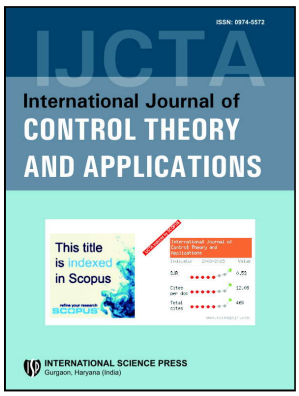

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 6 • 2017

Speaker Independent Emotion Recognition using Combined Feature and Spectral Analysis

A. Revathi[1] and C. Jeyalakshmi[2], M. Karthick[2]

[1] School of EEE SastraUniversity, Tanjore, TamilNadu, India, Email: revathidhanabal@rediffmail.com
[2] Department of Electronics and Communication Engineering K. Ramakrishnan College of Engineering, Trichy, TamilNadu, India, Emails: lakshmikrce.2016@gmail.com,winnerkarthick@gmail.com

Abstract: Affective computing has become a key issue in ensuring the better and effective human-machine interaction. In the recent years, researchers have done the emotion recognition on various databases. This paper mainly discusses the effectiveness of feature selection and spectral analysis in evaluating the performance of multi speaker independent emotion recognition system by using an iterative clustering technique. Since, EMO-DB Berlin database used in this work contains only ten speeches uttered by ten speakers in different emotions, it has become a challenging task to improve the performance of the system for the emotions such as Anger, Boredom, Disgust, Fear, Happy, Neutral and Sad. Speaker independent emotion recognition is done by creating clustering models for all emotions and evaluation is done with the speeches of a speaker not considered for training. The emotion recognition system is also evaluated for the combination of perceptual feature with probability feature and formants. The proposed features are captured and trained using clustering technique and testing is done on clean test speeches with and without spectral analysis by using minimum distance criterion. The perceptual features, perceptual features with probability and formant frequencies provide complimentary evidence in assessing the performance of the system. This algorithm provides 73% and 65% as overall weighted accuracy recall for using the combination of these features with and without spectral analysis. It also provides 80% and 68% as weighted accuracy recall for the feature combination with and without spectral analysis if the two group classification is done followed by emotion classification ensuring the increase in accuracy by 12%. Further increase in accuracy is by 5%, if the set of emotions is divided into three group models and spectral analysis is applied as additional preprocessing technique.

Keywords: Emotion *recognition, Vector quantization (VQ), Mel frequency perceptual linear predictive cepstrum (MFPLPC), Probability, Formants, Spectral analysis, Short-time energy, Zero crossing rate.*

1. INTRODUCTION

The speech signal contains the information regarding age, gender, social status, accent and emotional state of a speaker in addition to the linguistic information. It has become a challenging task to recognize the type of emotion from the database containing the same set of speeches uttered by same set of speakers. Each speaker expresses different emotions in different ways. Speech recognition on emotional speeches has found applications in call centers. People working in call centers may not behave in same manner when attending calls of the customers. When a customer experiences a negative emotion, the system has to adjust itself to the needs of the

customer or pass the control to the human agents for giving alternate convenient reply to the customers. It also has found applications in controlling the hazardous processes where physical presence of humans is not possible. These systems can also be applied in health care systems for which treatments could be extended to the patients with depression and anxiety. It also finds applications in web interactive services, information retrieval, medical analysis and text to speech synthesis. These systems would find applications in human-robot interaction where robots will behave according to the emotional state of the operator. Tin Lay New et.al [1] have used short-time log frequency power coefficient as a feature and discrete HMM as a classifier in evaluating the performance of the emotion recognition system. Donn Morrison et.al [2] have compared accuracy of emotion recognition system evaluated by using different classification techniques. Modulation spectral feature is used as a new feature by Siging Wu et.al [3] for emotion recognition. Chi-Chun Lee et.al [4] have used hierarchical binary classifier and acoustic & statistical feature for emotion recognition. Thurid Vogt et.al [5] have used combination of pitch, energy and MFCC as feature for emotion recognition and they have done gender detection. K. SreenivasaRao et.al [6] have used MFCC and GMM for recognizing emotions. AnkurSapra et.al [7] has used modified MFCC feature and NN classifier for emotion recognition. Speaker identification in emotional environment has been done by Ismail Sahini [8] and he has used log frequency power coefficients as feature and evaluated the system using HMM, CHMM and SPHMM. Shashidar G. Koolakudi et.al [9] have used MFCC and GMM for speaker recognition in emotional environment. In this paper, iterative clustering technique is used to generate the models for each emotion. For evaluating the performance of the system, features of the speech of the test emotion is applied to the individual emotion models and emotion classification is done. System is also evaluated by performing spectral analysis on the speeches of test emotion and features are extracted. These features are applied to the individual training models and improvement in the accuracy of the emotion classification is noticed. Performance has been substantially enhanced by doing two and three group classification and then respective emotion classification with spectral analysis as additional preprocessing technique .

2. FEATURE BASED ON CEPSTRUM

The short-time speech spectrum for voiced speech sound has two components: 1) harmonic peaks due to the periodicity of voiced speech 2) glottal pulse shape. The excitation source decides the periodicity of voiced speech. It reflects the characteristics of speaker. The spectral envelope is shaped by formants which reflect the resonances of vocal tract. The variations among speakers are indicated by formant locations and bandwidth .

2.1. MFPLPC Extraction

PLP (perceptual linear predictive cepstrum) speech analysis method [10-12] is for modeling the speech auditory spectrum by the spectrum of low order all pole model. This perceptual feature mainly emphasizes the need for critical band analysis which integrates the energy spectral density in the frequency range (0-8) kHz to get the speech auditory spectrum. Loudness equalization is done to emphasize the spectrum in the upper and middle frequencies and cube root compression is performed to reduce the dynamics of the speech spectrum. Then inverse fast Fourier transform is done to get the signal in time domain. Autocorrelation method is used to find linear prediction coefficients. These prediction coefficients are converted into cepstral coefficients by using recursive procedure. Critical band analysis is done using 47 critical bands, when the frequencies are spaced in mel scale. The relationship between frequency in Mel and frequency in Hz is specified as in (1)

$$f(mel) = 2595 * \log\left(1 + f(Hz)/700\right) \tag{1}$$

3. EMOTION RECOGNITION BASED ON ITERATIVE CLUSTERING TECHNIQUE

Emotional speech database considered in this work is a Berlin database which contains about 500 utterances spoken by actors in happy, angry, anxious, fearful, bored and disgusted way as well as in a neutral version.

Utterances are chosen from 10 different actors and ten different texts. Ten emotional utterances are collected from five male and female speakers respectively in the age ranging from 21 to 35 years. They are required to utter ten different utterances in Berlin in seven different emotions such as anger, boredom, disgust, fear, happy, neutral and sad. Since the database contains only ten sentences uttered by ten speakers in different emotions, it has become more challenging to implement techniques for improving the accuracy of the system. For creating a training model, speech signal is first passing through the silence and low energy frames removal block followed by pre-emphasis block for spectral flattening the signal. Hamming window is subsequently applied on differenced speech frames of 16 msecs duration with overlapping of 8 msecs to reduce the signal discontinuities at the beginning and end of the frame. Then feature vectors are extracted. Feature vectors are applied to develop the set of clusters for each emotion. During training, set of ten utterances uttered by nine speakers are used, where each utterance constitutes an observation sequence of some appropriate spectral or temporal representation. Utterances of tenth speaker in the respective emotion have been used for testing. In order to improve the accuracy of the system, spectral analysis is done to improve the strength of the frames with high spectral energy before extracting the features. Accuracy can be improved by creating group models for arousal and soft emotions separately. If the test speech corresponds to the soft or arousal emotion, group is correctly identified and testing is done with training models of the emotions corresponding to that group. Further enhancement in accuracy is ensured by using three group models.

3.1. Experimental Analysis Based on Clustering Technique

The way in which L training vectors can be clustered into a set of M code book vectors is by K-means clustering algorithm [13]. Classification procedure for arbitrary spectral analysis vectors that chooses the codebook vector is by computing Euclidean distance between each of the test vectors and M cluster centroids. Clusters are formed in such a way that they capture the characteristics of the training data distribution. Minimum distances are extracted for each test vectors and test speech is classified corresponding to the model which produces minimum of average of minimum distances.

4. CHARACTERISTICS OF EMOTIONAL SPEECH – FREQUENCY DOMAIN ANALYSIS

The semantic part of the speech contains linguistic information which reveals the characteristics of the pronunciation of the utterances based on the rules of the language. Speeches of the emotions such as anger, fear and happy are displaying the psychological behavior of the speaker such as high blood pressure and high heart rate. These speeches are loud, fast and enunciated with strong high frequency energy. On the other hand, speeches of the emotion sad reveal the characteristics of the speaker such as low blood pressure and low heart rate. These speeches are slow, low volume and possess little high frequency energy. Frequency analysis is done on the emotions of the speaker uttering the same sentence. It is found that emotions such as anger, fear and happy have more number of frames with high frequency energy and the emotion sadness has very few frames with high frequency energy. From the plot shown in Fig.1, it is indicated that emotions such as anger, fear and happy have more number of frames with high frequency energy and the emotion sadness has very few frames with high frequency energy.

5. RESULTS AND DISCUSSION

The performance of emotion recognition system based on perceptual features is evaluated by applying test speech vectors to the training models corresponding to the emotions. MFPLPC feature extraction is dealt in many literatures [10-12]. Block diagram of a parallel group classifier and parallel specific emotional pattern classifier is shown in Fig. 2.

Feature vectors of the speech of the test emotion are applied to the group models and group is identified based on the comparison with reference to the minimum of average of minimum distances. Then, subsequently

testing is done with models of emotions corresponding to the group. Speech of the test emotion is identified by first computing average of minimum distances for all the models corresponding to the group and classification is done with model pertinent to the emotions in a group.

Classification is done based on minimum distance for clustering approach and the emotion recognition accuracy is the number of correct choices over the total number of test speech segments considered for each

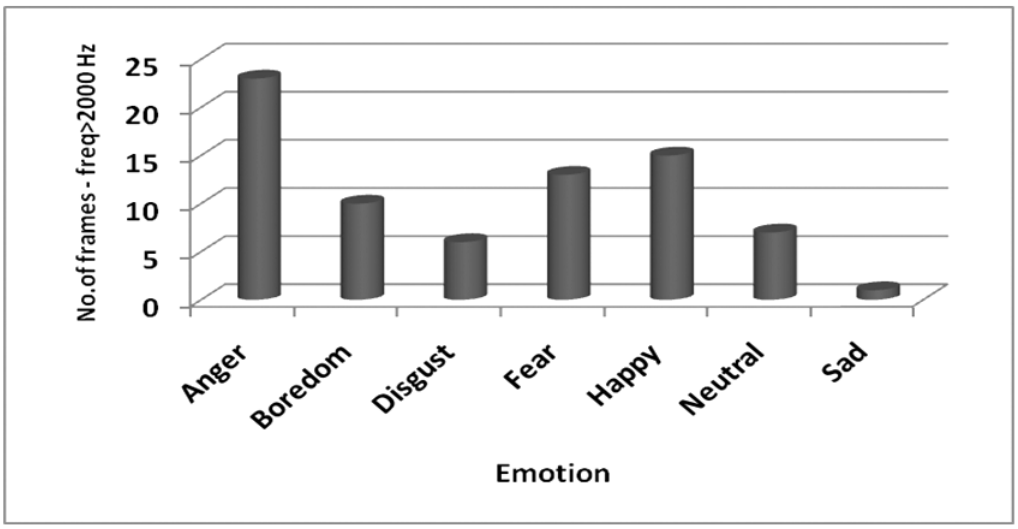

Figure 1: Frequency analysis on emotions

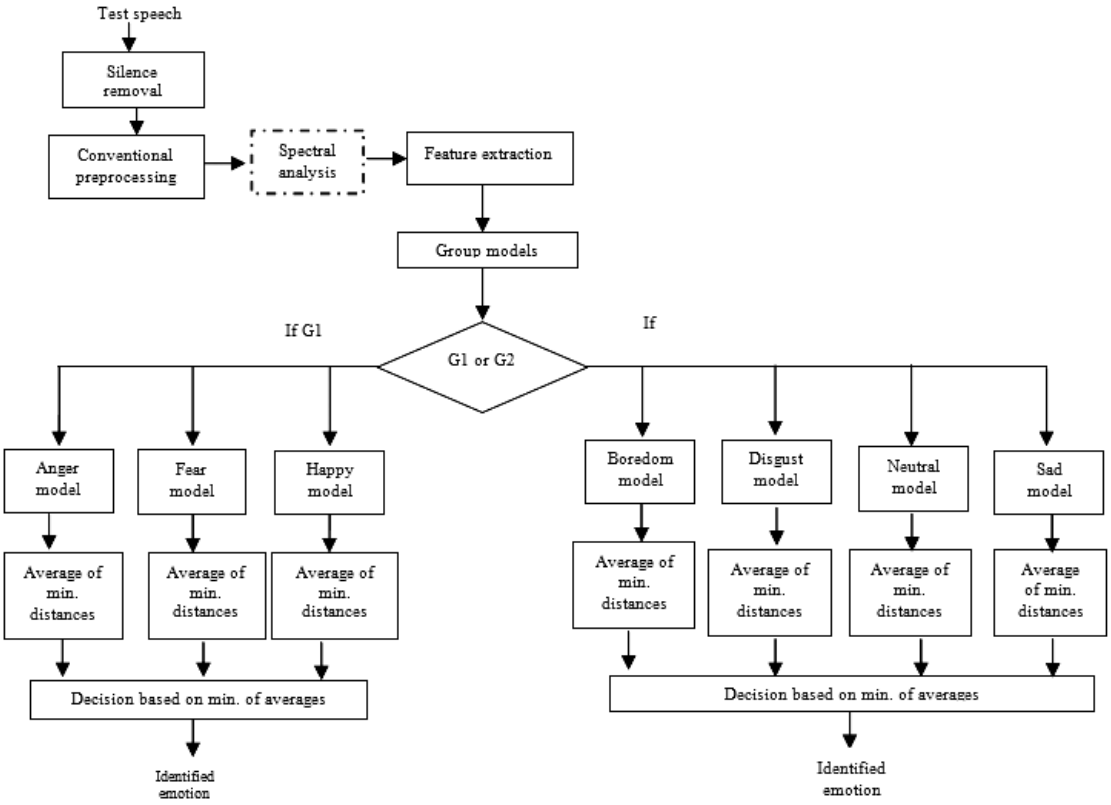

Figure 2: Parallel Group and specific emotional pattern classifier

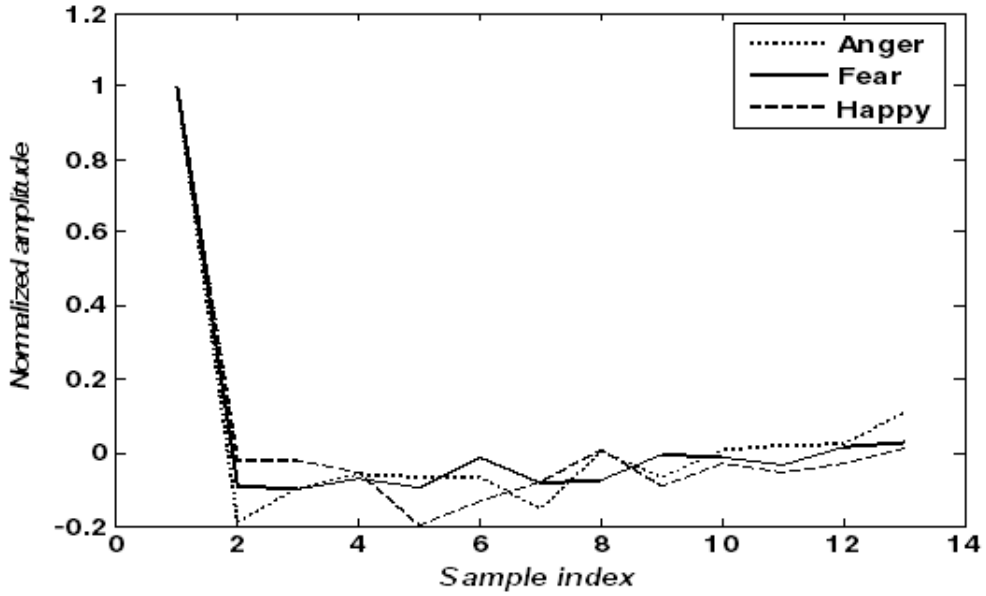

Figure 3: Feature variation – Arousal Emotions

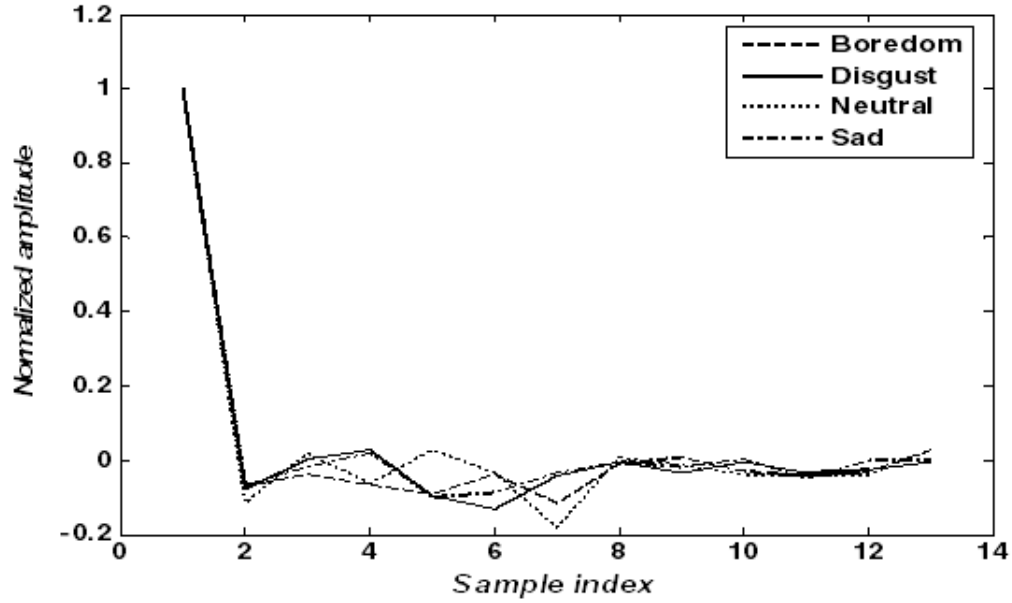

Figure 4: MFPLPC Feature variation – Soft emotions

emotion. Fig. 3 and Fig. 4 depict the nature of the feature variation for arousal and soft emotions for the same speech spoken by same speaker.

From the Figures 3 and 4, it is clear that three lines are distinct and there is less overlapping in values for Arousal emotions, whereas, the lines are more overlapping for soft emotions. Fig.5 and Fig.6 indicate the spectral variation for the speech uttered by same speaker in different emotions corresponding to the group of arousal and soft emotions.

Table. 1 gives the details of the performance of the system without spectral analysis by applying feature vectors of the test speech to the training models of individual emotions for MFPLPC, MFPLPC+Prob, Formants and the combination of all the three features.

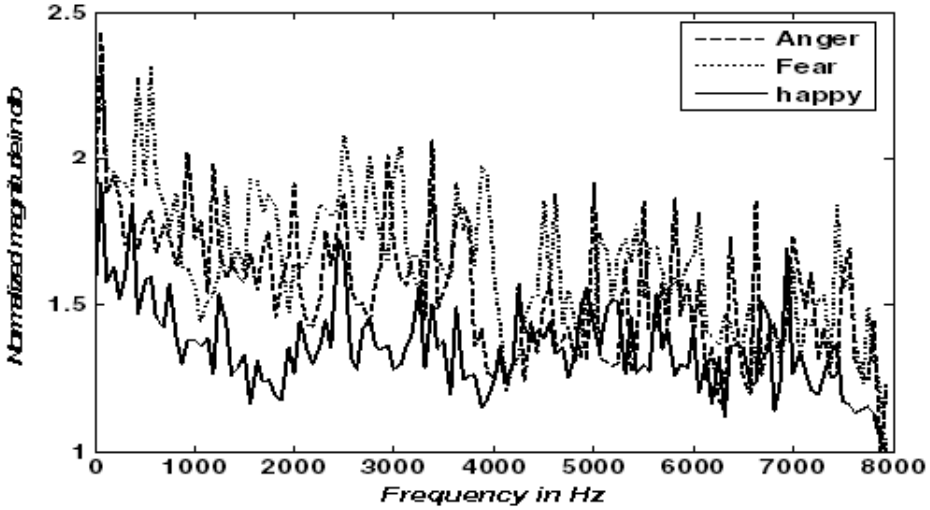

Figure 5: Spectral variation – Arousal emotions

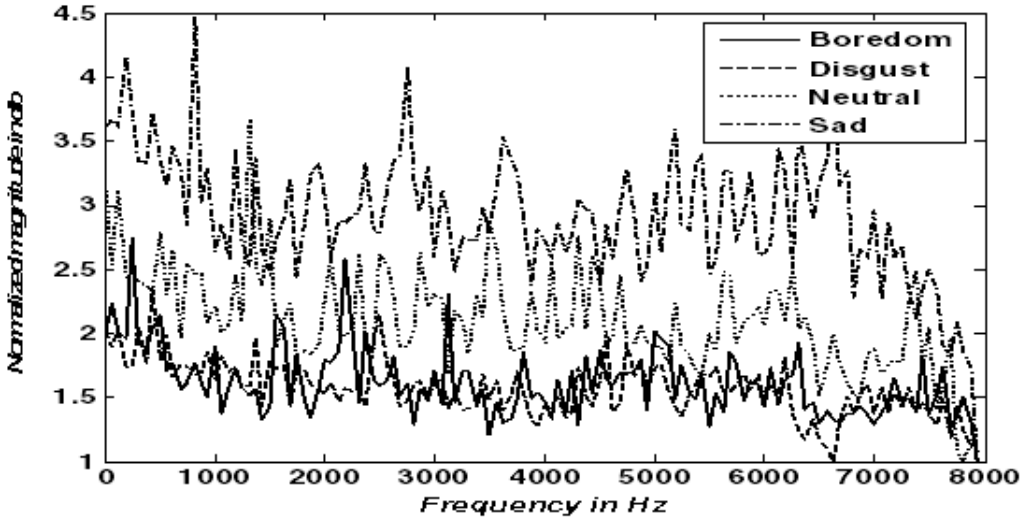

Figure 6: Spectral variation – Soft emotions

Table 1
Performance analysis of individual models without spectral analysis A-Concerned emotion and B-Other emotions

Emotions	*Classification of test segments – individual accuracy*							
	MFPLPC+PROB+FORMANTs		*MFPLPC*		*MFPLPC+PROB*		*FORMANTs*	
	A	*B*	*A*	*B*	*A*	*B*	*A*	*B*
Anger	353-100%	0	353-100%	0	345-98%	8	271-77%	82
Boredom	175-45%	218	97-25%	296	146-37%	247	17-4%	376
Disgust	290-75%	97	204-53%	183	188-49%	199	161-42%	226
Fear	12-7%	152	4-3%	160	0-0%	164	9-5%	155
Happy	122-40%	186	17-6%	291	5-2%	303	107-35%	201
Neutral	92-71%	37	52-40%	77	67-52%	62	19-15%	110
Sad	269-94%	17	248-87%	38	254-89%	32	37-13%	249
Avg. UA recall	62%		45%		47%		27%	
accu-WA recall racy	65%		48%		50%		31%	

Table 2 indicates the performance analysis of emotion recognition by applying test speeches to the individual models with spectral analysis done on test speeches.

Table 2
Performance analysis of individual models with spectral analysis
A-Concerned emotion and B-Other emotions

Emotions	*Classification of test segments – individual accuracy*							
	MFPLPC+PROB+FORMANTs		*MFPLPC*		*MFPLPC+PROB*		*FORMANTs*	
	A	*B*	*A*	*B*	*A*	*B*	*A*	*B*
Anger	353-100%	0	353-100%	0	344-98%	9	263-75%	90
Boredom	249-64%	144	151-39%	242	214-55%	179	2-1%	391
Disgust	306-79%	81	210-54%	177	200-52%	187	171-44%	216
Fear	43-26%	121	37-23%	127	27-17%	137	5-3%	159
Happy	145-47%	163	25-8%	283	10-3%	298	118-39%	190
Neutral	96-75%	33	40—31%	89	76-59%	53	34-26%	95
Sad	279-98%	7	261-91%	25	270-95%	16	92-32%	194
Avg. UA recall	70%		50%		54%		32%	
accu-WA recall racy	73%		54%		57%		34%	

From Table 1 and 2, it is evident that there is 8% increase in weighted average recall between the systems with and without spectral analysis. Table 3 shows the performance analysis on the system with respect to group classification first and then respective arousal or soft emotion classification without spectral analysis.

Table 3 Performance analysis of two group models without spectral analysis
A-Concerned emotion and B-Other emotions

Emotions	*Classification of test segments – individual accuracy*							
	MFPLPC+PROB+FORMANTs		*MFPLPC*		*MFPLPC+PROB*		*FORMANTs*	
	A	*B*	*A*	*B*	*A*	*B*	*A*	*B*
Anger	353-100%	0	353-100%	0	345-98%	8	286-81%	67
Boredom	273-70%	120	148-38%	245	177-45%	216	135-34%	258
Disgust	187-48%	200	142-37%	245	136-35%	251	39-10%	348
Fear	27-17%	137	4-3%	160	0-0%	164	25-15%	139
Happy	147-48%	161	17-6%	291	5-2%	303	136-44%	176
Neutral	99-77%	30	63-49%	66	73-57%	56	23-18%	106
Sad	275-96%	11	274-96%	12	256-90%	30	38-13%	248
Avg. UA recall	65%		47%		47%		31%	
accu-WA recall racy	68%		50%		50%		34%	

Table 4 indicates the performance analysis on the emotion recognition with group classification and subsequently the respective emotion classification with spectral analysis done as additional preprocessing to improve the strength of the frames of the signal which are significant.

Table 4
Performance analysis of two group models with spectral analysis
A-Concerned emotion and B-Other emotions

Emotions	*Classification of test segments – individual accuracy*							
	MFPLPC+PROB+FORMANTs		*MFPLPC*		*MFPLPC+PROB*		*FORMANTs*	
	A	*B*	*A*	*B*	*A*	*B*	*A*	*B*
Anger	353-100%	0	353-100%	0	344-98%	9	278-79%	75
Boredom	311-79%	82	193-49%	200	230-59%	163	108-28%	285
Disgust	350-91%	37	290-75%	97	286-74%	101	242-63%	145
Fear	57-38%	107	38-23%	126	30-18%	134	19-12%	145
Happy	157-51%	151	25-8%	283	10-3%	298	130-42%	178
Neutral	105-82%	24	57-44%	72	68-53%	61	42-33%	87
Sad	283-99%	3	283-99%	3	272-95%	14	92-32%	194
Avg. UA recall	77%		57%		57%		41%	
accu-WA recall racy	80%		61%		62%		45%	

From the tables 3 and 4, it is clear that the process of identifying a group and subsequently the respective type of soft or arousal emotion gives better results in comparison with the system on individual training models for emotions. Especially, formants perform relatively better than MFPLPC and MFPLPC+Prob. The combination of MFPLPC with probability as feature ensures better accuracy for two of the emotions namely boredom and neutral. Combination of these features performs better for the case of emotion recognition system with spectral analysis. If the test segment is correctly identified for any one of these features, it is counted as a correct classification corresponding to the particular emotion. Since the individual accuracy for the emotions Fear and Happy are relatively poor and the test segments corresponding to these emotions are misclassified with the emotion Anger. Accuracy for soft emotions is relatively better than that of arousal emotions. Achieving accuracy for the database which is containing same set of speeches uttered by same set of speakers is really challenging because of the nature of the database and also the emotional speech revealing the characteristics of the speech and speaker. Table 5 gives the details of evaluation by using three group models comprising anger, fear and happy, boredom and disgust, neutral

Table 5
Performance analysis of three group models with spectral analysis
A-Concerned emotion and B-Other emotions

Emotions	*Classification of test segments – individual accuracy*							
	MFPLPC+PROB+FORMANTs		*MFPLPC*		*MFPLPC+PROB*		*FORMANTs*	
	A	*B*	*A*	*B*	*A*	*B*	*A*	*B*
Anger	353-100%	0	353-100%	0	349-99%	4	278-79%	75
Boredom	358-91%	35	310-79%	83	337-86%	56	157-40%	236
Disgust	376-97%	11	304-79%	83	324-84%	63	289-75%	98
Fear	58-36%	106	38-23%	126	26-16%	138	19-12%	145
Happy	165-54%	143	25-8%	283	20-7%	288	130-42%	178
Neutral	129-100%	0	107-83%	12	126-98%	3	105-82%	24
Sad	284-99.3%	2	283-99%	3	280-98%	6	148-52%	138
Avg. UA recall	83%		67%		70%		55%	
accu-WA recall racy	85.3%		70%		73%		56%	

and sad and it reveals that accuracy is still improving by 5% for the combination of MFPLPC, MFPLPC+Prob and formants as compared to that of classification using two group models. MFPLPC+prob combination gives better accuracy for the emotions boredom, disgust and neutral than the basic perceptual features.

Features extracted after spectral analysis have proved to be better in achieving better performance by using F-ratio (Fisher Discriminant ratio) as a measure. F-ratio is high for the best features and this computation is done before actually applying these features to test the performance. F-ratio is calculated as per equation (2). Table 6 indicates how F-ratio is used as a measure to validate the feature selection.

$$F-ratio = \frac{Inter \quad emotion \quad \text{var}\, iability}{Intra \quad emotion \quad \text{var}\, iability} \tag{2}$$

Table 6
Validation of spectral analysis using F-ratio

Emotions	*Features*	*F-ratio*	
		Without spectral analysis	*With spectral analysis*
Anger and Boredom	MFPLPC	0.1624	0.1649
	MFPLPC+prob	0.1380	0.1395
	Formants	0.0076	0.0090

Features with high F-ratio ensure better results. From the F-ratio calculation as shown in Table 6, it is evident that features after performing spectral analysis have produced better results in comparison with the case of features without spectral analysis.

6. CONCLUSIONS

This paper proposes the use of additional preprocessing technique using spectral analysis and different feature combination such as MFPLPC, MFPLPC combined with probability and formants for evaluating multi speaker independent emotion recognition by using iterative clustering technique. Training models are developed using clustering technique and test speeches are applied to the individual models and performance is evaluated in terms of recognition accuracy. Since, perceptual based features normally perform well in developing robust speech recognition system; they are used in this work to evaluate the performance of the emotion recognition system. Probability has been computed and it is combined with perceptual features and testing is done on the training models developed for this combination. This combined feature performs well for boredom and neutral emotions as compared to the basic perceptual features. Formants provide complimentary evidence for happy emotion. Accuracy of the system is relatively good for the system implemented using the combination of these features which has spectral analysis as additional preprocessing technique. Accuracy of the system increases if the group classification is done first, and subsequently the respective emotion classification. Once the group out of two groups is correctly identified and subsequent testing is done on the respective group containing models for less number of emotions. Performance is still better if three groups are used with combination of emotions such as anger, fear and happy, boredom and disgust, neutral and sad respectively. In all the cases, performance of the system using spectral analysis as additional preprocessing technique is relatively better than the system without spectral analysis. F-ratio is computed and used as a measure to validate the efficiency of the feature selection after performing the spectral analysis as additional preprocessing technique.

REFERENCES

[1] Tin Lay Nwe , Say Wei Foo, Liyanage C. De Silva , "Speech emotion recognition using hidden Markov models" *Speech Communication*, 41, 2003, pp.603–623.

[2] Donn Morrison, Ruili Wang, Liyanage C. De Silva "Ensemble methods for spoken emotion recognition in call- centres", *Speech Communication*, 49,(2007, pp. 98–112.

[3] SiqingWua, Tiago H. Falk b, Wai-Yip Chan "Automatic speech emotion recognition using modulation spectral features", *Speech Communication* ,53, 2011, pp.768–785.

[4] Chi-Chun Lee , Emily Mower , Carlos Busso , Sungbok Lee , Shrikanth Narayanan , "Emotion recognition using a hierarchical binary decision tree approach" *Speech Communication*, 53, 2011, pp.1162–1171.

[5] ThuridVogt, Elisabeth Andr, "Improving Automatic Emotion Recognition from Speech via Gender Differentiation"*Proc. Language Resources and Evaluation Conference* (LREC 2006).

[6] K. SreenivasaRao, TummalaPavan Kumar, KusamAnusha, BathinaLeela, IngilelaBhavana and Singavarapu V.S.K. Gowtham, "Emotion Recognition from Speech", *International Journal of Computer Science and Information Technologies*, Vol. 3 (2), 2012, pp. 3603-3607.

[7] AnkurSapra, Nikhil Panwar, SohanPanwar , " Emotion Recognition from Speech",*International Journal of Emerging Technology and Advanced Engineering*, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Vol.3, Issue 2, February 2013.

[8] Ismail Shahin, "Speaker Identification in Emotional Environments"*Iranian Journal of Electrical and Computer Engineering*, Vol. 8, No. 1, Winter-Spring 2009

[9] Shashidhar G. Koolagudi, Kritika Sharma, K. SreenivasaRao "Speaker Recognition in Emotional Environment" Communications in Computer and Information Science, Vol. 305, 2012, pp.117-124.

[10] HynekHermansky, Kazuhiro Tsuga, Shozo Makino and HisashiWakita, "Perceptually based processing in automatic speech recognition", *Proc. IEEE int. conf. on Acoustics, speech and signal processing*, Tokyo, April 1986, 11, pp.1971-1974

[11] HynekHermansky, Nelson Margon, ArunaBayya and Phil Kohn, "The challenge of Inverse E: The RASTA PLP method", *Proc. Twenty fifth IEEE Asilomar conf. on signals, systems and computers*, Pacific Grove, CA, USA, November 1991,2, pp.800-804

[12] HynekHermansky and Nelson Morgan, "RASTA processing of speech",*IEEE transactions on speech and audio processing*, 1994, 2, (4), pp.578-589

[13] Rabiner.L.&Juang B.H., *Fundamentals of speech recognition*, Prentice Hall, NJ 1993.