I.J. Information Engineering and Electronic Business, 2015, 4, 51-57
Published Online July 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2015.04.07

Emotion Recognition System Based On Skew Gaussian Mixture Model and MFCC Coefficients

M.ChinnaRao
Research scholar JNTU, Kakinada-533003, India,
Email: chinnarao.mortha@gmail.com

A.V.S.N.Murthy
Professor of Mathematics Dept, AEC, Surampalem, India.
Email:avsnmurthy2005@gmail.com

Ch.Satyanarayana
Professors of Computer Science Dept, JNTU, Kakinada-533003, India.
Email:chsatyanarayana@yahoo.com

Abstract—Emotion recognition is an important research area in speech recognition. The features of the emotions will affect the recognition efficiency of the speech recognition systems. Various techniques are used in identifying the emotions. In this paper a novel methodology for identification of emotions generated from speech signals has been addressed. This system is proposed using Skew Gaussian mixture model. The proposed model has been experimented over a gender independent emotion database. In order to extract the features from the speech signals cepstral coefficients are used. The developed model is tested using real-time speech data set and also using the standard and data set of Berlin. This model is evaluated in the presence of noise and without noise the efficiency of the model is evaluated and is presented by using confusion matrix.

Index Terms—Emotion recognition, Skew Gaussian mixture model, Cepstral coefficients, confusion matrix, Berlin data set.

I. Introduction

A lot of useful implicit information is present in the speech signal like speaker gender, age, race and accent of speaking, etc. The interest to develop natural and effective interfaces for human machine communication applications has increased in attempting problems like dialect recognition and emotion recognition Speech consists of words spoken in a particular way. It always conceals the information about emotions that are available in the way the words are spoken. Emotions play significant role while interpreting the intrinsic behavior of a personage. It helps to identify the physical state of mind of person at particular instance of time and during a particular incident. Every individual exhibit his own emotion during a particular incident. These emotions will play a dominating step in applications ranging from BPO, Telemedia and also in police stations/emergency ambulance calls. Every presenter has his own verbalization rate by which the distinctiveness of a speaker can be established [7].

Emotion comprises one of the most basic factors with respect to the communication between humans. It would be ideal to have human emotions automatically recognized by machines, mainly for improving human machine interaction [12].

The emotion specific characteristics of the speech can be attributed to (1) characteristics of the excitation source, (2) shape of the vocal tract system, while producing different emotions,(3)supra-segmental characteristics(prosodic parameters : energy, pitch and energy), (4) linguistic information and (5) emotional behavior of the speaker. Emotion specific characteristics of vocal tract are represented by its unique shapes while producing sound units in different emotions[13].

Chung-Hsien Wu et al presented an approach to emotion recognition of affective speech based on multiple classifiers using acoustic prosodic information (AP) and semantic labels (SLs). For AP-based recognition, acoustic and prosodic features including spectrum, formant, and pitch-related features are extracted from the detected emotional salient segments of the input speech. Three types of models, GMMs, SVMs, and MLPs, are adopted as the base-level classifiers[14].

Yakun Hu, Dapeng Wu, and Antonio Nucci.[15] have published a method on large population of Speaker Identification under noisy conditions was addressed. The major techniques Mel Frequency cepstral coefficients, Gaussian mixture model and universal background model are performed well for small population identification under low noise conditions and it degrades as population increases. To overcome this Fuzzy clustering based decision tree approach it uses a decision tree approach that hierarchically partitions into groups of small size and applies MFCC+GMM and FCC+GMM+UBM thereby increases the accuracy. By applying this method it increases the accuracy about 15% using the decision tree with the proposed techniques than applying directly the

above techniques .Emotion specific information is represented by spectral features such as linear prediction cepstral coefficients (LPCCs), Mel frequency cepstral coefficients (MFCCs) and their derivatives.

Many models have been accessible in the literature to identify the emotions. Most of these models are based on generative approaches and degenerative approaches among which models based on SVM, ANN, HMM are mostly focused [3][4][5][6][8]. However degenerative models are more effective than non-generative models [2]. This has given direction to carry out further research using generative model based approaches, in particular using Gaussian Mixture Models (GMM the emotion more accurately from a speech sample, the Distribution which is Asymmetric in nature will be more practical. Hence in this paper Skew Gaussian distribution is considered, the advantage of considering Skew Gaussian as is that it can handle large data sets and also GMM is a particular case of this distribution, which enables to identify the even speech samples which may exhibit symmetric nature.

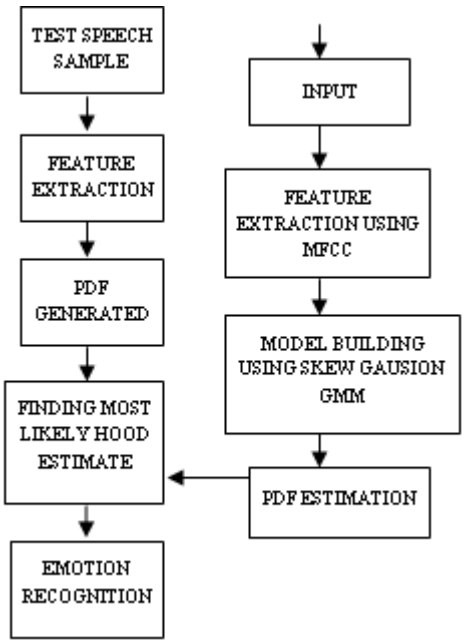

Fig. 1. Basic structure of the Recognition System

II. Feature Extraction

For effective recognition of the emotions, Feature extraction plays a dominating role. This paper is highlighted using the MFCC features, together with Formants. The spirit of considering MFCC is due to the fact that it can identify the emotions even from small sample rates also effectively.

We are proposing a set of novel acoustic features in this experiment. Most researchers use prosodic features and their statistical characteristics to classify the emotions [8][11][13][14]. In this contribution we are using the set of features listed in Table I. Among these features only Mel Frequency Cepstrum Coefficients (MFCC) and Zero Crossing Rate (ZCR) have been used for speech emotion recognition in the past [9][10][11], while the rest are being used for the first time in this application. All the features are extracted from each frame and then the mean and standard deviation for each feature is considered to constitute the feature vector. C. Feature Selection The performance of a pattern recognition system highly depends on the discriminate ability of the features. Selecting the most relevant subset from the original feature set, we can increase the performance of the classifier and on the other Figure 1

A. Mel frequency cepstral coefficients (MFCC)

Human auditory system is nonlinear. The MFCC features can match to human auditory systems. The MFCC contain both time and frequency information of the speech signal and this makes them more useful for feature extraction. MFCC features have been used in the field of speech recognition widely and have managed to handle the dynamic features as they extract both linear and non-linear properties of the signal.

In order to present the ideology with proposed model, experimentation is conducted on generated data set with 200 speakers of both the genders with acted sequences of 5 different emotions, namely happy, sad, angry, boredom, neutral. In order to test the data 50 samples are considered and a database of audio voice is generated in .wav format. The performance of the developed method is compared to that of the GMM. And a model is also tested using standard emotion speech data set namely “Berlin data set”. In order to identify the emotions the features from the emotion speech sample are extracted using MFCC. Rest of the paper is presented as follows. Section 2 of the paper highlights the concepts of Feature extraction and MFCC coefficients, Skew Gaussian mixture model is presented in section 3, section 4 of the paper describes about the Gender identification using K-Means algorithm, and the section 5 highlights the methodology along with experimental results and in the concluding section 6 results derived are tabulated.

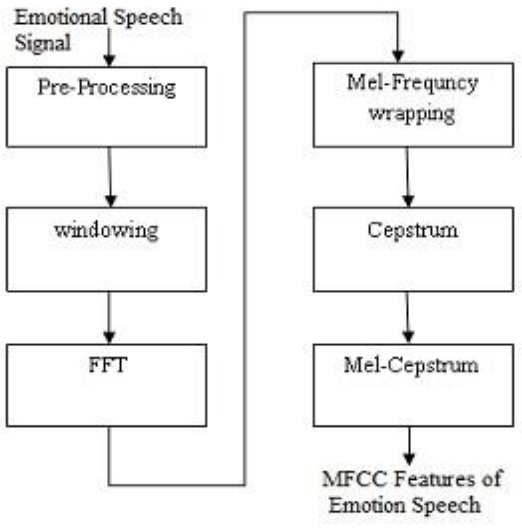

Fig. 2. MFCC feature extraction

Usually the emotion speech signals are measured in Mel-Scale rather than using the linear scale, which is computed using Mel (f) = $2595*\log_{10}$ (1+f/700)The subsequently step is to compute the Mel frequency cepstral coefficients, where the log Mel spectrum coefficients are transformed to time domain via the discrete cosine transform (DCT). The MFCC [9] is the progression of converting back the log Mel spectrum into time frequency.

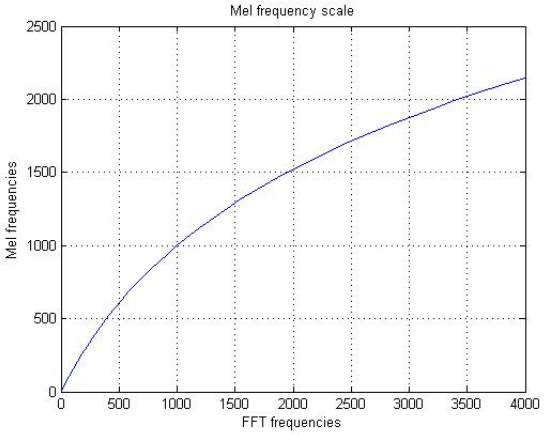

Fig. 3. Mel-frequency Scale

B. Formants

These are defined as the spectral peaks of the sound spectrum of the voice. The frequency components of human speech formants are represented with F1, F2, F3 [10]. The arranging of formants starting from increasing order with low frequency F1 to high frequency F3. F1 and F2 are the distinguish vowels. The two determine the quality of vowels open or close front or back. F1 is assigned higher frequency for 'a' and lower frequency for close vowel 'i' and 'u'. F2 is assigned as higher frequency for front vowel 'i' and lower frequency for back vowel 'u'.

III. Skew Gaussian Distribution

Gaussian mixture models can perform a difficult task of speaker identification. Gaussian mixture model maintains high identification performance for increasing population. Mixture models and their typical parameter estimation methods can approximate variety of probability density functions From a practical point of view it is often sound to form the mixture using one predefined distribution type, a basic distribution. Generally the distribution function can be of any type, but the multivariate normal distribution, the Gaussian distribution, is undoubtedly one of the most well-known and useful distributions in statistics, playing a predominant role in many areas of applications [11]. For instance, in multivariate analysis most of the existing inference procedures have been developed under the assumption of normality and in linear model problems the error vector is often assumed to be normally distributed. In addition to appearing in these areas, the multivariate normal distribution also appears in multiple comparisons, in the studies of dependence of random variables, and in many other related fields. So, if there no prior knowledge of a probability density function of phenomenon, only a general model can be used and the Gaussian distribution is a good candidate due to the enormous research effort in the past.

Skew Gaussian mixture model is an asymmetric model which belongs to a class of Gaussian mixture models and the main advantage of this model is that GMM is its particular case.

The probability density function of the Skew Gaussian mixture model is given by

$$f(z) = 2.\phi(z).\Phi(\alpha z); \quad -\infty < z < \infty \tag{1}$$

$$\text{Where, } \Phi(\alpha z) = \int_{-\infty}^{\alpha z} \phi(t)dt \tag{2}$$

$$\text{And, } \phi(z) = \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}} \tag{3}$$

Let, $y = \mu + \sigma z$

$$z = \frac{y-\mu}{\sigma} \tag{4}$$

Substituting equations (2),(3), and (4) in equation (1),

$$f(z) = \sqrt{\frac{2}{\pi}}.e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}\left[\int_{-\infty}^{\alpha\left(\frac{y-\mu}{\sigma}\right)} \frac{e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}}{\sqrt{2\pi}}dt\right] \tag{5}$$

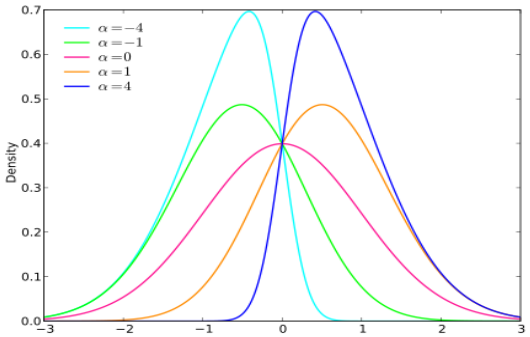

Fig. 4. Frequency curves of Skew Normal Distributions

IV. Gender Identification Using K- Means Clustering

The important modules considered in this paper are

1. Segmenting the speech into frames of Voice
2. Extraction of features from different emotion
3. Classification of emotion using Right Truncated GMM
4. Determining the accuracy of emotion

The emotion features like happy, sad, angry, neutral, boredom are extracted from the speech samples and are trained using Right Truncated Gaussian Mixture model. The feature extraction steps include, generating the emotion samples in .wav format and converting these into amplitude values, after transforming these signals into amplitude sequence, we get the values for the emotions like, happy, sad, angry, neutral, boredom. Using these amplitude values, the Probability Density Function (PDF) values of the Skew Gaussian mixture are generated, the test signal is considered and the PDF values of the test signals are classified to ascertain the emotion.

In order to extract the feature vectors from the database, K-Means clustering algorithm is used. The most important aspect considered for any speaker identification is Gender identification. Unsupervised machine learning algorithm, such as K-Means algorithm is preferred, due to the fact that no gender information is available in the database. The dataset is clustered basing on the speech sample size for both male &female speech samples. Two different cancroids are considered, for male and female. Based on the distance between the pitch values and each of the cancroids (μ) the male or the female data is classified. The new mean μ_c of each cluster C_c is calculated by using equation (6).

$$\mu_c = \frac{\sum x_i C_c x_i}{[C_c]} \tag{6}$$

Where x_i is the pitch value of the i[th] sample [6], the process is applied iteratively until cluster convergence is attained. Once the training process of k-means clustering is completed, the classification and identification is carried out by using feature vector.

V. Experimental Evaluation

To demonstrate our method we have used a database with 200 different speakers with different dialects, having five different emotions namely Happy, Sad, Boredom, Neutral and Angry.

The algorithm for our model is given as

Phase -1: Extract the MFCC coefficients.

Phase -2: cluster the data with different sample sizes.

Phase -3: train the data by PDF of Skew Gaussian Distribution

Consider the test emotion and follow steps 2 and 3. The speeches are recorded each containing of 30 sec for training and minimum of one sec of data for testing.

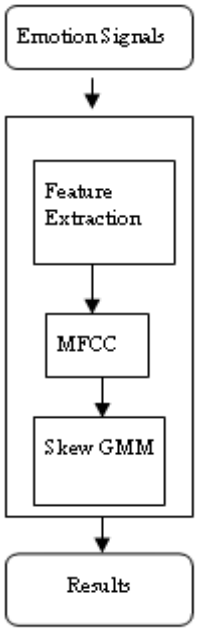

Fig. 5. The Emotion Recognition process Model

VI. Results

After extracting the emotion features and training the features using Skew Gaussian distribution, the results obtained are stored in the database in Excel format. The emotion speech signal to be tested is trained as specified in section-4 of the paper, and the obtained features are compared with the existing emotions, based on MFCC coefficients. The features of the test emotion are classified using Right Truncated Gaussian distribution using the emotions in the database and the results obtained are tabulated using a confusion matrix and are presented in Table-1 and Table-2 and Barchart 1 &2. The developed model is also compared using the Standard data Set of Berlin and the results obtained are presented in Table-3 and Table-4

Table 1. Comparison of Confusion Matrix for identify different Emotions of Male

stimulation	**Recognition Emotion (%)/proposed model**					**Recognition Emotion (%)/GMM**				
	Angry	**Boredom**	**Happy**	**Sadness**	**Neutral**	**Angry**	**Boredom**	**Happy**	**Sadness**	**Neutral**
Angry	90	10	0	0	0	80	0	10	10	0
Boredom	8	82	0	10	0	10	70	0	20	0
Happy	0	0	90	0	10	10	10	70	0	10
Sadness	0	10	10	80	0	10	0	10	60	20
Neutral	0	10	0	10	80	0	10	0	20	70

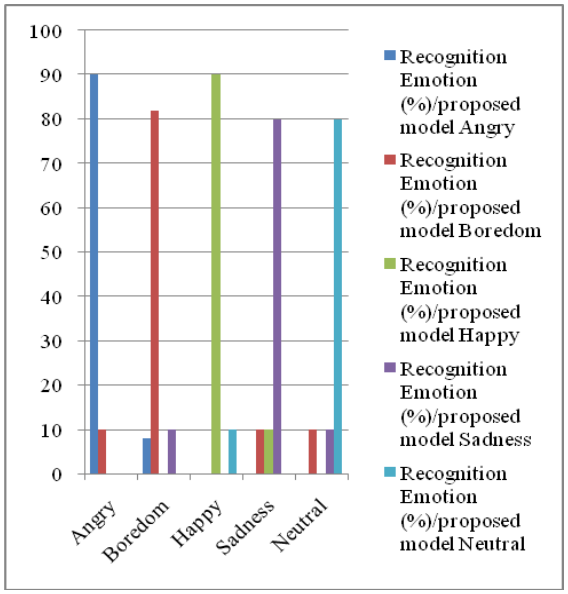

Fig. 6. Representing the recognition rates from Male database

Table 2. Confusion Matrix for identify different emotions of Female

stimulation	Recognition Emotion (%)/proposed model					Recognition Emotion (%)/GMM				
	Angry	Boredom	Happy	Sadness	Neutral	Angry	Boredom	Happy	Sadness	Neutral
Angry	85	0	10	5	0	92	0	8	0	0
Boredom	0	80	10	10	0	10	70	20	0	0
Happy	0	10	80	0	10	10	0	82	0	08
Sadness	0	0	10	90	0	10	10	0	70	10
Neutral	0	8	10	0	82	10	0	10	0	80

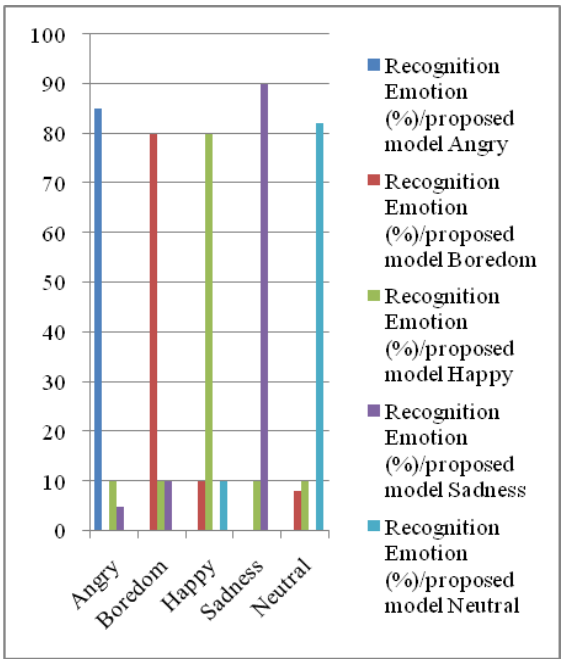

Fig. 7. Representing the recognition rates from Female database

Table 3. Comparison of Confusion Matrix for identify different Emotions of Male using BERLIN Data Set

stimulation	Recognition Emotion (%)/proposed model					Recognition Emotion (%)/Berlin				
	Angry	Boredom	Happy	Sadness	Neutral	Angry	Boredom	Happy	Sadness	Neutral
Angry	80	10	10	0	0	90	0	5	5	0
Boredom	6	84	0	10	0	5	80	0	15	0
Happy	0	2	88	0	10	0	0	90	0	10
Sadness	0	10	10	78	2	9	0	8	83	0
Neutral	0	10	0	12	78	0	10	0	7	83

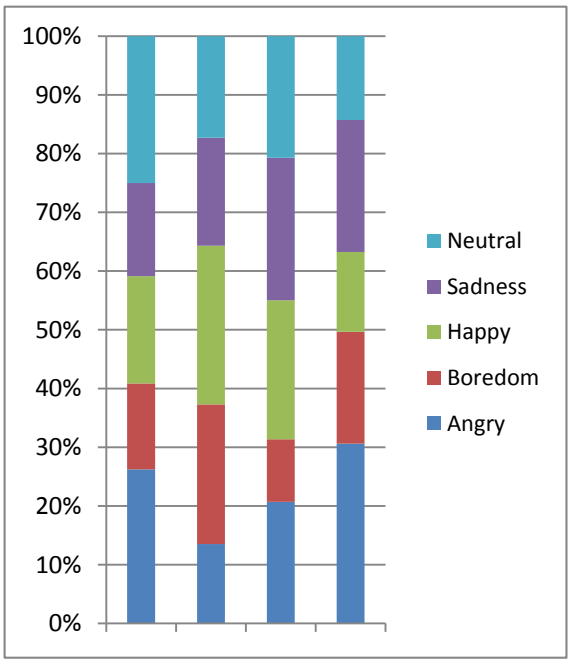

Fig. 8. Representing the recognition rates from Male database

Table 4. Confusion Matrix for identify different emotions of Female using BERLIN Data Set

stimulation	Recognition Emotion (%)/proposed model					Recognition Emotion (%)/GMM				
	Angry	Boredom	Happy	Sadness	Neutral	Angry	Boredom	Happy	Sadness	Neutral
Angry	79	0	10	10	0	88	0	12	0	0
Boredom	0	75	5	20	0	10	88	2	0	0
Happy	0	10	80	0	10	10	0	82	0	08
Sadness	0	0	20	80	0	10	5	0	85	0
Neutral	0	8	10	0	82	10	0	10	0	80

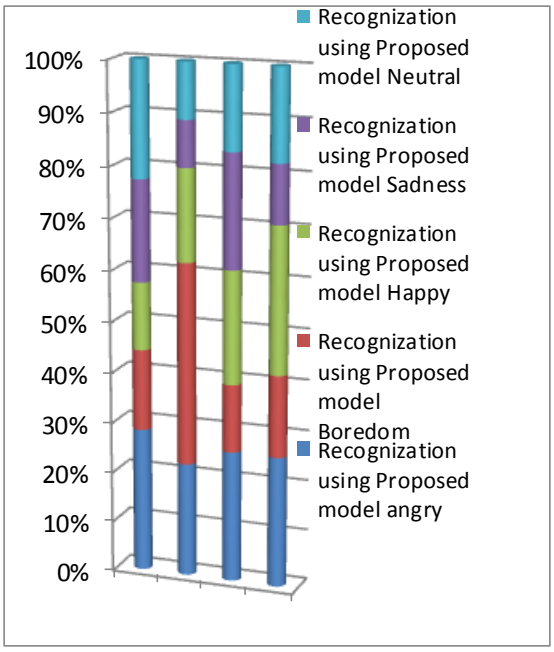

Fig. 9. Representing the recognition rates from Female database

VII. Conclusion

In this paper a novel methodology for emotion recognition is using Right Truncated Gaussian Distribution is developed. The emotions were considered from the students of Kakinada Institute of Engineering & Technology College with different dialects. These emotions are recorded at 30 ms with five different emotions. The speech database is generated from the acting sequence of one short emotionally based speech sentence comprising of 5 different emotions from 200 students (speakers) from different parts of India. The features are extracted and for recognizing, the test speaker's emotion is considered and classified using Right Truncated Distribution. The results obtained are presented in the confusion matrix for both genders in Table-1 and in Table -2, and bargraphs-1 &2, from the above tables and graphs, it can be see that the recognition rate is 88%.in case of certain emotion and for the other emotion, the recognition rate is almost 80%. The developed method is also tested by varying the sample sizes. The output is compared with that of the existing model based on GMM and from the Table-1 & Table -2, it can be clearly seen that our method outperforms the existing model. The overall emotion rate is above 80% .The results are compared with that of the standard dataset of BERLIN and the developed model performs in par with the results of the Standard data set. This shows that, the developed model performs well in identifying the emotion.

Acknowledgment

Authors wish to thankful to Dr.Y.Srinivas R&D Director of GITAM University, Vishakhapatnam and KIET Organization, Korangi, India for thier valuable ideas supported to completion of this paper.

References

[1] Arvid C. Johnson, *"Characteristics and Tables of The Left-Truncated Normal Distribution"* International *Journal* of Advanced Computer Science and Applications (IJACSA), pp133-139, May 2001.

[2] Forsyth M. and Jack M., *"Discriminating Semi-continuous HMM for Speaker Verification"* IEEE Int.conf.Acoust., speech and signal processing, Vol.1, pp313-316, 1994.

[3] Forsyth M., "Discrimination observation probability hmm for speaker verification, speech communication", Vol.17, pp.117-129, 1995.

[4] George A and Constantine K *"Phonemic Segmentation Using the Generalized Gamma Distribution and Small Sample Bayesian Information Criterion, speech communication"* DOI: 10.1016/j.specom.2007.06.005, June-2007.

[5] Gregor D et al *"Emotion Recognition in Borderline Personality Disorder- A review of the literature"* journal of personality disorders, 23(1), pp6-9, 2009.

[6] 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.

[7] Meena K, Subramanian U, and Muthusamy G "*Gender Classification in Speech Recognition using Fuzzy Logic and Neural Network*" The International Arab Journal of Information Technology, Vol. 10, No. 5, September 2013, PP477-485.

[8] Prasad A., Prasad Reddy P.V.G.D., Srinivas Y. and Suvarna Kumar G *"An Emotion Recognition System based on LIBSVM from telugu rural Dialects of Andhra Pradesh"* journal of advanced research in computer engineering: An International journal ,volume 3, Number 2, july-December 2009.

[9] Vibha T "*MFCC and its application in speaker recognition*" international journal of emerging technology ISSN: 0975-8364 pp19-22.

[10] Kasiprasad Mannepalli, Panyam Narahari Sastryand V. Rajesh "*Modelling And Analysis Of Accent Based Recognition And Speaker Identification System*" ISSN 1819-6608, Dec 2014 pages 2807-2815.

[11] GSuvarna Kumar et. Al "*SPEAKER RECOGNITION USING GMM.*" International Journal of Engineering Science and Technology, Vol. 2(6), 2010, 2428-2436.

[12] Stavros Ntalampiras and Nikos Fakotakis" Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition" IEEE Transactions on affective computing, vol. 3, no. 1, january-march 2012.

[13] K.Sreenivasa Rao Hindi Dialects and Emotions using Spectral and Prosodic features of Speech" Systems, Cybernetics and Informatics Volume 9 - Number 4 .ISSN: 1690-4524.

[14] 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 ffective computing, vol. 2, no. 1, january-march 2011.

[15] Fuzzy-Clustering-Based Decision Tree Approach for Large Population Speaker Identification. Yakun Hu, Dapeng Wu, Fellow, IEEE, and Antonio Nucci. IEEE Transactions on audio, speech, and language processing, vol. 21, no. 4, april 2013.

Authors Profiles

Mortha.ChinnaRao received the B.Tech degree in Computer Science and Engineering from JNTU, Hyderabad, Andhra Pradesh, India, in 2004. And also received M.Tech, in Software Engineering from JNTU, Hyderabad, Andhra Pradesh, India, in 2008. Presently He is pursuing PhD in JNTUK, Kakinada, and Andhra Pradesh, India. He is working as Associate Professor in KIET, Kakinada. His research interests including Data mining and Emotion recognition and speech recognition in Image processing.He is member of Computer society of India and Indian Society for Technical Education.

Dr. Akella. V. S. N. Murthy is working as Professor in Mathematics at Aditya Engineering College, Adityanagar, Surampalem, Near Kakinada, and Andhra Pradesh, India. He received PhD from Andhra University in March 2007. He has 17 years of Teaching experience and more than 10 years of Research experience. He published Research papers in reputed National and International Journals.

Dr.Ch.Satyanarayana working as a professor of computer science Engineering Dept, JNT University Kakinada .He did Doctoral Degree in the field of Image processing from JNT University, Hyderabad. He has successfully guided 5 doctoral students in the fields of Image Processing, Neural Networks and Pattern Recognition while several other 25 students are being supervised by him in a wide variety of other fields like Data Mining, Medical Image Processing and Object Recognition etc. He served as an academic supervisor to more than 250 Master Degree dissertations towards the award of M.Tech Degree and degree in Master of Computer Applications. He has published more than 100 research papers in reputed International Journals. He shared his research experience more than 200 podiums like conferences, workshops, seminars and sym

How to cite this paper: M.ChinnaRao, A.V.S.N.Murthy, Ch.Satyanarayana,"Emotion Recognition System Based On Skew Gaussian Mixture Model and MFCC Coefficients", IJIEEB, vol.7, no.4, pp.51-57, 2015. DOI: 10.5815/ijieeb.2015.04.07

