Emotion Recognition in Speech Using MFCC and Wavelet Features

K.V.Krishna Kishore
Computer Science and Engineering
Vignan University
Andhra Pradesh, India
kishorekvk_1@yahoo.com

P.Krishna Satish
Computer Science and Engineering
Vignan University
Andhra Pradesh, India
Krishnasatish2007@gmail.com

***Abstract*—Recognition of emotions from speech is one of the most important sub domains in the field of affective computing. Six basic emotional states are considered for classification of emotions from speech in this work. In this work, features are extracted from audio characteristics of emotional speech by Mel-frequency Cepstral Coefficient (MFCC), and Subband based Cepstral Parameter (SBC) method. Further these features are classified using Gaussian Mixture Model (GMM). SAVEE audio database is used in this work for testing of Emotions. In the experimental results, SBC method out performs with 70% in recognition compared to 51% of recognition in MFCC algorithm.**

Keywords: **Mel-frequency Cepstral Coefficient (MFCC), Subband based Cepstral Parameter (SBC), Gaussian Mixture Model (GMM).**

I. INTRODUCTION

Biometrics are automated methods of recognizing a person based on a physiological or behavioral characteristic. Face, fingerprints, hand geometry, palm, iris, retinal, vein, and voice are the features used in Biometrics. Biometric technologies are becoming the foundation of an extensive array of highly secure identification and personal verification solutions. There are many applications in biometrics beyond Homeland Security such as Enterprise-wide network security infrastructure, secure electronic banking, investing and other financial transactions, retail sales, law enforcement, health and social services. Humans often use faces and voice to recognize individuals and advancements in computing capability over the past few decades enabled similar recognitions automatically.

Human speech contains not only the linguistic content but also contains some emotions of the speaker. Even though the emotion does not alter the linguistic content, it also carries some important information about the speaker and their responses to the outside world. With the increasing demand for spoken language understanding in human computer interface, automatic identification of emotional states from speech gain importance in realizing natural and effective communications between human and computers. Previous studies are mostly focused on recognizing emotions from face, while research on speech emotion recognition has not been well developed. In the area of emotion classification, many pattern recognition algorithms have been used.

Some of the approaches for speech emotion classifications are K-nearest Neighbors (KNN), Hidden markov Model (HMM), Gaussian Mixture Model (GMM), Support vector machine and artificial Neural network (ANN). HMM, with dynamic time warping capabilities, has been used for long time for speech recognition. Support Vector Machine (SVM) has also attracted attention in the speech recognition field for its high generalization capability due to its structural risk minimization oriented training. Real time speech recognition system is very different from the environment of recognition from speech database. Actual environment have problems that do not exist in the experimental database. Hence, the study of emotion recognition in noisy environments is needed. The speech signal varies for a given word both between speakers and for multiple utterances by the same speaker.

The problem of speech emotion recognition is divided into two stages. Feature extraction is first stage and classification is second stage. Although these two components appear to be independent they are highly coupled. There are several feature extraction algorithms are available for extract features from the speech signal. Most popularly used approaches are Mel-Frequency Cepstral coefficients (MFCC) [1], Linear Prediction Coding (LPC)[1], and Eigen-FFT [2]. Even though MFCC's gives good results it's not immune to noise. So Subband based cepstral (SBC) parameters are used for feature extraction in this paper, as it has embedded denosing or enhancement at feature extraction stage to improve the results of MFCC. SBC parameters are less sensitive to the noisy data when compared to MFCC. So SBC is expected to give better performance and recognition accuracy.

Different pitch features that might be useful for emotion recognition, such as the steepness of rising and falling of the pitch, and direction of the pitch contour is considered in feature extraction by Paeschke et.al.,[4]. Likewise, in [5] the differences in the global trend of the pitch were considered. Using perceptual experiments by Ladd et al., [6] proved that pitch range was more salient than pitch shape and explained by making the distinction between linguistic and paralinguistic pitch features.

Another interesting question is whether the emotional variations in the pitch contour change in terms of specific emotional categories or general activation levels. The mean and range of the pitch contour change as a function of

978-1-4673-4529-3/12/$31.00 ©2012 IEEE

emotional arousal as stated in [7], on the other hand, they did not find evidence for specific pitch shapes for different emotional categories. Song et al., [8,9] extracted pitch and energy as audio features, and used the motion of eyebrow, eyelid, and cheek as expression features, while that of lips and jaw as the visual speech ones and around 85% recognition rate was claimed. Pitch was extracted as the features and a nearest-neighbour method was used for classification in [10]. The acoustic signals were recognized by a neural network approach on statistics of low level features such as pitch, power, formants and duration of voiced segments [11].

Statistics of the pitch contour, energy envelope and their derivatives have used to represent the characteristics of emotional speech and classified six principal emotions by using the nearest mean classifier [12].

Maximum likelihood Bayes classification, kernel regression and k-NN (Nearest Neighbor) with four emotion categories are compared using 17 features by Deallert et. al.,[13]. Scherer et al., [14] extracted 16 features and achieved overall accuracy for fourteen emotional states. As a first attempt for emotion recognition in call center application by Zhou G et. al., [15] have used nonlinear Teager Energy Operator (TEO) feature for stressed/neutral classification. Then compared feature performance with the traditional pitch and MFCC feature. Angry versus Neutral speech for call center applications was focused in [16] to achieve maximum accuracy.

In every speech frame, three kinds of features- pitch, energy, MFCC with 13 coefficients, and their delta values are extracted as emotional features. The mean and standard deviation of these original features and their delta values are computed and normalized over each frame to form a total of 56-dimensional feature vectors. SHR (Sub harmonic-to-Harmonic Ratio) transforms the speech signal into the FFT domain, and then decides two candidate pitches by detecting the peak magnitude of the log spectrum. Short time energy of the speech signal provides a convenient representation that reflects amplitude variations. On the other hand, MFCC is the most widely used feature in speech recognition. Essentially, the DCT step in the calculation of MFCC features decorrelates filter bank energies. It has been shown in [17] that the wavelet transform is better decorrelater in coding applications. Gaussian mixture densities typically used to model the emotions. The degree to which this assumption is satisfied will be depending upon transformation which makes decorrelation.

Three mainstream approaches including parallel phone recognition language modeling (PPRLM), support vector machine (SVM) and the general Gaussian mixture models (GMMs) were proposed in [18] and the experimental results shows that the SVM framework achieved an equal Error rate (EER) of 4.0%, outperforming the state-of-art systems by more than 30% relative error reduction. A generalized technique by using GMM in [19] had obtained an error of 17%. A description of the major elements of MIT Lincoln Laboratory's Gaussian mixture model (GMM)-based speaker verification system built around the likelihood ratio test for verification, using simple but effective GMMs for likelihood functions, a universal background model (UBM) for alternative speaker representation, and a form of Bayesian adaptation to derive speaker models from the UBM were presented in [20].

The remainder of this paper is organized as follows: In section II, we propose an emotion recognition system using MFCC and SBC. In section III, we describe feature extraction algorithms. In section IV, we describe GMM algorithm. In section V, we show the experimental results and comparative study. We compare results of MFCC method with SBC method. Finally, in section VI we present conclusions.

II. PROPOSED ARCHITECTURE

Speech emotion recognition is one of challenging task. This is due to lack of proper definition of speech emotions [19]. Most of the existing algorithms have concentrated on only speech recognition [4], [6] or emotion recognition from facial features only. The methods or algorithms towards speech emotion recognition had not been well developed. In this work, proposing a system that deals with the speech emotion recognition and targets in improvement of results using MFCC and SBC. The proposed architecture is shown below Fig. 1.

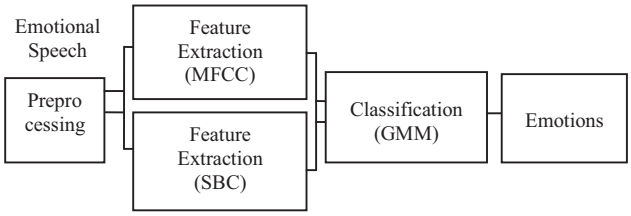

Fig 1. Architecture of emotional model

In this paper text dependent emotional speech samples are taken as inputs. The processing has been performed on training utterances. Windowing is one of the preprocessing technique which can be performed on emotional speech samples before extracting the feature. In this method feature extraction has done using MFCC and SBC. After the extraction of MFCC/SBC features, save them as feature vectors. The basic acoustic features extracted directly from the original speech signals, e.g. pitch and intensity related features, are widely used in speech emotion recognition. Some features derived from mathematical transformation of basic acoustic features, e.g. Mel-Frequency Cepstrum Coefficients (MFCC) and Linear Prediction-based Cepstral Coefficients (LPCC) [17], are also employed in some studies. By selecting some of these features modifications are applied and features which are having equal importance are used for emotion recognition. This prompts the need of feature selection in emotion recognition. Feature vectors are input to the classification algorithm. Gaussian Mixture model (GMM) is used for classification of emotions. Based on GMM parameters the emotions can be classified. GMM models the probability density function (pdf) of the data as weighted sum

of several different Gaussian density functions. Expectation Maximization (EM) algorithm is used to estimate the parameters of GMM, including probability, mean, and covariance matrix of each component. For classification, GMM is usually performed in a modular architecture, which involves a separate GMM being trained for each individual Class.

III. Feature Extraction Algorithms

The process of Speech Emotion Recognition is mainly three phases. First phase is feature extraction, second is training based on extracted features and third is classification of emotions. For Feature Extraction we use two algorithms MFCC (Mel-Frequency Cepstral Coefficients) and SBC (Subband based Cepstral coefficients).

A. MFCC Algorithm

Mel-Frequency Cepstral Coefficient (MFCC) is a popular and powerful analytical tool in the field of speech recognition. The purpose of MFCC is to mimic the behavior of human ears by applying cepstral analysis. The MFCCs are computed based on speech frames. However, the lengths of the utterances are different, and total number of coefficients extracted is different. Furthermore, with a feature vector of high dimension, the computational cost is high. Usually, in speech recognition, the total number of coefficients being used is between nine and thirteen. This is because most of the signal energy is compacted in the first few coefficients due to the properties of the cosine transform. In this work, the first 13 coefficients are considered as the useful features. From the 13 coefficients calculated mean, median, standard deviation, max, and min of coefficients as the extracted features, which produce a total number of 65 MFCC features.

The procedure to find MFCCs is mainly with the following steps. In first step, apply the Fourier Transform on input signal. In next step, map the power of the spectrum obtained in above step to the Mel scale. In next step take the logs of powers at each of the Mel frequencies of speech signal. Then take Discrete Cosine Transform on bank of Mel log powers. In this final step, we convert the log Mel spectrum back to time. The result is called the Mel frequency cepstrum coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis.

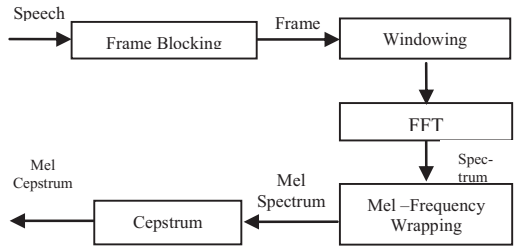

Fig 2. Block Diagram of MFCC

Because the Mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT). The algorithm divides the each speech sample into frames and computes MFCCs of each frame and stores in matrix. The coefficients represented in frames with constant sampling.

In Frame Blocking step the continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M ($M < N$). The first frame consists of the first N samples. The second frame begins M samples after the first frame, and overlaps it by $N - M$ samples. Similarly, the third frame begins $2M$ samples after the first frame (or M samples after the second frame) and overlaps it by $N - 2M$ samples. This process continues until all the speech is accounted for within one or more frames. Typical values for N and M are $N = 256$ (which is equivalent to ~ 30 msec windowing and facilitate the fast radix-2 FFT) and $M = 100$. The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame.

The next processing step is the Fast Fourier Transform, which converts each frame of N samples from the time domain into the frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT) which is defined on the set of N samples $\{x_n\}$, as follow:

$$X_n = \sum_{k=0}^{N-1} x_k \, e^{-2\pi jn/N} \quad \text{here } 0, 1, 2\ldots\ldots\ldots\text{N-1}$$

Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'Mel' scale. The Mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels. In this final step, convert the log Mel spectrum back to time. The result is called the Mel frequency cepstrum coefficients (MFCC).

The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the Mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT). The Mel spaced filter bank which is shown below.

Finally calculate the score. If the score is greater than 6.8 it is considered as a perfect match. Otherwise it is considered as poor quality.

B. SBC Algorithm

In MFCC algorithm the derivation of parameters is done in two stages. In the first stage, compute the filter bank energies and in the second stage, decorrelation of log filter bank energies with DCT has to be done to obtain the MFCC. The derivation of SBC is also similar to MFCC but deviated in the calculation of filter bank energies using wavelet packet transform rather than using Fourier transform. So this

technique gives better results than MFCC. While computing parameters of MFCC the spectrum of signal is filtered with cosine type filters to obtain the filter bank energies. The effect of filtering as a result of tracing low pass/high pass filter of wavelet packet tree. The SBC parameters are derived from Subband energies by applying the Discrete Cosine Transformation.

$$\sum_{i=1}^{l} log\ S_i\ \mathrm{Cos}\left(\frac{n(i-0.5)}{L} * \pi\right)$$

Here n=1 …….n`. Where n` is the number SBC parameters and L is total number of frequency bands. Because of similarity in root–cepstral analysis, they are called as SBC parameters. Just like in MFCC here also the score is computed. But here in this method features extracted are completely different from MFCC. In SBC method, if the score is greater than 21.5 then the sample is considered as a perfect sample.

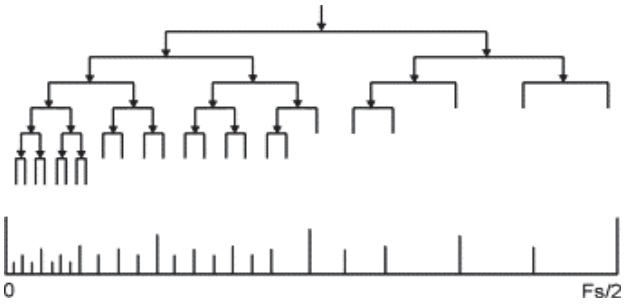

Fig 3: 24 sub band wavelet packet tree

The samples which score less than the 21.5 are considered as a poor quality. So SBC method outperforms the MFCC feature extraction method in extraction of features. The wavelet packet tree is shown below. The sampling rate taken for conversion of speech corpus is 8000 and frame size taken is 192. Next step is to calculate the energy for all the levels in subbands. After calculating energies decision on quality is considered.

IV. GMM ALGORITHM

Gaussian Mixture Model is represented as a mixture of Gaussian densities. The Gaussian mixture model is a linear combination of M Gaussian, and given by the equation,

$$p = \left(\frac{\vec{x}}{\lambda}\right) = \sum_{i=1}^{M} p_i b_{i(\vec{x})}$$

where $\vec{x}$ is a D- dimensional random vector ,$b_{i\vec{x}}$,and i= 1,.2,…….M are the component densities and p_i ,i=1,2,……..M are mixture weights. Each component density is a D-dimensional Gaussian function of the form

$$b_{i(\vec{x})} = \frac{1}{(2\pi)^{D/2}|\sum i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x}-\mu_i)^T \sum_i^{-1}(\vec{x}-\mu_i)\right\}$$

where **μ** denotes the mean vector and $\sum_i$ denotes the covariance vector matrix. The mixture weights satisfy the law of total probability, $\sum_{i=1}^{M} p_i = 1$.

V. EXPERIMENTS AND RESULTS

The SAVEE database consists of six different emotions called Anger, Disgust, Fear, Happy and Sad. Various experiments are conducted on the different emotional speech signals in SAVEE database. Each of the emotional class consists of 15 emotional speech samples. The experiments are performed on three subjects called DC, JE and JK. Each emotion class contains 15 samples, and a total of 90 samples per class available. Training and Testing of speech database is done using both the feature extraction algorithms.

Initially 5 samples are used for training in MFCC algorithm and calculated an overall accuracy about 50% for DC class and it is shown in Table 1. The performance in classification of anger emotion is 60%, disgust emotion is 10%, fear is 30%, happy is 60%, neutral is 80% and sad is 60%. By using SBC algorithm with 5 samples of training, the recognition result is about 70% for DC class and it is shown in Table 2. In that anger emotion is classified with 90%, disgust is 70%, fear is 20%, happy is 90% and neutral is 60% and sad is 90%. Similarly results for cases of 6 training samples and 7 training samples are tabulated in Tables 3,4,5, and 6.

Table 1. MFCC with 5 training samples (50%) for DC.

Emotion	Anger	Disgust	Fear	Happy	Neutral	Sad
Anger	60%	20%	20%	0%	0%	0%
Disgust	20%	10%	0%	0%	0%	70%
Fear	40%	10%	30%	0%	0%	20%
Happy	40%	0%	0%	60%	0%	0%
Neutral	10%	10%	0%	0%	80%	0%
Sad	20%	20%	0%	0%	0%	60%

Table2. SBC with 5 training samples (70%) for DC

Emotions	Anger	Disgust	Fear	Happy	Neutral	Sad
Anger	90%	0%	0%	0%	0%	10%
Disgust	0%	70%	0%	0%	0%	30%
Fear	10%	0%	20%	20%	0%	50%
Happy	0%	0%	0%	90%	0%	10%
Neutral	0%	0%	0%	0%	60%	40%
Sad	0%	10%	0%	0%	0%	90%

Table 3. MFCC with 6 training samples (51.85%) for DC

Emotions	Anger	Disgust	Fear	Happy	Neutral	Sad
Anger	77.77%	0%	11.11%	11.12%	0%	0%
Disgust	11.12%	22.22%	0%	0%	0%	66.66%
Fear	55.55%	11.11%	22.22%	11.12%	0%	0%
Happy	44.45%	0%	0%	55.55%	0%	0%

Neutral	0%	0%	0%	0%	66.66%	33.34%
Sad	0%	33.34%	0%	0%	0%	66.66%

Table 4.SBC with 6 training samples (75.92%) for DC

Emotions	**Anger**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**
Anger	88.88%	0%	0%	0%	0%	11.12%
Disgust	11.12%	44.44%	0%	0%	0%	44.44%
Fear	22.22%	11.12%	33.33%	0%	0%	33.33%
Happy	0%	0%	0%	100%	0%	0%
Neutral	0%	0%	0%	0%	100%	0%
Sad	0%	0%	0%	11.12%	0%	88.88%

Table5. MFCC with 7 training samples (58.33%) for DC

Emotions	**Anger**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**
Anger	87.5%	0%	12.5%	0%	0%	0%
Disgust	0%	25%	12.5%	0%	25%	37.5%
Fear	25%	0%	12.5%	25%	0%	37.5%
Happy	50%	0%	0%	50%	0%	0%
Neutral	12.5%	0%	0%	0%	87.5%	0%
Sad	0%	12.5%	0%	0%	0%	87.5%

Table 6.SBC with 7 training samples (79.16%) for DC

Emotions	**Anger**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**
Anger	87.5%	0%	0%	0%	0%	12.5%
Disgust	0%	50%	0%	0%	25%	25%
Fear	0%	0%	50%	25%	0%	25%
Happy	0%	0%	0%	100%	0%	0%
Neutral	0%	0%	0%	0%	100%	0%
Sad	0%	0%	0%	12.5%	0%	87.5%

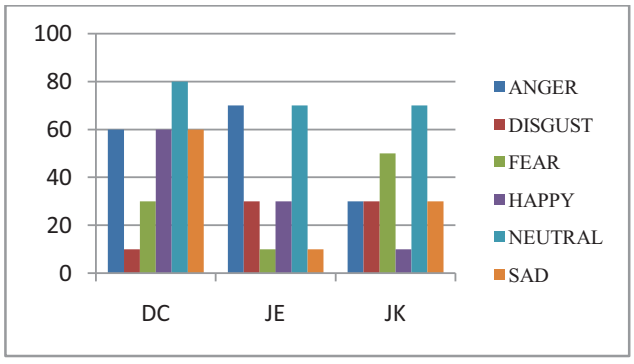

Fig . 4. DC with 50% accuracy, JE with 36.66% accuracy and JK with 36.66% accuracy for 5 training samples using MFCC.

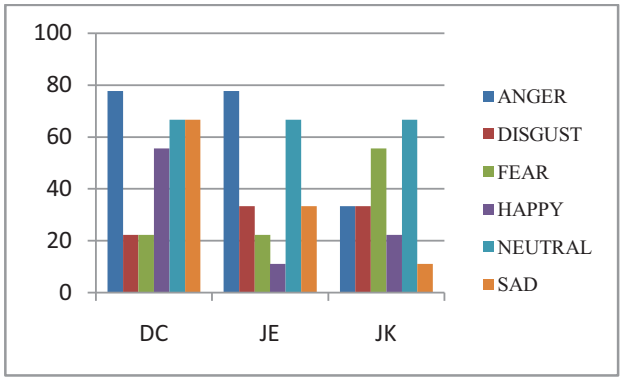

Fig. 5. DC with 51.85% accuracy, JE with 40.74% accuracy and JK with 37.03% accuracy for 6 training samples using MFCC

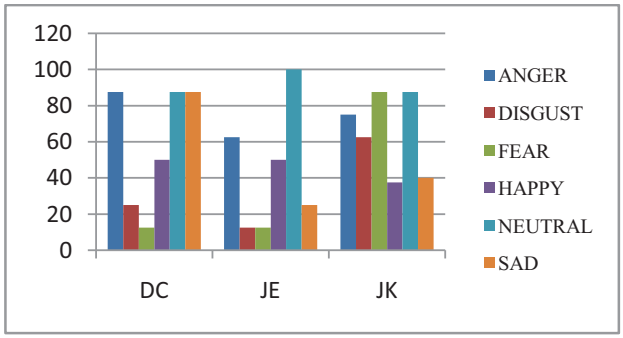

Fig. 6. DC with 58.33% accuracy, JE with 43.75% accuracy and JK with 56.25%for 7 training samples using MFCC

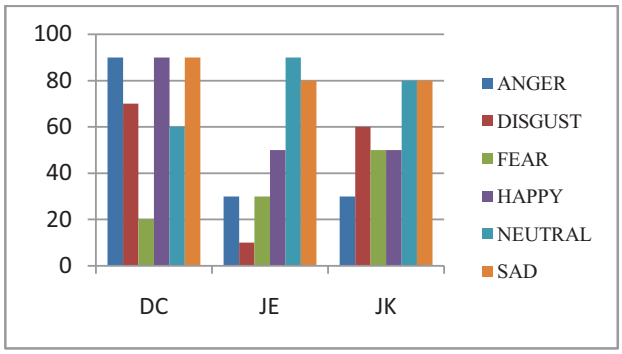

Fig. 7. DC with 70% accuracy, JE with 48.33% accuracy and JK with 58.33% for 5 training samples using SBC

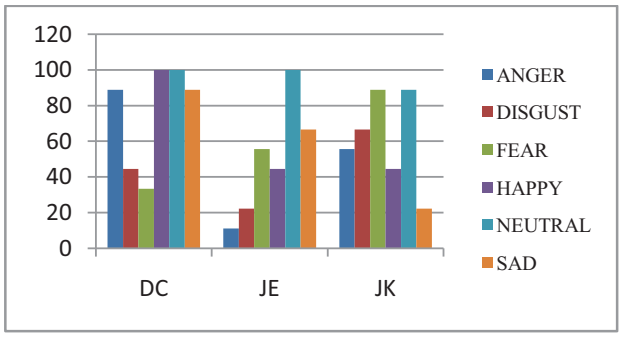

Fig. 8. DC with 75.92% accuracy, JE with 49.99% accuracy and JK with 61.10% accuracy for 6 training samples using SBC.

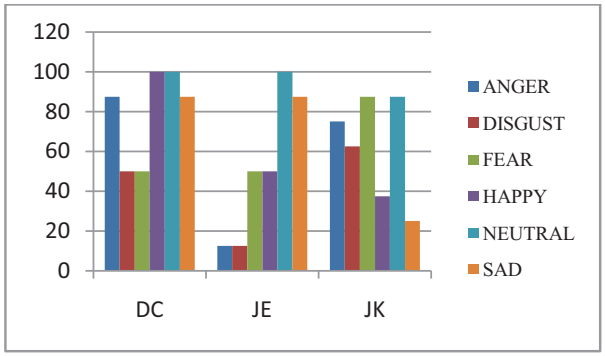

Fig. 9. DC with 79.16% accuracy, JE with 52.88% accuracy and JK with 62.5% accuracy for 7 training samples using SBC

DC, JE, and JK are three subjects which are tested using both MFCC and SBC algorithms. The System is tested with 5, 6 and 7 training samples and performance comparison is shown in Figures 4, 5, 6, 7, 8, and 9. The mean accuracy about 75% is achieved in this paper. After experimental analysis it is observed that DC subject is recognized with 79% accuracy and it outperforms the other two subjects JE and JK having SBC as a feature extraction algorithm.

VI. Conclusion

This paper describes recognition of human emotional state from speech. In the field of speech emotion recognition, the performance depends on the emotional feature. For commercial use, it is important that stable performance can be realized not only in a clean environment but also noisy environments. The experimental result shows that the performance SBC is better when compared to the MFCC method. Gaussian Mixture Model is used as classifier for recognition of emotions in this work. SAVEE database have used for performance evaluation. The performance can be improved by fusion of features extracted from multimodal such as face data and speech data and also fusion at decision level.

References

[1] Xia Mao, Lijiang Chen, Liqin Fu "Multi-Level Speech Emotion Recognition based on HMM and ANN" in 2009 World Congress on Computer Science and Information Engineering.

[2] Eun Ho Kim, Kyung Hak Hyun, Soo Hyun Kim and Yoon Keun Kwak "Speech Emotion Recognition Using Eigen-FFT in Clean and Noisy Environments" in16th IEEE International Conference on Robot & Human Interactive Communication, August 26, 2007, Korea.

[3] Ruhi Sarikaya, Bryan L. Pellom and John H.L.Hansen "Wavelet packet Transform Features With application to speaker identification".

[4] A. Paeschke, M. Kienast, and W.Sendlmeier,"F0-Contours in Emotional Speech," in Proc. 14th Int. Conf. Phonetic Sci. (ICPh'99), San Francisco,CA, Aug. 1999, pp. 929–932.

[5] A. Paeschke, "Global trend of fundamental Frequency in emotional Speech," in Proc. Speech Prosody (SP'04), Japan, Mar. 2004, pp. 671–674.

[6] K. Scherer, D. Ladd, and K. Silverman, "Vocal Cues to speaker Affect: Testing two models," J. Acoust. Soc. Amer., vol. 76, no. 5, pp. 1346–1356, Nov. 1984.

[7] T. Bänziger and K. Scherer, "The role of Intonation in emotional expressions, "Speech Commune. vol. 46, no. 3–4, pp. 252–267, 2005.

[8] M. Song, C. Chen, and M. You, "Audio-visual Based Emotion recognition using triples hidden Markov Model," in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, May vol. 5, pp. 877–880, 2004.

[9] M. Song, J. Bu, C. Chen, and N. Li, "Audio- Visual Based emotion recognition: A new Approach," in Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 1020– 1025. 2004.

[10] L. C. De Silva and P. C. Ng, "Bimodal emotion Recognition," in Proc. 4th IEEE Int. Conf. on Automatic Face and Gesture Recognition, France, pp. 332–335, Mar. 2000.

[11] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll, "Bimodal fusion of emotional data in an Automotive Environment," in Proc. IEEE Int. Conf. on Acoustic, Speech, and Signal Processing, vol. 2,pp.1085–1088, Mar. 2000.

[12] L. S. Chen, H. Tao, T. S. Huang, T. Miyasato, And R. Nakatsu, "Emotion recognition from Audiovisual Information," in Proc. IEEE 2nd Workshop on Multimedia Signal Processing, CA, pp. 83–88, Dec. 1998.

[13] Dellaert, F., Polzin, T., Waibel, A.: Recognizing Emotion in Speech. Proc. International Conf. on Spoken Language Processing, pp. 1970–1973, 1996.

[14] Scherer, K.R.: Adding the affective dimension: A New look in Speech analysis and synthesis In Proc. International Conf. on Spoken Language Processing, pp. 1808–1811, 1996.

[15] Zhou, G., Hansen, J.H.L., Kaiser, and J.F.: Nonlinear Feature Based Classification of Speech under Stress. IEEE Transactions on Speech and audio processing, IEEE Computer Society Press, Los Alamitos, vol. 9(3), 2001.

[16] Yacoub, S., Simske, S., Lin, X., Burns, Recognition of emotions in interactive voice Response system.Eurospeech 2003 Proc. 2003.

[17] M. Antonini, M. barlaud, P. Mathieu and I .Daubechies," Image Coding using Wavele Transform" IEEE Transaction on image Process. vol. 2, pp.205-220, 1992.

[18] Hongbin SUO1, 2, Ming LI1, Ping LU1 and Yonghong YAN, Automatic Language Identification with Discriminative Language Characterization Based on SVM, IEICE Transactions on Info and Systems, Volume E91-D, Number 3 , Pp. 567-575, 2008.

[19] C. Lee and S. Narayanan, "Toward detecting Emotions in spoken Dialogs," *IEEE* Transactions On Speech and Audio Processing, vol. 13, no. 2, Pp293-303, March 2005.

[20] M. Song, J. Bu, C. Chen, and N. Li, "Audio-Visual based emotion recognition – a new Approach," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, no. 2, pp. 1020-1025, 2004

.