Speaker Dependent Emotion Recognition from Speech

Biswajit Nayak, Mitali Madhusmita, Debendra Kumar Sahu, Rajendra Kumar Behera, Kamalakanta Shaw

Abstract— ***The speech signal is the fastest and the most natural method of communication between humans. Hence speech can use for fast and efficient way of interaction between human and machine. Speech is attractive and effective medium due to its several features expressing attitude and emotions through speech is possible. In human machine interaction automatic speech emotion recognition is so far challenging but important task which paid close attention in current research area. In this paper we have analysed emotion recognition performance on eight different speakers. IITKGP-SEHSC emotional speech database used for emotions recognition. The emotions used in this study are anger, fear, happy, neutral, sarcastic, and surprise. The classifications were carried out using Gaussian Mixture Model (GMM). Mel Frequency Cepstral Coefficients (MFCCs) features are used for identifying the emotions. It can be observed that, the percentage of accuracy is 75.00% for 32 centered GMM, 72.00% for 16 centered GMM and 66.67% for 8 centered GMM.***

Keywords— ***Emotion Recognition, Gaussian Mixture Model (GMM), Male-scale Frequency Cepstral Coefficient (MFCC), IITKGP-SEHSC (Indian Institute of Technology Kharagpur Simulated Hindi Emotional Speech Corpus).***

I. INTRODUCTION

An important issue in speech emotion recognition is the need to determine a set of the important emotions to be classified by an automatic emotion recognizer. Human being express there feelings through emotions, and the way of expression may be through face, gesture and speech. Emotions are essential for conveying crucial information. Presence of emotion makes speech more natural. Human being use emotion extensively for expressing their intention through speech. It is observed that same message can be conveyed in different way by using appropriate emotion. Speech signal contain information like intended message, speaker identity, emotion state of speaker.

The speech signal is the fastest and the most natural method of communication between humans. Hence speech can use for fast and efficient way of interaction between human and machine. However, this requires that the machine should have the sufficient intelligence to recognize uman voices. We are still far from having a natural interaction between man and machine because the machine unable to understand the emotional state of the speaker. Therefore the identification of emotion present in the speech is necessary to understand the emotional state of human interpret the message properly.

So there is needed to develop speech system that recognizes emotion efficiently.There are different types of emotions present it’s very difficult to classify all these emotions. Many researchers agree with the ‘palette theory’, which states that any emotion can be decomposed into primary emotions similar to the way that any color is a combination of some basic colors. Primary emotions are Anger, Fear, Happy, Neutral, Sadness, and Surprise.

The task of speech emotion recognition is very challenging for the following reasons. First, it is not clear which speech features are most powerful in distinguishing between emotions. The acoustic variability introduced by the existence of different sentences, speakers, speaking styles, and speaking rates adds another obstacle because these properties directly affect most of the common extracted speech features such as pitch, and energy contours. Moreover, there may be more than one perceived emotion in the same utterance; each emotion corresponds to a different portion of the spoken utterance.[1][2][5]

The structure of the paper is as follows. Next section, section II describes the emotional speech database used in this work. Section III provides the male frequency cepstral coefficient. Section IV describes the Gaussian Mixture Model for speech emotion recognition. Section V describes the Architecture of Emotion recognition system. Section VI presents the experiment study, results got in experiment and observations on those results. And finally section VII gives conclusion and at the end references.

II. EMOTIONAL SPEECH DATABASE

For characterizing the emotions, both for synthesis or for recognition, a suitable emotional speech database is a necessary prerequisite.The design and collection of emotional speech corpora mainly depends on the research goals. For example a single speaker emotional speech corpus would be enough for the purpose of emotional speech synthesis, whereas recognizing emotions needs a database with multiple speakers and various styles of expressing the emotions.

The survey presented in this section critically analyzes the emotional speech databases based on the language, number of emotions and the method of collection. This approach has been verified using IITKGP-SEHSC database to carry out the emotion classification. This database is particularly designed and developed at the Indian Institute of Technology, Kharagpur, to support the study on speech emotion recognition. The proposed speech database is the first one developed for analyzing the common emotions present in day-to-day conversations.

Manuscript Received November, 2013.

Biswajit Nayak, Computer Science and Engineering, Bhubaneswar Engineering College, Bhubaneswar, India.

Mitali Madhusmita, Computer Science and Engineering, The TECHNO SCHOOL, Bhubaneswar, India.

Debendra Kumar Sahu, Computer Science and Engineering, Eastern Academy Of Science and Technology, Phulnakhara, Bhubaneswar, India.

Rajendra Kumar Behera, Computer Science and Engineering, Eastern Academy of Science and Technology, Phulnakhara, Bhubaneswar, India.

Kamalakanta Shaw, Computer Science and Engineering, Eastern Academy of Science and Technology, Phulnakhara, Bhubaneswar, India.

This corpus is sufficiently large to analyze the emotions in view of speaker.IITKGP-SEHSC (Indian Institute of Technology Kharagpur Simulated Hindi Emotional Speech Corpus) is a Hindi speech database recorded using 10 (5 males and 5 females) professional artists from All India Radio (AIR) Varanasi, India. The eight emotions considered for recording this database are anger, disgust, fear, happy, neutral, sadness, sarcastic and surprise. Each of the artists has to speak 15 sentences in 8 given emotions in one session. The number of sessions recorded for preparing the database is 10. The total number of utterances in the database is 12000 (15textprompts × 8emotions× 10speakers × 10sessions). Each emotion has 1500 utterances. The total duration of the database is around 7 hours.

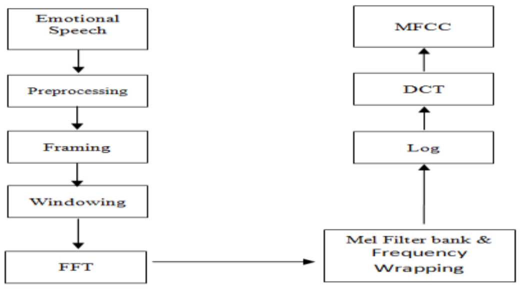
The proposed approach is using six emotion states such as Anger, Fear, Happy, Neutral, Sarcastic and Surprise of eight different speakers from this database. The data samples of speech are separated into two groups. One group for training and other group for testing. The first group is used for training the data samples and the second group is used for testing purpose. The GMM classifier is used to classify different emotions from these testing data samples. The data samples which were used for testing purpose is to be compared with the data samples which is already trained. This comparison gives the detection of emotion from these data samples. [1][3][4]

III. MEL FREQUENCY CEPSTRAL COEFFICIENT

Mel-frequency cepstral coefficients (MFCCs) are based on the known variation of the human ear`s critical bandwidth with frequency filter spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the Mel frequency scale, which is linear frequency spacing below 1000 Hz and a logarithmic spacing above the 1000 Hz. This frequency warping can allow for better representation of sound. Fig 1 shows the process to calculate Mel-frequency cepstral coefficients.

Steps to calculate MFCCs are as follows:

1. Pre-emphasize the speech signal.
2. Signal divided into sequence of frames with frame size 20 ms and frame shift 10 ms. Apply hamming window for each frame.
3. Compute magnitude spectrum for each windowed frame by applying Fourier transform.
4. Mel spectrum is computed by passing the Fourier transform signal through Mel filter bank.
5. Discrete cosine transform is applied to the log Mel frequency coefficients (log Mel spectrum) to derive the desired MFCCs.[8][9]



(Fig 1. Steps to calculate MFCC)

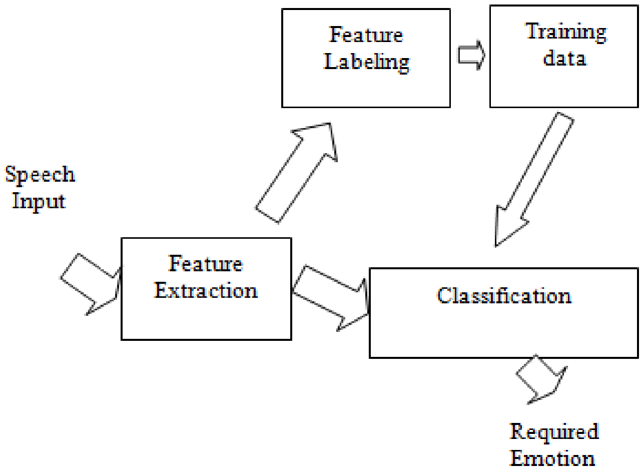
IV. GAUSSIAN MIXTURE MODEL

Gaussian mixture model is a probabilistic model for density clustering and estimation.GMMs are very efficient in modelling multi-modal distributions and their training and testing requirements are much less.GMMs cannot model temporal structure of the training data since all the training and testing equations are based on the assumption that all vectors are independent. Determining the optimum number of Gaussian components is an important but difficult problem.

In this study GMMs are used as classification tools to develop emotion recognition models. They model the probability density function of observed variables using a multivariate Gaussian mixture density. Given a series of inputs, GMM refines the weights of each distribution through expectation-maximization algorithm. Mixture models are a type of density model which comprise a number of component functions, usually Gausses. These component functions are combined to provide a multimodal density.In this work for experimental purpose we used 3 different variations of GMM 8 centred, 16 centred and 32 centred GMM. For each test case we can create the 3 model for each emotion using GMM. [6][7][10]

V. EMOTION RECOGNITION SYSTEM ARCHITECTURE

System architecture used for recognition is as shown in figure. Basically this architecture has two phases. (1)Training phase. (2)Testing phase. In training phase the models are trained for different emotion. In this first we classify the training data according to the class / emotion it belongs. During feature extraction features of speech utterance are extracted. The emotion recognition model (GMM) is trained using MFCC features vectors. After completion of training phase we have the model for each emotion. In testing phase passes testing feature vectors to all models which gives the probability value .Whichever model having highest probability value the speech utterance can classified according to that model. For example suppose for particular speech sample Anger emotion model gives highest probability value compare to other models hence we can conclude that emotion present in speech sample is Anger emotion.[7][10]



(Fig 2. Block Diagram of Emotion recognition System)

VI. EXPERIMENTAL RESULT

In this emotion recognition system, emotion recognition occurs for eight speaker data and 6 emotions for each speaker is used for both training and testing. Here training text data for each speaker is different from testing text data. Here out of 15 text prompts, 10 prompts were used for training the models and 5 prompts were used for testing purpose.Training and testing data were 67% and 33% respectively. Table 6 shows the text independent emotion recognition performance of speaker 3 from IITKGP-SEHSC (Indian Institute of Technology KharaGPur- Simulated Emotion Hindi Speech Corpus), Hindi language database.

In this experiment, we have explored different number of components 8, 16 and 32. First, we build the GMM model in basic way. Emotion recognition accuracy is observed to be 75.00 % for 32 centered GMM Model, 72.00% for 16 centered GMM model and 67.66 % for 8 centered GMM Model of speaker 3. Table-II shows the Confusion Matrix for data using 32 centered GMM model. Diagonal values of table show the correctly recognized samples. Table-III shows the Confusion Matrix for data using 16 centered GMM model. Table-IV shows the Confusion Matrix for data using 8 centered GMM model. Table V shows the average speaker depependent emotion recognition performance of all 8 speaker.

(Table I: Speaker Dependent emotion recognition performance of a speaker)

	8 Centered GMM	16 Centered GMM	32 Centered GMM
Accuracy	67.66	72.0	75.0

(Table II: Confusion matrix of Speaker Dependent emotion recognition performance for 32-centered GMM)

	Anger	Fear	Happy	Neutral	Sarcastic	Surprise
Anger	70	2	24	2	2	0
Fear	0	100	0	0	0	0
Happy	14	0	72	0	14	0
Neutral	12	10	2	64	10	2
Sarca stic	6	0	14	4	74	2
Surpri se	16	0	2	0	12	70
	Average percentage of accuracy=75.00					

(Table III: Confusion matrix Speaker Dependent emotion recognition performance for 16-centered GMM)

	Anger	Fear	Happy	Neutral	Sarcastic	Surprise
Anger	62	2	26	8	0	2
Fear	0	100	0	0	0	0
Happy	16	2	68	0	10	4
Neutral	20	10	4	52	12	2
Sarca stic	12	0	12	0	72	4
Surpri se	10	0	2	4	6	78
	Average percentage of accuracy=72.00					

(Table IV: Confusion matrix Speaker Dependent emotion recognition performance for 8-centered GMM)

	Anger	Fear	Happy	Neutral	Sarcastic	Surprise
Anger	48	2	22	20	2	6
Fear	2	94	2	2	0	0
Happy	10	0	70	2	12	6
Neutral	4	10	10	52	22	2
Sarca stic	4	0	24	6	58	8
Surpri se	6	0	0	6	4	84
	Average percentage of accuracy=67.66					

(Table V: Average Speaker dependent emotion recognition performance)

	8 Centered GMM	16 Centered GMM	32 Centered GMM
Accuracy	65.96	70.43	73.68

VII. CONCLUSION

Although it is difficult to get a accurate result, but we can show the variations that occur when emotion changes.MFCC features of speech sample used to extract speech feature for the recognition. Emotional speech database of Hindi language were used for experiment. Performance of the emotion recognition is depending on the speaker, emotion and language used for recognition. We use GMM to classify six different emotions as: Anger, Fear, Happy, Neutral, Sarcastic and Surprise of eight different speakers from database. Three type of GMM model used namely 8 centred GMM, 16 centred GMM, and 32 centred GMM. It is observed that, the average recognition accuracy is observed to be 75.00% for 32 centered GMM, 72.00% for 16 centered GMM and 66.67% for 8 centered GMM,when the training and testing data were 67% and 33% respectively.

REFERENCES

1. Shashidhar G. Koolagudi, K. Sreenivasa Rao ," *Emotion Recognition from Speech using Source,System, and Prosodic Features*" , November 2011.
2. D. Ververidis and C. Kotropoulos, "*Emotional speech recognition: Resources,features, and methods*," SPC, vol. 48, p. 11621181, 2006.
3. S. Koolagudi, R. Reddy, J. Yadav, and K. Sreenivasa Rao, "*IITKGP-SEHSC : Hindi speech corpus for emotion analysis*," in International Conference on Devices and Communications (ICDeCom), pp. 1 –5, Feb. 2011.
4. S. G. Koolagudi, S. Maity, V. A. Kumar, S. Chakrabarti, and K. Sreenivasa. Rao, *"IITKGP-SESC: Speech database for emotion analysis," Springer-Verlag Berlin Heidelberg*, vol. 40, pp. 485–492, 2009.
5. L. R. Rabiner and B. H. Juang, *"Fundamentals of Speech Recognition"*. Englewood Cliffs,New Jersy: Prentice-Hall, 1993.
6. Douglas Reynolds, "Gaussian *Mixture Models"*, MIT Lincoln Laboratory, 244 St Wood, . Emotion Recognition Using Support Vector Regression" 10th International Society for Mus ic Information Retrieval Conference (ISMIR 2009).
7. Xianglin Cheng, Qiong Duan," *Speech Emotion Recognition Using Gaussian Mixture Model*", The 2nd International Conference on Computer Application and System Modeling (2012).
8. Bhoomika Panda, Debananda Padh, Kshamamayee Dash, Prof. Sanghamitra Mohanty "*Use of SVM Classifier & MFCC in Speech Emotion Recognition System"* ,IJARCSSE-Volume 2,Issue-3,M.arch-2012,ISSN:2277128X..
9. Jesus Olivares-Mercado, Gualberto Aguilar, Karina Toscano-Medina, Mariko Nakano and Hector Perez Meana , "*GMM vs SVM for Face Recognition and Face verification*",(2011) Reviews, Refinements and New Ideas in Face Recognition, Dr. Peter Corcoran (Ed.), ISBN: 978-953-307-368-2.
10. Nitin Thapliyal, Gargi Amoli," *Speech based Emotion Recognition with Gaussian Mixture Model*", International Journal of Advanced Research in Computer Engineering & Technology, July2012.