the computation and is used to obtain the Mel frequency cepstrum coefficients. Here we extract the first 13-order of the MFCC coefficients [2]-[10].

V. Classification

The most important aspect of emotion recognition system through speech is classification of an emotion. The performance of the system influenced by the accuracy of classification, on the basis of different features extracted from the utterances of emotion speech samples emotions can be classified by providing significant features to the classifier. In introduction section describes many type of classifiers, out of which Gaussian mixture model (GMM) and Hidden markov model (HMM) classifiers were used for emotion recognition.

1] Gaussian mixture model classifier

GMM is parametric probability density function represented as a weighted sum of Gaussian component densities. it is a probabilistic model for density estimation using a convex combination of multivariate normal densities. GMMs estimated from training data using the iterative Expectation-Maximization (EM) algorithm and using a convex combination of multivariate normal Densities. GMMs are widely used as probability distribution features, such as vocal-tract related spectral features in a speaker recognition or emotion recognition systems. GMMs having advantage that are more appropriate and efficient for speech emotion recognition using spectral feature of speech.GMM is parameterized by the mean vectors, covariance matrices and mixture weights from all component 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. Here we have considered Five emotional states namely Happy, Angry, Sad , surprise and Neutral[3]-[11].

2] Hidden markov model classifier

Hidden Markov Model is widely used classifier in the field of speech recognition and emotion recognition system. Hidden Markov Model is having the long history in the field of speech applications. The HMM consist of the first order markov chain whose states are hidden from the observer therefore the internal behaviour of the model remains hidden. The hidden states of the model capture the temporal structure of the data [2]. Hidden Markov Models are statistical models that describe the sequences of events. HMM is having the advantage that the temporal dynamics of the speech features can be trapped due to the presence of the state transition matrix. During classification, the probability for each speech signal calculated from a speech signal which is provided to the model. An output of the classifier is based on the maximum probability that the model has been generated this signal [10].

VI. Experimental Results

1] Experimental Results using GMM

While performing emotion recognition using Gaussian Mixture Model (GMM), first the database is sort out according to the mode of classification. In this study for five modes for five different emotional states then the features were extracted from input waveform. These extracted features were added to the database. According to the modes the emission matrix and transition matrix has been made , which generates the emissions from the model and the random sequence of states then finally estimates the probability of multivariate normal densities of state sequence using iterative Expectation-Maximization (EM) algorithm , 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.The recognition rate for emotion by using GMM is calculated by passing test input to classifier , which is as shown in the Table 1 after passing test samples to classifier. For the happy state test sample classifier correctly classified at the recognition rate of 74.37% as happy whereas they were misclassified 16.57% as surprise and 15.26% as sad state. Test samples for angry state were classified as angry state at 78.27% and misclassified 12.45% as happy state. The neutral state were correctly classified at 73.00% and misclassified 26.89% as sad state. The test sample for sad state is correctly classified as 75.26% and also classified as neutral state as 15.77% and 9.56% for surprise state. The test samples of the surprise state were classified as surprise at 68.39% and also classified angry and happy state as 11.69% and 18.29% respectively. Therefore from this results which were calculated using Gaussian mixture model one can observe that there was confusion between two or three emotional state.

Table 2. Recognition Rate of Emotions Using Hidden Markov Model

EMOTION STATE	EMOTIONS RECOGNIZED (%)				
	HAPPY	ANGRY	NEUTRAL	SAD	SURPRISE
HAPPY	74.37	0	0	15.26	16.57
ANGRY	12.45	78.27	0	0	0
NEUTRAL	0	0	73.00	26.89	0
SAD	0	0	15.77	75.26	9.56
SURPRISE	18.29	11.69	0	0	68.39

2] Experimental Results using HMM

According to the mode of classification first the database which created of emotional speech samples is provided to Hidden Markov Model (HMM), For the emotion recognitio*n* Then the features were extracted from input waveform These features were added to the database. According to this modes The transition matrix and emission matrix has been made , which generates the random sequence of states and emissions from the model. Finally by using Viterbi algorithm Hidden markov model classifier estimated the state sequence probabilities .From this probability of this HMM define the matching of mode with the database from the calculated output of hmm we can put the result comparing mode that is most match for emotion recognition.

Table 2. Recognition Rate of Emotions Using Hidden Markov Model

EMOTION STATE	EMOTIONS RECOGNIZED (%)				
	HAPPY	ANGRY	NEUTRAL	SAD	SURPRISE
HAPPY	71.14	0	0	12.19	15.57
ANGRY	16.29	82.49	0	0	0
NEUTRAL	0	0	74.00	26.00	0
SAD	0	0	27.47	69.68	0
SURPRISE	10.28	20.32	0	0	67.39

As shown in the table, For the happy state test sample classifier correctly classified at the recognition rate of 71.14% as happy whereas they were misclassified 15.57% as surprise and 12.19% as sad state. Test samples for angry state were classified as angry at 82.49% and misclassified 16.29% as happy state. The neutral state were correctly classified at 74.00% and misclassified 26.00% as sad state. The test sample for sad state is correctly classified as 69.68% and also classified as neutral state as 27.47%. The test samples of the surprise state were classified as surprise at 67.39% and also classified angry and happy state as 20.32% and 10.28% respectively

VII. CONCLUSION

In this paper, Emotion recognition through speech using two classification methods viz. Gaussian mixture model an Hidden markov model were studied speech features such as spectral and prosodic feature were extracted from emotional speech samples such as pitch ,energy , formant frequency, speech rate , Mel frequency cepstrum coefficient (MFCC).by using combined features performance of system get increased. Both the classifiers provide relatively similar accuracy for classification. The efficiency of system is highly depending on database of emotional speech sample used in system. Therefore it is necessary to create a proper and correct emotional speech database. For accurate emotional speech database system will provide more efficiency

REFERENCES

[1] Chiriacescu I., ‘Automatic Emotion Analysis Based On Speech’, M.Sc. *Thesis, Department of Electrical Engineering, Delft University of Technology, 2009.*

[2] Ashish B. Ingale, D. S. Chaudhari “Speech Emotion Recognition” *International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012*

[3] Nitin Thapliyal, Gargi Amoli “Speech based Emotion Recognition with Gaussian Mixture Model” *international Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 5, July 2012*

[4] 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.*

[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] Schuller B., Rigoll G., Lang M., ‘Hidden Markov Model Based Speech Emotion Recognition’, IEEE ICASSP, 1-3, 2003.

[7] Dimitrios Ververidis and Constantine Kotropoulo, “ A Review of Emotional Speech Databases”.

[8] 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.*

[9]. Rabiner L. R. and Juang, B., ‘Fundamentals of Speech Recognition’, Pearson *Education Press, Singapore, 2nd edition, 2005.*

[10] Albornoz E. M., Crolla M. B. and Milone D. H. “Recognition of Emotions in Speech”. *Proceedings of 17th European Signal Processing Conference, 2009.*

[11] Xianglin Cheng, Qiong Duan, “Speech Emotion Recognition Using Gaussian Mixture Model” *The 2nd International Conference on Computer Application and System Modelling (2012).*