Volume 4, Issue 5, May 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Emotional Speech Recognition with Gaussian Mixture Model with Reference to Bodo Language

Barnali Kalita*
Department of Instrumentation & USIC
Gauhati University, Guwahati Assam.

Prof. P.H.Talukdar
Department of Instrumentation & USIC
Gauhati University, Guwahati Assam.

Abstract— The research paper is basically with emotion recognition of Bodo speech using Gaussian Mixture Model (GMM) which allows training the desired data set from the main databases. Since GMM are suitable for developing emotion recognition model when large number of feature vector is available and hence GMM are known to capture distribution of data point from the input feature space. From a given a set of inputs, GMM refines the weights of each distribution through expectation-maximization algorithm which is used for finding maximum likelihood estimates of parameters in probabilistic models. The Linear Predictive (LP) analysis method is for extracting the emotional features since it is one of the powerful speech analysis techniques for calculating the basic speech parameter such as pitch, formants, 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.

Keywords— Bodo, 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[**1**]. Speech is a complex signal containing information about the 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 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. Block diagram of Emotion recognition system

The features which are derived from residual of LP analysis are referred to as Excitation source or these are simply source features. During LP analysis, the residual signal contains the higher order correlations. Among its samples as the first and second order correlations are filtered out. By using the features like strength of excitation, characteristics of glottal volume velocity waveform, characteristics of open and closed phases of glottis etc, these higher order correlations may be captured to some extent. The Pitch data which are extracted from LP residual signal is used for speaker recognition**[6]**. For vowel and speaker recognition LP residual energy emotion is used. Again for capturing the speaker specific information**[7] ,** Cepstral features that are derived from LP residual signal are used. Again in case of speaker recognition, the combination of features that are derived from LP residual and LP residual cepstrum has been used to minimize the equal error rate by processing LP residual signal. The instants of significant excitation are accurately determined by using Hilbert envelope and group delay function.

Bodo- Before 1954, the Bodo language had no standard form of writing. It had a history of using Deodhai, Roman and Assamese scripts. At present, Bodos adopted the Devanagari script. But, there is a huge difference in the usage of the letters in Bodo language from the Devanagari script. Bodo language shares some common salient features with other languages belonging to the Bodo group. These features are similar in terms of phonology, morphology,

syntax, and vocabulary. Bodo language is closely associated with the Dimasa language of the state of Assam and with the Garo language of the state of Meghalaya, and also with Kokborok language of Tripura. It important to note that ,among the four districts of present Bodo land , namely, Kokrajhar, Chirang, Baksa and Udalguri, the language is heard in pure form only in the district of Udalguri. The language is affected by other communities , mostly, Assamese, Bengali and Hindi speaking communities.

***A)* Dialects in Bodo Language**

The Bodo language has four clear-cut dialects-areas with a sufficient number of dialectal variations (P.C.Bhattacharya, A Descriptive Analysis of the Bodo Language):

- ✓ The North-West dialect area which covers the northern regions of Goalpara and Kamrup district of Assam.
- ✓ The South-West dialect area which covers South Goalpara, Garo Hills and a few places of south Kamrup of Assam.
- ✓ The North-Central Assam dialect area covering the entire district ofDarrang, Lakhimpur and a few places of Arunachal pradesh.
- ✓ The southern Assam dialect area covering district of Nowgong, North Cachar and Mikir Hills along with some adjacent areas.

***B)* Phonological Structure**

The Phonology is the study of speech sound and their functions within the sound system of a particular language. Phonemes are considered as the basic unit of a language.

The Bodo phonemes consists of 6(six) vowels and 16(sixteen) consonants. Out of these 16 consonants 2(two) are semi vowel. They are as shown below-

a. Vowels : अ, आ, इ, उ, ए, औ

b. Consonants : ख, ग, ङ, ज, थ, द, न, फ, ब, म, र, ल, स, ह

c. Semi Vowels : य, व

Applications-Now a days the Speech emotion recognition has several different applications. Here we have discussed few of them. That are:

I. In aircraft cockpits, speech recognition systems trained to recognize stressed speech.

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

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

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

II. PROPOSED WORK

By using excitation features, we have used GMMs to develop an Bodo based Emotional Speech Recognition systems of Bodo language. GMMs are known to capture distribution of data points from the input feature space and hence for developing emotion recognition models GMMs are suitable. For clustering and for density estimation, Gaussian Mixture Models (GMMs) are among the most statistically matured methods. Using a multivariate Gaussian mixture density, we model the probability density function of observed data points. GMM refines the weights of each distribution through expectation-maximization algorithm from a given set of inputs. Here we have considered only four emotions namely Neutral, Happy, Anger and Sad.

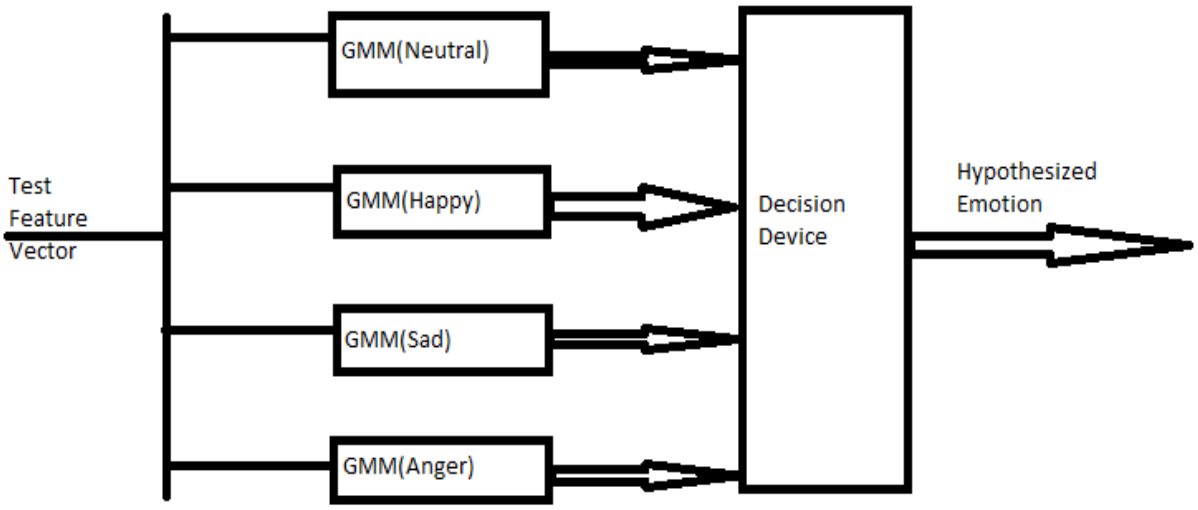

Fig2. GMM model

III. FEATURE EXTRACTION

Recognition of Emotions from Speech- To accomplish different speech tasks Speech features are extracted from excitation source, vocal tract or prosodic points of view. We have concentrated its scope to spectral features used for recognizing Bodo emotions. Using block processing approach excitation features are extracted and so the entire speech signal is processed block by considering the block size of around 20 ms. Normally the blocks are also known as frames.

We have assumed that the speech signal is stationary in nature within a block. In Block processing approach we suffer from some logical problems i.e. it is difficult to find relationships among the neighboring feature vectors that physical blocking of speech signal may not be suitable for extracting features. The Block processing approach blindly processes entire speech signal. Here 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. The Bodo emotion recognition systems is found out using semi natural database collected from Bodo tele movies and simulated database.

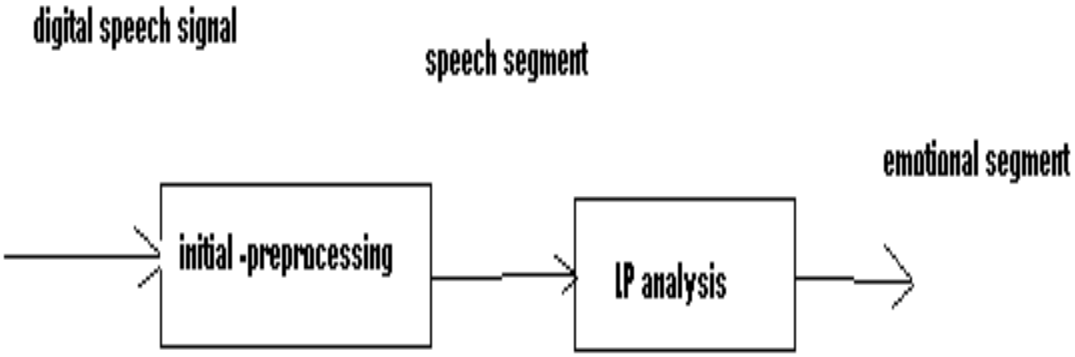

Fig3. Emotion feature Extraction process

Following are the basic stages involved in Feature extraction:

1. **Preprocessing**- In the first phase, the digitized sample speech form is normalized by its maximum amplitude and remove the D.C. component of the speech and after that the sample speech is divided into 20 msec frames. Normally one needs to extract the features from different levels for example sub-segmental, segmental and supra-segmental in order to use the emotion specific information from speech. In our experiment, sub-segmental i.e. excitation source, features are used for analyzing the emotions.

2. **Linear predictive analysis**- It is one of the most powerful speech analysis techniques for estimating the basic speech parameter in speech recognition. The vocal tract system is modeled as a time varying filter and presence or absence of excitation source causes voiced or unvoiced speech. For analysis and processing of speech signal, the equation for **Linear predictive** residual obtained by inverse filtering is as follows:

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

Here S(n) is current speech sample, p is order of prediction a'k are the filter coefficient and S(n-k) is the (n-k)th sample of speech.

The excitation source signal may contain the emotion specific information. It is in the form of unique features such as higher order relations among linear prediction (LP) residual samples. **Linear predictive** residual signal is obtained by first extracting the vocal tract information from the speech signal. After LP residual signal obtained it is then suppressed by inverse filter formulation and which is termed as LP residual and contains mostly information about the excitation source. We have used MATLAB7 function to analyze LP residual signal, where input is the segment speech signal and the output are the LPC coefficients.

IV. DATABASES

To have a clear representation of acoustics correlates of one emotion the collection of different emotion voices are very important and for this the utterance should be expressed skillfully for the intended emotion. There have a number of drawbacks for speech materials that are spontaneously produced. The main drawback is that recordings are usually not free of background noises. For doing work in speech recognition a database with the appropriate materials for training and testing the system is more important. The size of database is crucial to achieve and intended results. So collecting and processing data to build a useful database is not trivial. The speech emotion recognition system is completely based on the level of naturalness of database which is used as an input to speech emotion recognition system. The database as an input to the speech emotion recognition system may contain the real world emotions or the acted ones. The database that we have proposed in our experiment is collected from Bodo tele movies by analyzing the emotions from the dialogues being delivered by the film actors/actresses. Here the database is considered as a semi natural one. As simulation the appropriate emotions is close to the real and practical situations. But in movies, the emotions expressed are more realistic. So it is easy to categorize them based on context and by listening the dialogues being spoken by the

speaker for an observer Male and female dialogues are separately extracted from the movies of popular actors/actresses to collect desired emotions. At the time of collecting the database, first the audio is extracted from the video with the help of Adobe Audition, in which the sampling rate we have consider is of 16 KHz and the channel with 16 bit resolution.

TABLE I
Multi Speaker

Sl. No	Speaker	No of speaker	Data (in min)
1	Male	50	63
2	Female	45	23

Table 1: Details of multi speaker.

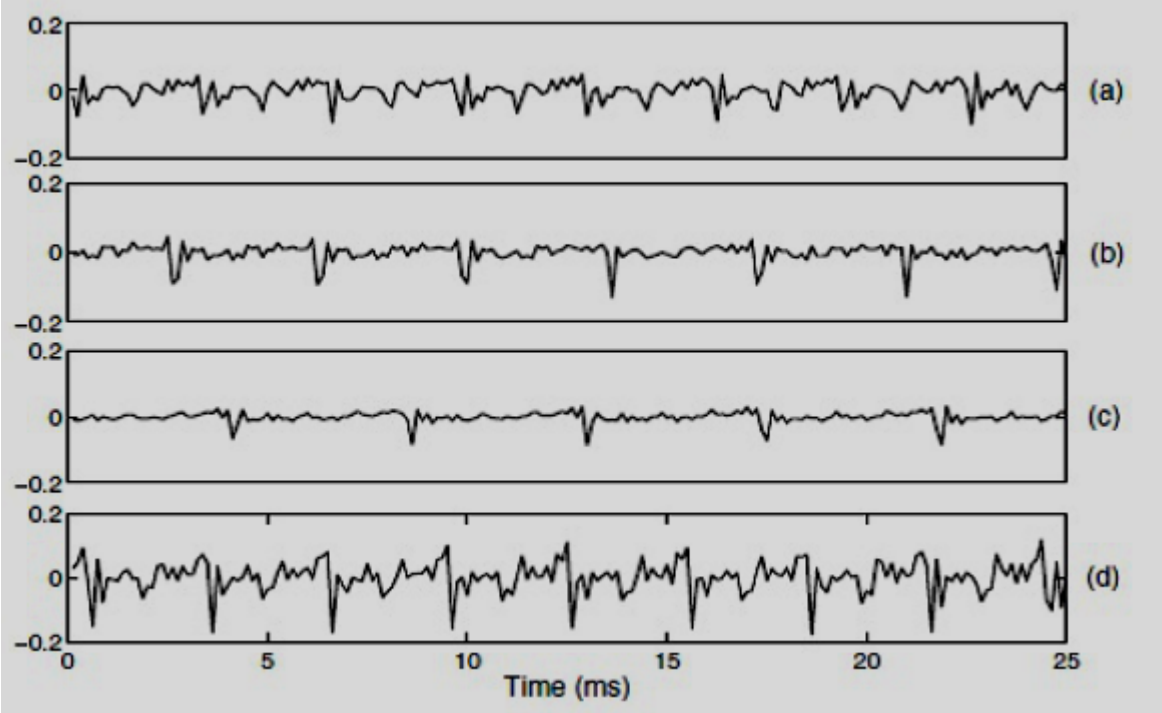

Fig1. LP residual for Anger, Happy, Sad and Neutral

V. RESULTS

The results so obtained are in form of 4x4 matrix after training and testing the emotion recognition models. Each row represents the test data recognized by different models and each column represents the trained model. The diagonal elements show the correct classification. The other elements in row indicate miss-classification percentages.

Emotion	Anger	Happy	Neutral	Sad
Anger	**55**	**12**	**11**	**26**
Happy	**1**	**66**	**23**	**12**
Neutral	**7**	**6**	**53**	**38**
sad	**21**	**17**	**5**	**62**

Table2:Emotion classification performance (in percentage)

From the observation we have obtained that 55%, 68%, 54% and 62% of anger , happy, neutral and sad emotions for male speaker respectively. So the overall value that we have obtained 61%. The same process is used for female and male+female speakers. We have observed in two modes i.e closed and open set. Where close 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 91.5%, 83.65% and 93% for male, female and male+female speakers respectively. The emotion recognition performance in case of open set utterances has been observed to be 60.25%, 62.05% and 58.5% for male, female and male+female speakers respectively. We have 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 our study Excitation Source features are basically used for characterizing the emotions present in a speech. To develop a sophisticated emotion recognition system the emotion recognition performance using LP residual is not sufficient. By combining the different features such as spectral and prosodic, the performance can be improved. In our experiment the emotion recognition performance is observed to be about 55-64%.

REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[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.

