ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology
Volume 1, Issue 5, July 2012

Speech based Emotion Recognition with Gaussian Mixture Model

Nitin Thapliyal[1]
Assistant Professor UIT, Dehradun
thapliyal.nitin@gmail.com

Gargi Amoli[2]
Assistant Professor DIT, Dehradun
gargi.amoli@gmail.com

Abstract— **This paper is mainly concerned with speech based emotion recognition. The main work is concerned with Gaussian mixture model (GMM model) which allows training the desired data set from the databases. GMM are known to capture distribution of data point from the input feature space, therefore GMM are suitable for developing emotion recognition model when large number of feature vector is available. Given a set of inputs, GMM refines the weights of each distribution through expectation-maximization algorithm. Once a model is generated, conditional probabilities can be computed for test patterns (unknown data points). Expectation maximization (EM) algorithm is used for finding maximum likelihood estimates of parameters in probabilistic models. Moreover Linear Predictive (LP) analysis method has been chosen for extracting the emotional features because it is one of the most powerful speech analysis techniques for estimating the basic speech analysis techniques for estimating the basic speech parameter such as pitch, formants, spectra, vocal tract functions and for representing speech by low bit rate transmission for storage. Speakers are made to involve in emotional conversation with the anchor, where different contextual situations are created by the anchor through the conversation to elicit different emotions from the subject, without his/her knowledge.**

Index Terms— **Speech, Gaussian Mixture Model, Vocal, Emotion Recognition, Linear predictive.**

I. Introduction

Emotional speech reorganization is basically identifying the emotional or physical state of human being from his or her voice[**10**]. Speech is a complex signal containing information about the message, speaker, language and emotions .Speech is produced from a time varying vocal tract system excited by a time varying excitation source. Emotion on other side is an individual mental state that arises spontaneously rather than through conscious effort. There are various kinds of emotions which are present in a speech. Some are ANGER, COMPASSION, DISGUST, FEAR, HAPPY, NEUTRAL, SARCASTIC and SURPRISE. Recognition of Emotions from Speech-Speech features may be basically extracted from excitation source, vocal tract or prosodic points of view to accomplish different speech tasks. Speech features derived from excitation source signal are known as source features. Excitation source signal is obtained from speech, after suppressing vocal tract (VT) characteristics. This is achieved by-

1. First predicting the VT information using filter coefficients (linear prediction coefficients (LPCs)) from speech signal.
2. Then separating it by inverse filter formulation the resulting signal is known as linear prediction residual. It contains mostly the information about the excitation source.

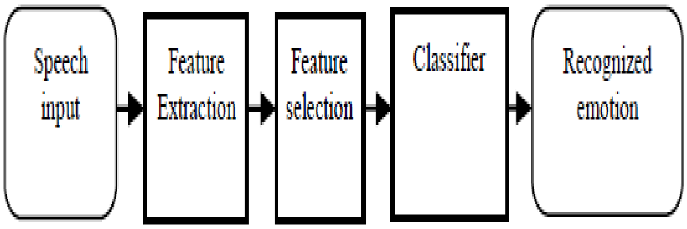


Fig1. Emotion recognition system block

The features derived from LP residual are referred to as Excitation source, sub-segmental, simply source features. LP residual signal basically contains the higher order correlations among its samples as the first and second order correlations are filtered out during LP analysis. These higher order correlations may be captured to some extent, by using the features like strength of excitation, characteristics of glottal volume velocity waveform, shapes of the glottal pulse, characteristics of open and closed phases of glottis and so on. The excitation source information contains all flavors of speech such as message, speaker, language, and emotion specific information. Pitch information extracted from LP residual signal is successfully used in for speaker recognition**[5]**. LP residual energy is used for vowel and speaker recognition. Cepstral features derived from LP residual signal are used for capturing the speaker specific information**[4].** The combination of features derived from LP residual and LP residual cepstrum has been used to minimize the equal error rate in case of speaker recognition. By processing LP residual signal using Hilbert envelope and group delay function, the instants of significant excitation are accurately determined. **Applications**-Speech emotion recognition has several applications in day-to-day life. Some of these are:

I. It is Useful for enhancing the naturalness in speech based human machine interaction.

II. Call center conversation may be used to analyze behavioral study of call attendants with the customers which helps to improve quality of service of a call attendant.

III. Interactive movie, storytelling and E-tutoring applications would be more practical, if they can adapt themselves to listeners' or students' emotional states.

IV. Emotion analysis of telephone conversation between criminals would help crime investigation department.

V. Conversation with robotic pets and humanoid partners would be more realistic and enjoyable, if they are able to understand and express emotions like humans.

VI. In aircraft cockpits, speech recognition systems trained to recognize stressed speech are used for better performance.

II. PROPOSED WORK

In this work, GMMs are used to develop emotion recognition systems using excitation features. GMMs are known to capture distribution of data points from the input feature space. Therefore, GMMs are suitable for developing emotion recognition models using spectral features, as the decision regarding the emotion category of the feature vector is taken based on its probability of coming from the feature vectors of the specific model. Gaussian Mixture Models (GMMs) are among the most statistically matured methods for clustering and for density estimation. They model the probability density function of observed data points using a multivariate Gaussian mixture density. Given a set of inputs, GMM refines the weights of each distribution through expectation-maximization algorithm. Once a model is generated, conditional probabilities can be computed for test patterns (unknown data points). Here we have considered only on four emotions namely Happy, Anger, Sad and Neutral.

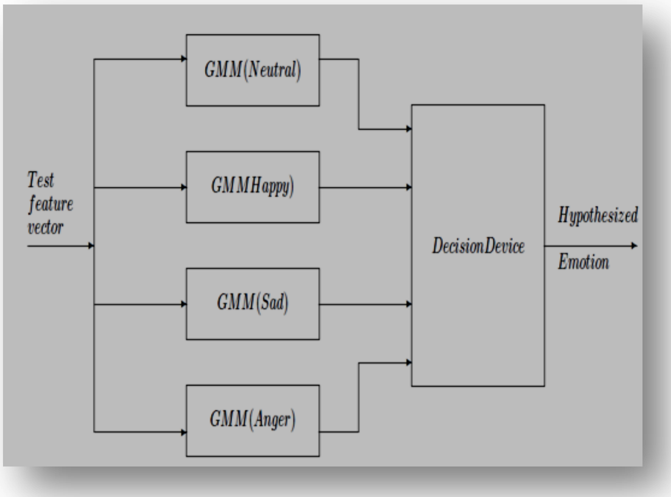


Fig2. GMM model

III. FEATURE EXTRACTION.

Recognition of Emotions from Speech-Speech features may be basically extracted from excitation source, vocal tract or prosodic points of view to accomplish different speech tasks. This work confines its scope to spectral features used for recognizing emotions. Normally excitation features are extracted through block processing approach. Therefore entire speech signal is processed block by block considering the block size of around 20 ms. Blocks are also known as frames. It is assumed that within a block, speech signal is stationary in nature. Block processing approach suffers from some logical problems, they are: physical blocking of speech signal may not be suitable for extracting features, as it is difficult to find relationships among the neighboring feature vectors. Block processing approach blindly processes entire speech signal, but the redundant information present in the regions like steady vowel portion may be exempted from feature extraction. Generally most of the languages in Indian context are syllabic in nature. Performance of emotion recognition systems is found out using semi natural database collected from Hindi movies and simulated database.

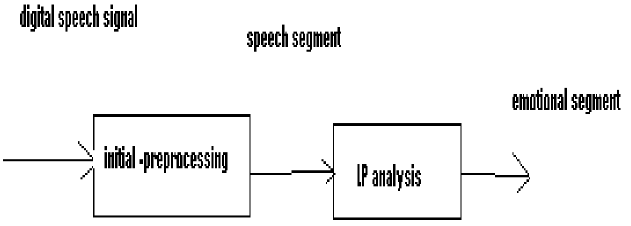


Fig3. Emotion feature Extraction process

Feature extraction basically involves following stages:

1. **Preprocessing**- In this the sample speech which is in digitized form is first normalized by its maximum amplitude and the D.C. component is removed. After this sample is to be divided into 20 msec frames. In order to use the emotion specific information from speech, one needs to extract the features from different levels (sub-segmental, segmental and supra-segmental). Here sub-segmental (excitation source) features are used for analyzing the emotions present in the speech.
2. **Linear predictive analysis**- LP analysis method is one of the most powerful speech analysis techniques for estimating the basic speech parameter. In general, for analysis and processing of speech signal, vocal tract system is modeled as a time varying filter, and presence or absence of excitation causes voiced or unvoiced speech. The equation for LP residual obtained by inverse filtering is as follows:

$$S(n) = 1 + \sum_{k=1}^{p} a_k S(n-k)$$

Where S(n) is current speech sample, p is Oder of prediction a’$_k$ are the filter coefficient and S(n-k) is the (n-k)th sample of speech. Excitation source signal may contain the emotion specific information, in the form of unique features such as higher order relations among linear prediction (LP) residual samples. LP residual signal is obtained by first extracting the vocal tract information from the speech signal and then suppressing it by inverse filter formulation. Resulting signal is termed as LP residual and contains mostly information about the excitation source LP residual is then derived by inverse filtering of the speech signal. All the calculation for LP analysis is developed using MATLAB7 function where input is the segment speech signal and the output are the LPC coefficients which later are used to determine the final parameter for results.

IV. DATABASES

It’s very important that the collection of various emotion voices should have a clear representation of acoustics correlates of one emotion. For this the utterance should be expressed skillfully for the intended emotion. Good recordings of spontaneously produced emotional utterances are difficult to collect. Speech materials that are spontaneously produced have number of drawbacks. The recordings are usually not free of background noises. One of the most important requirements for doing work in speech recognition is a database with the appropriate materials for training and testing the system under development. The size of database is crucial to the achieve and intended results, so collecting and processing data to build a useful database is not trivial. The evaluation of speech emotion recognition system is based on the level of naturalness of database which is used as an input to speech emotion recognition system. If the inferior database is used as an input to the system then incorrect conclusion may be drawn. The database as an input to the speech emotion recognition system may contain the real world emotions or the acted ones. It is more practical to use database that is collected from the real life situations. The database proposed in this paper is collected from Hindi movies by analyzing the emotions from the dialogues being delivered by the film actors/actresses. The database is considered as a semi natural one as simulation the appropriate emotions is close to the real and practical situations. In movies, the emotions expressed are more realistic. Hence, it become easier for an observer to categorize them based on context and by listening the dialogues being spoken by the speaker. Male and female dialogues are separately extracted from the movies of popular actors/actresses to collect desired emotions. While collecting the database, initially the audio is extracted from the video with the help of Adobe Audition, in which the sampling rate of 16 KHz and mono channel with 16 bit resolution are chosen. Speech utterances without background music are extracted carefully to be part of the database based on the contextual emotion present.

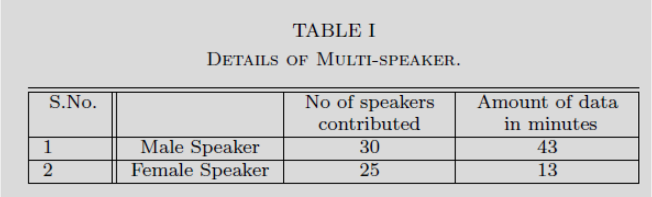


TABLE I
DETAILS OF MULTI-SPEAKER.

S.No.		No of speakers contributed	Amount of data in minutes
1	Male Speaker	30	43
2	Female Speaker	25	13

Table 1: Details of multi speaker.

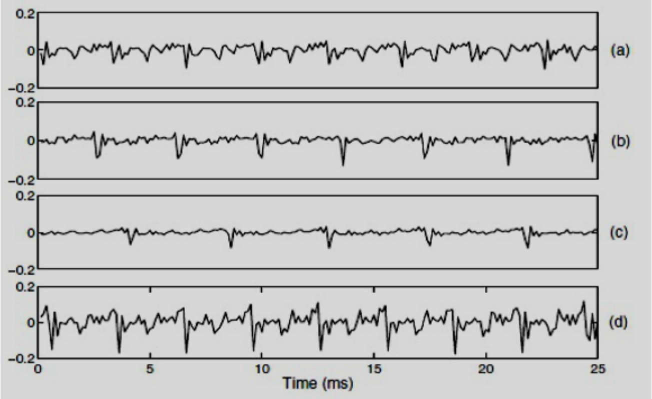
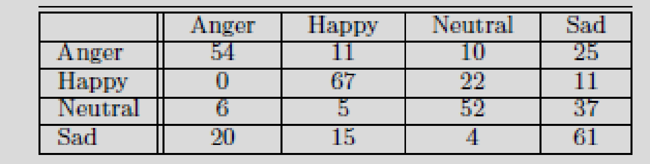


Fig4. LP residual for ANGER, HAPPY, SAD, NEUTRAL.

V. RESULTS

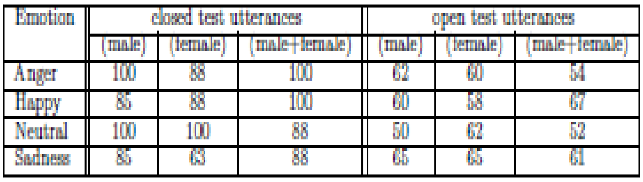
After training and testing the emotion recognition models results are obtained in form of 4X4 matrix form. Each row of matrix represents the test data recognized by different models in form of percentage and each column of the matrix represents the trained model. In this matrix, diagonal elements show the correct classification and other elements in row indicate miss-classification percentages. From table below it is analyzed that 54% of anger is classified as anger similarly 67%, 52% and 61% of happy, neutral and sad emotions are recognized correctly for single male speaker. So the total result of this matrix is analyzed is 59%.



	Anger	Happy	Neutral	Sad
Anger	54	11	10	25
Happy	0	67	22	11
Neutral	6	5	52	37
Sad	20	15	4	61

Table2:Emotion classification performance (in percentage)

Later, the same process is used for female and male+female (male and female utterances for training and testing are used simultaneously) speakers. Table III shows the comparison of emotion recognition performance for different 4 emotions for closed and open set. Where closed set means same utterances are used for training as well as testing whereas open set means different set of utterances are used for training and testing. The emotion recognition performance in case of closed set utterances has been observed to be 92.5%, 84.75% and 94% for male, female and male+female speakers respectively.



Emotion	closed test utterances			open test utterances		
	(male)	(female)	(male+female)	(male)	(female)	(male+female)
Anger	100	88	100	62	60	54
Happy	85	88	100	60	58	67
Neutral	100	100	88	50	62	52
Sadness	85	63	88	65	65	61

Table3. Comparison of emotion recognition

The emotion recognition performance in case of open set utterances has been observed to be 59.25%, 61.25% and 58.5% for male, female and male+female speakers respectively. It is observed that in case of closed set utterances, better results are obtained as compared to open set utterances. This is so because in case of closed set utterances, same utterances are used for training as well as testing the emotion recognition models. While in case of open set utterances, different utterances have been used for training and testing the emotion recognition models. Another possible reason for low recognition rate for open set utterances could be: use of less speech utterances for training and testing the emotion recognition models.

VI. CONCLUSION

In this paper work Excitation Source features are used for characterizing the emotions present in a speech. As usual, the emotion recognition performance using LP residual is not sufficient enough to develop sophisticated emotion recognition system. The performance may be improved by combining the different features such as spectral and prosodic. Overall emotion recognition performance is observed to be about 52-60%.

.

REFERENCES

[1] S. Prasanna, C. Gupta, and B. Yegnanarayana, ”Extraction of speakerspecific information from linear prediction residual of speech,” J. Acoust., Soc. , Amer. Speech Communication, vol. 48, pp. 1243-1261, Oct. 2006.

[2] B. AtaI, ”Automatic speaker recognition based on pitch contours,” J. Acoust. Soc. Amer., vol. 52, pp. 1687- 1697, March 1972.

[3] H. Wakita, ”Residual energy of linear prediction to vowel and speaker recognition,” IEEE Trans. Acoust. Speech Signal Process. vol. 24, pp. 270-271, April 1976.

[4] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, ”Speech enhancement using excitation source information,” Proc. IEEE Int. Con! Acoust. , Speech, Signal Processing, vol. 1, pp. 541-544, May 2002.

[5] A. Bajpai and B. Yegnanarayana, ”Combining evidence from subsegmental and segmental features for audio clip classification,” TENCONIEEE region 10 corifTences, pp. 1-5, Nov 2008.

[6] J. Benesty, M. M. Sondhi, and Y. Huang,”Springer handbook on speech processing,” Springer Publisher, 2008.

[7] C. M. Lee and S. S. Narayanan, toward detecting emotions in spoken dialogs, IEEE Trans. Speech and Audio Processing,vol. 13, pp. 293303, Mar. 2005.

[8] J. Nicholson, K. Takahashi, and R. Nakatsu,”Emotion recognition in speech using neural networks,” Neural Computing and Applications, vol. 9, pp. 290-296, Dec. 2000.

[9] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," Neural Computing and Applications,vol. 9, pp.290-296, Dec. 2000.

[10] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition.Englewood Cliffs, New Jersy: Prentice- Hall, 1993.

[11] T. V. Ananathapadmanabha and B. Yegnanarayana, "Epoch exctraction from linear prediction residual for identification of closed glottis interval," IEEE Trans. Acoustics, Speech and Signal Processing, vol. 27, pp. 309-319, 1979 1997.

[12] F. Charles, D. Pizzi, M. Cavazza, T. Vogt, and E. Andr, Emoemma: Emotional speech input for interactive storytelling, in 8th Int. Conf. on Autonomous Agents and MultiagentSystems (AAMAS 2009) (Decker, Sichman, Sierra, and Castelfranchi, eds.), (Budapest, Hungary), pp. 13811382, May 2009.