ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

Gender Dependent and Independent Emotion Recognition System for Telugu Speeches Using Gaussian Mixture Models

Kalyana Kumar Inakollu [1], Sreenath Kocharla[2]

(M.Tech), Department of CSE, QIS College of Engineering and Technology, Ongole, India [1]

Asst.Professor, Department of IT, QIS College of Engineering and Technology, Ongole, India [2]

Abstract--Speech based emotion recognition has its utility in the real world applications. Gaussian Mixture Model (GMM) is used to achieve it. This model is used to train dataset required for the task and suitable for recognizing speech based emotions. GMM makes use of Exception Maximization (EM) algorithm for estimation maximum likelihood of parameters. Based on the GMM model, conditional probabilities are computed for data points which are not known priori. To extract emotional features from the speech Mel Frequency Cepstral Coefficients (MFCCs) method is used. The parameters used for this include vocal tract functions, spectra and pitch formants. In this paper we build a prototype application to demonstrate the concept of speech based emotion recognition. The empirical results revealed that the proposed application is effective.

Index Terms – Emotion recognition, GMM, MFCCs, EM

I. INTRODUCTION

Emotion recognition in speech can be used in various applications as they have utility. Essentially it is used to identify the physical state of human being who produces emotional speech [1]. Among the signals speech is the complex thing that has emotions that can be interpreted to know human behavior. Vocal tract system is excited in order to produce speech. Emotion of human being represents the mental state of the human being. The emotions are of many types. They include disgust, fear, anger, compassion, surprise, sarcastic, happy and so on. The emotions are extracted from the source and vocal tract points in order to achieve speech tasks. Recognition needs certain systematic steps in order to make it useful in applications[2]. From speech source of excitation is obtained which is attributed to the suppression of vocal tract (VT)[3] and its characteristics.

First of all, filter coefficients are used in order to predict VT information. Then inverse filter is used to produce signal that is known as linear prediction residual. The result contains excitation source. The LP residual [4] is used to derive features which are known as source features of excitation sources. Higher order correlations are found in LP residual which helps in LP analysis. The higher order correlations can be obtained using features like glottal pulse, velocity waveform, and strength of excitation and so on. The information such as excitation source contains many flavors of speech such as emotion, language, speaker, and message. These can be used in order to recognize emotions of human beings successfully [5]. For vowel and speaker recognition LP residual energy is sued. LP residual signal is also used to derive Cepstral features which can be used for the purpose of emotion recognition [6]. The combination of LP residual cestrum and LP residual features are used to minimize error rate in recognition. Hilbert envelope is used to process LP residual signals[7], [8]. There are many applications which can benefit from emotion recognition in speeches.

- Human machine interaction can be improved.
- Call centers can use it to analyze behavior of customers for serving them better or to create strategies to improve business.
- Interactive applications such as E-touring, storytelling and movie can be built using this for better understanding.
- Conversation between criminals can be intercepted in order to make use of it in legal applications.
- Conversation with robotic animals can help to achieve realistic results.
- In aircrafts it can be used for better performance.

In this paper we built a prototype application that demonstrates the proof of concept. Empirical results revealed that the application is very useful. The remainder of the paper is structured as follows. Section II provides review of literature. Section III describes the proposed system. Section IV presents experimental results while the section V concludes the paper.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

II. PRILIMINARIES

This section provides required details before going to proposed emotion recognition system. It throws light into GMM, MFCC and EM methods in some detail.

Gaussian Mixture Model (GMM)

GMM is a model which is widely used to built emotion recognition systems that interpret emotions in human speech. This model is best used to capture the data point distribution from the given feature space. With this features GMMs are made suitable for captutuing human emotions from speech[9] and [10]. This model can support spectral features which will help in making well informed decisions while recognizing emotions. GMM is one of the most matured methods that can interpret statistics in best way. They are used for estimation of density using probability density function of data points that have been observed. Given a set of inputs, the GMMs compute distribution weights using an algorithm known as expectation maximization[11] and [12]. After generating model it is possible to extract patterns from the data that can eventually be recognized as an emotion such as natural, sad, anger or happy.

Mel Frequency Cepstral Coefficients (MFCCs)

Mel Frequency Cepstral Coefficients is one of the techniques used to recognize speech and speaker automatically. In 1980's it was introduced by Davis and Mermelstein that have been widely used for speaker recognition. In fact it is one of the state-of-the-art techniques available. Before this model came into existence, LPCs (Linear Prediction Coefficients) and LPCCs (Linear Prediction Cepstral Coefficients) were used as features for recognizing human speeches.

Expectation Maximization

It is a widely used algorithm used to obtain maximum likelihood estimation of given parameters. It is used with statistical data to make certain models which depend on hidden variables that have not been observed. It is basically an iterative model that has an expectation step and a maximization step. The expectation step is used to create a function obtaining likelihood values and evaluate them. The latter step is used for parameter maximization. This will help in finding the distribution of hidden values in the data. This it can be used in various applications including speech recognition.

III. PROOPSED EMOTION RECOGNITION SYSTEM

We propose a general framework that can be used to recognize emotions. The framework takes speech of human beings as input and generates output in the form of emotion recognition. The resultant emotions include disgust, fear, anger, compassion, surprise, sarcastic, happy and so on.

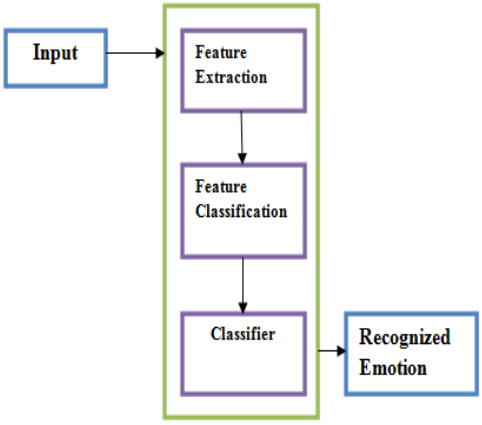

Fig. 1 – Block diagram for emotion recognition

As can be seen in figure 1, the system takes speech as input and extracts features available. As discussed earlier, it chooses the features that are required by the system. Then a classifier is applied which classifies the features in order to recognize the emotions. Towards practical implementation GMMs are used. They are nothing but capture distribution of data points. They are best used in emotion recognition. They are statistically matured when compared to their counterparts for emotion recognition. Given the set of inputs, the GMMs make use of algorithm such as expectation maximization algorithm in order to refine weights of each distribution. Thus it generates a model that can be used further. The model contains test patterns. Only four emotions are considered for experiments. They include Neutral, Sad, Anger and Happy. The GMM model is used to achieve this as given in figure 2.

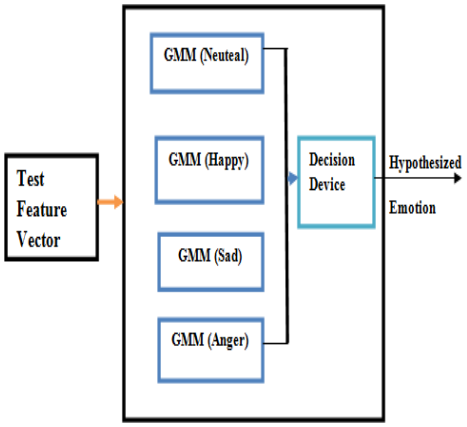

Fig. 2 - GMM models for four emotions

The GMM models take test feature vector as input and process them. The results are sent to decision device which can recognize the emotion. The features required by the GMMs are generally extracted from prosodic, vocal tract

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

and excitation sources. However, in this paper we focused on spectral featured in order to recognize emotions. Block processing is generally used to obtain excitation features. Thus the speech signal is processed around 20 ms block by block. However this kind of processing has some drawback as it suffers from physical blocking speech signals. As it blindly processed entire signal, it is not suitable for optimal performance. For this reason in this paper we use different emption feature extraction process which is as shown in figure 3.

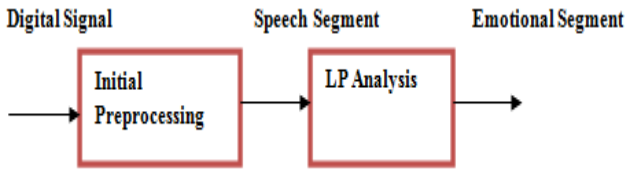

Fig. 3 – Illustrates process of extraction of emotional features

As can be viewed in figure 3, original speech signal is taken by the process and it is subjected to initial preprocessing. After preprocessing the speech segment is given to LP (Linear Productive) analysis which will produce emotional segment. In preprocessing amplitude is normalized and the gets rid of D.C. After this work, the sample is divided into frames of width 20msec. After wards excitation source features are extracted for further processing. The LP analysis is very useful for speech analysis. It can estimate basic speech estimation parameter. For obtaining LP residual the following equation is used.

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

Here the current speech sample is represented by S(n). Order or prediction is represented by p. emption specific information is found in features such as higher order relations. Vocal tract information is taken from speech signal and then inverse filter is applied. The resultant signal is known as LP residual. Then the LP analysis is made in order to produce emotional segments.

IV. EXPERIMENTAL RESUTLS

Experiments are made by building a custom simulator application. The application takes human voices and recognizes emotions in the voice. The emotions considered are anger, happy, neutral and sadness. The environment used for building the application is a PC with 4 GB RAM, core 2 dual processor running Windows 7 operating system. The classification performance and the experimental results are presented below. Male and female voices are collected in order to make experiments. The experimental results with respect to comparison of emotion recognition are presented in table 1.

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

Table 1 – Experimental results

As can be seen in figure the comparison of emotion recognition is given in the form of statistics. The results include closed test utterances and open test utterances. The results are visualized as shown in following figures.

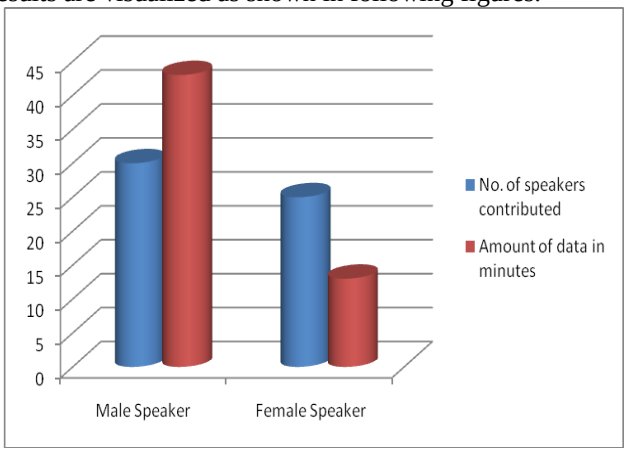

Fig. 4 - Multi-speaker contribution details

As seen in figure 4 multiple speakers have contributed for experiments. The number of male speakers is 30 and the number of female speaker is 25. The amount of data collected from male speakers is 43 minutes while 13 minutes data is collected from female speakers.

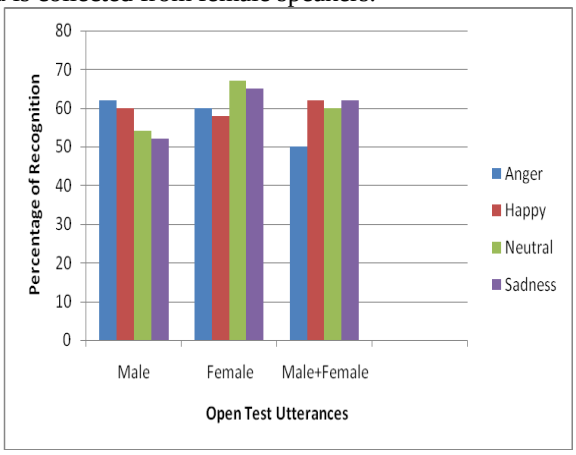

Fig. 4 – Results of closed test utterances

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

As can be seen in figure 4, it is evident that the four emotions are recorded for male and female voices. The horizontal axis represents voice samples of male, female and male plus female respectively with respect to closed test utterances. The vertical axis represents percentage of recognition.

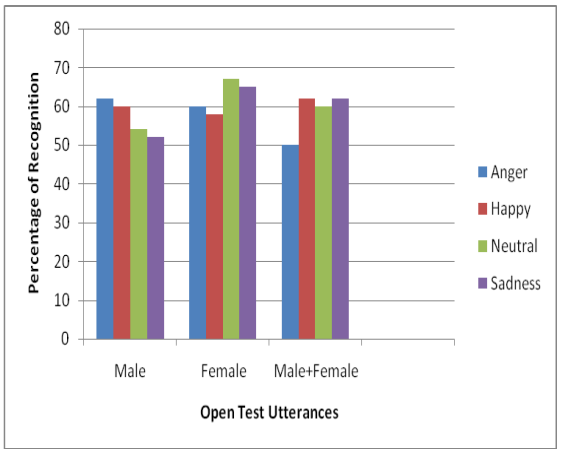

Fig. 5 – Results of closed test utterances

As can be seen in figure 5, it is evident that the four emotions are recorded for male and female voices. The horizontal axis represents voice samples of male, female and male plus female respectively with respect to open test utterances. The vertical axis represents percentage of recognition.

V. CONCLUSION

In this paper we studied the utility of speech based emotion recognition and its advantages in the real world applications. The emotions present in speech can help to interpret human behaviors so as to make expert decisions. LP residual is not adequate in order to recognize emotions in speech. With LP residual it is not possible to build a sophisticated emotion recognition system that can be used in real world applications. However, it can be combined with the features such as prosodic and spectral in order to make it more robust and useful for emotion recognition. With this combination the performance is improved up to 60%. We built a prototype application that demonstrates the efficiency of the application. The empirical results revealed that the application is very useful.

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] Gregor Domes et al "Emotion Recognition in Borderline Personality Disorder- A review of the literature" journal of personality disorders,23(1),6-9,2009.

[10] George Almpanidis and Constantine Kotropoulos" Phonemic Segmentation Using the Generalised Gamma Distribution and Small Sample Bayesian Information Criterion.Vol 50,Issue-1,pp 38-55,January-2008.

[11] Lin Y.L and wei G" Speech Emotion Recognition based on HMM and SVM" 4th international conference on machine learning and cybernetics,Guangzhou,Vol.8,pp4898-4901,18-aug-2005.

[12] Prasad Reddy P.V.G.D et al "Gender Based Emotion Recognition System for Telugu Rural Dialects Using Hidden Markov Models" journal of advanced research in computer engineering: journal of computiong Vol-2,c`issue 6,pp 94-98 ,june 2010.