ISSN print: 1812-1837
ISSN Online: 2312-7791

RECOGNITION OF EMOTIONS IN BERLIN SPEECH: A HTK BASED APPROACH FOR SPEAKER AND TEXT INDEPENDENT EMOTION RECOGNITION

C. Jeyalakshmi, B. Murugeswari and M. Karthick[c]

Department of ECE, K. Ramakrishnan college of Engineering, Tamilnadu, India,

Article received 3.2.2017; Revised 10.3.2017; Accepted 20.3,2017

ABSTRACT

Emotion recognition is one of the recent research area in speech processing to recognize the emotions from the human's speech. It finds various applications in different fields such as, education, gaming and in call centers to improve Human machine interaction. Researchers utilized different data bases and achieved different recognition accuracies. In this paper, we have proposed HTK based emotion recognition using EMO_DB Berlin database. This contains only ten speakers, uttered each emotion 10 times and the emotions considered are Anger, Boredom, Disgust, Fear, Happy, Neutral and Sad. Speaker and text independent emotion recognition is done by using the HMM models with MFCC features, implemented by HTK. The no. of states and mixtures are varied to validate the performance of the system. This system provides recognition accuracy of 68% for HMM models with 3mimtures and 3states. The system performance is also evaluated for speaker dependent and text independent emotions which produces a recognition accuracy of 81.7%. Even with very small amount of database the system produces better accuracy which can be improved by large amount of database.

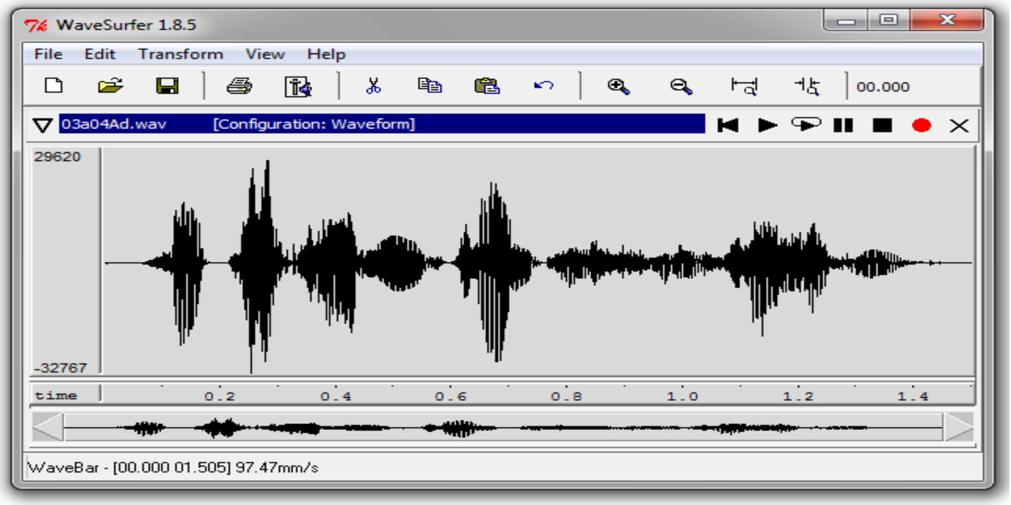
Key words: *Emotion recognition system (ERS), Mel frequency cepstral coefficients(MFCC), Formants, Spectral analysis, Hidden Markov Model tool kit.*

INTRODUCTION

In speech processing Emotion recognition is an emerging area in human-computer interaction (Cowie, et al., 2001). Recognizing the emotions from the speech is a challenging task since emotions are unique for each person corresponding to the affective state. Tremendous research is going on for recognizing emotions from various data bases. Speech signals does not convey us, not only the information, it also implicitly tells us about the affective state of a speaker. For recognizing the emotions, speech is one of the best method besides human facial expressions (Schuller, et al., 2003). It finds applications in web interactive services, information retrieval, medical analysis and text to speech synthesis. During the hazardous situations, it is used to find the emotional state of the affected persons where physical presence of the others is not possible. It can be utilized for psychiatric diagnosis, intelligent toys, lie detection, learning environ-ments, and educational software (Emerichand and Lupu, 2011). ERS is developed using discrete HMM models (Nwe, et al., 2003) and comparative analysis for ERS using various classification techniques has been done (Donn et al., 2007). Modulation spectral feature for ERS has been introduced (Nwe, et al., 2011). Acoustic, statistical feature and hierarchial binary classifier for ERS have been utilized (Lee, et al., 2011). Gender detection is done by combined features such as pitch, energy and MFCC as feature (Vogt and Andr, 2006). Emotion recognition has been done using MFCC and GMM (Rao, et al., 2012). MFCC feature and NN classifier for recognizing emotion is utilized (Ankur, et al., 2013). Speaker identification for various emotional environmental has been analyzed with the system that used log frequency power coefficients as feature and performance evaluation is done using HMM, CHMM and SPHMM (Shahin, 2009). Speaker recognition has been done using MFCC and GMM (Shashidhar, et al., 2012). In this proposed work, the development of emotion recognition is done by using berlin emotional speech database, with MFCC features and HMM modeling techniques to analyze the performance of the speaker dependent and speaker independent ERS.

EMOTIONAL SPEECH CORPUS: Mostly the data base for emotions are collected from actors and from spontaneous speech spoken by non-acted persons. In this work, the Berlin emotional speech database is considered has 500 utterances spoken by actors in happy, angry, anxious, fearful, bored, disgusted and in neutral state. The above database is uttered by various actors and various texts approximately 10 in which five are male and five are female. The text contents are very simple to be pronounced by the speaker. The following letters are used to represent different emotions in german: 'W' for anger, 'L' for boredom, 'E' for disgust, 'A' for anxiety or fear, 'F' for happiness, 'T' for sadness, 'N' for Neutral version. Each utterance is named according to a common scheme. i.e position 1-2 represents number of speakers, position 3-5 used for code of text, position 6 represents type of emotion and position 7 tells about database version.

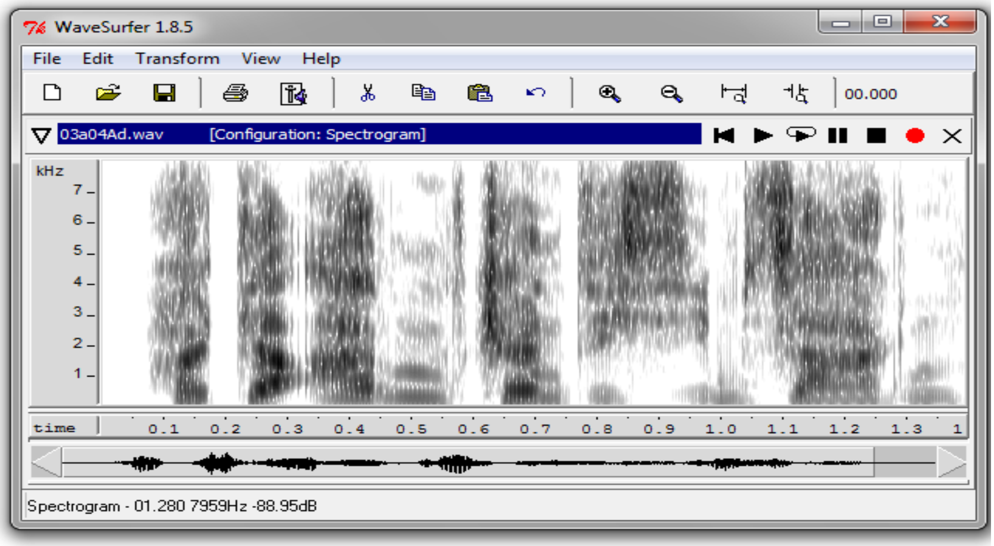
CHARACTERISTICS OF EMOTIONAL SPEECH: To show the variations in the characteristics between various emotions, spectrogram is shown for two emotions with its waveform. In spectrogram intensity of the input speech in different bands of frequency can be obtained. Figure 1and 2 shows the Emotional speech wave form and spectrogram for anger similarly, figure 3 and 4 shows the Emotional speech waveform and spectrogram of the emotion disgust. The dark portion of the figures indicates the spectral intensity at a specified frequency. Among all emotions, disgust has small amount of database since it is difficult to exhibit the emotion. So, for this emotion few speakers don't have the database. Neutral is similar to the ordinary speech and for anger, fear, sadness we have reasonable database.


WaveSurfer 1.8.5
File Edit Transform View Help
00.000
29620
-32767
time 0.2 0.4 0.6 0.8 1.0 1.2 1.4

03a04Ad.wav [Configuration: Waveform]

WaveBar - [00.000 01.505] 97.47mm/s

Figure-1: Waveform of emotion ‘anger’

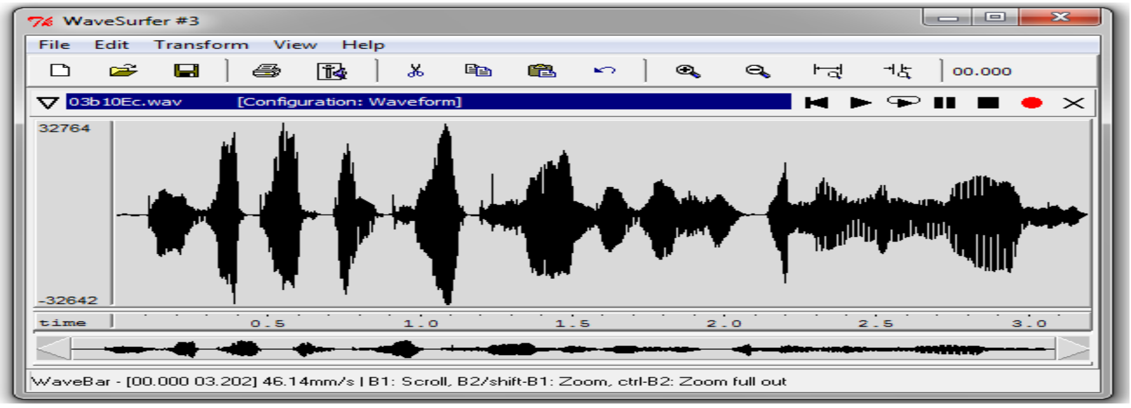

WaveSurfer 1.8.5
File Edit Transform View Help
00.000
kHz
7
6
5
4
3
2
1
time 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1

03a04Ad.wav [Configuration: Spectrogram]

Spectrogram - 01.280 7959Hz -88.95dB

Figure- 2: Spectrogram of emotion ‘anger’

In the spectrogram, the horizontal lines indicate the pitch of the input emotional speech during voiced regions (Rabiner and Juang, 1993). High frequency energies can be depicted in the spectrogram for unvoiced speech, and due to the low signal level at the time of silence, nothing can be seen in the spectrogram. Like time domain representation of the speech signal spectrogram plots are also completely different for the different emotions due to variation in the speaking rate, Pronunciation, abstraction, directness etc.



03b10Ec.wav [Configuration: Waveform]

WaveBar - [00.000 03.202] 46.14mm/s | B1: Scroll, B2/shift-B1: Zoom, ctrl-B2: Zoom full out

Figure -3: Waveform of emotion ‘disgust’

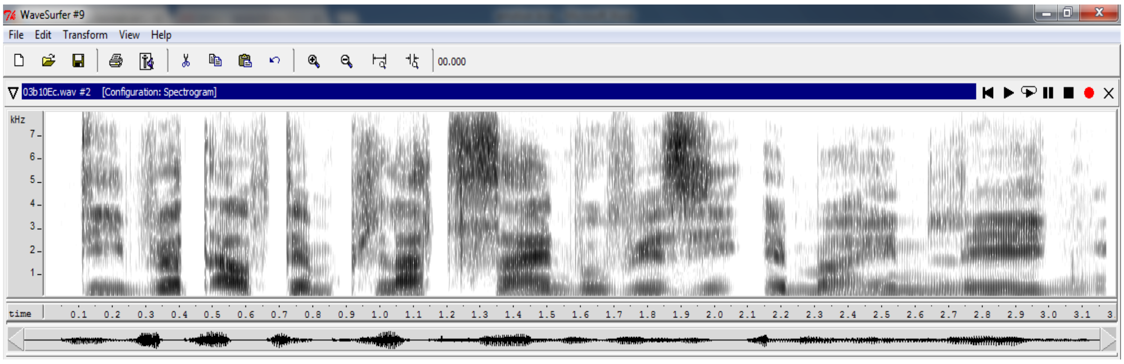


Figure- 4: Spectrogram of emotion 'disgust'

ERS USING MFCC FEATURES AND HMM MODELS: MFCC features are widely used for speech/ speaker/language identification. Murty and Yegnanarayana (2006) have developed emotion recognition system using MFCC and their detailed procedure used for MFCC feature extraction is shown in Figure 5.

Feature extraction: The front end spectral analysis of any pattern recognition is the feature extraction and they should exhibit statistics which are highly invariant across speakers and speaking environment. From the literature, it is evident that, MFCC is the commonly used features for recognizing the speech. For extraction of MFCCs, filters spaced linearly at low frequencies and logarithmically at high frequencies are used to capture the phonetically important characteristics of speech. Then process of computing MFCC step by step is shown in Figure 5.

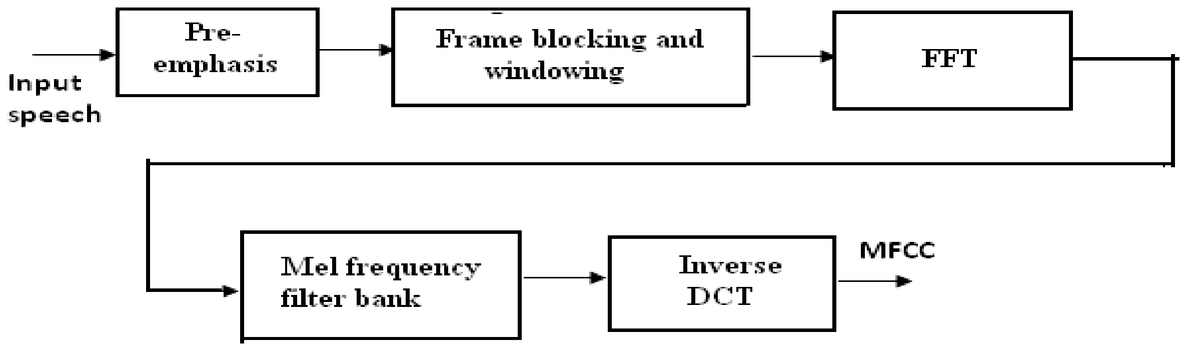


Figure- 5: Block diagram of MFCC feature extraction method

The pre-emphasis filter used is a first order FIR filter, to remove the noise which is converted into frames. The 20 msec. is taken as a frame length and 10 msec. overlap between the adjacent frames. This process corresponds to 320 samples per frame (N) with adjacent frame being separated by 160 samples (M). If M is very much less than N, then the spectral estimates from frame to frame will be quite smooth (Rabiner and Juang, 1993). Then Hamming window is used to reduce the signal discontinuities and FFT is applied for each frame. Forty band pass filters are used to filter the spectrum of each frame and then power of each band is calculated. For the given frequency 'F', mel frequency is calculated using equation (1) for conversion of linear frequency into mel frequency (Umesh, et al., 1999).

$$mel(f) = 2595 * \log(1 + f(Hz)/700) \qquad (1)$$

Using Equation (2) cepstral coefficients are calculated

$$C_n = \sum_{k=1}^{K} (\log S_k) \cos[n(k-0.5)\pi / K] \qquad (2) \quad n=1, 2 \ldots L$$

where S_k is the output power of the kth filter of the filter bank and L is the desired length of the cepstrum.

In this work on emotion recognition, MFCC features consisting of 13 MFCC+13 delta+13 acceleration coefficients are extracted and system is analyzed by applying the features to the correct emotion models.

HMM model creation for each Emotion: From the literature, it is evident that, HMM is utilized for all, to achieve high performance speech recognition system. It is a generator of vector sequences in which no. of states are connected by arcs used to derive maximum likelihood parameters (Rabiner, 1989). For classifying a test speech corresponding to a particular emotion, feature vectors of test speeches are applied to the emotion and log likelihood values are computed. Among the log likelihood outputs from all models, the model which produces the largest log likelihood score has been then identified as emotion associated with that model. In this work, HTK is used to implement HMM models. Hence a HMM can be defined by no. of states, observation vector type, no. and width of the input data and mixture weights for each emitting state are to be defined. Using the model definition, proto type HMM models are developed (Rabiner, 1989).

In this work, prototype models are initialized with number of mixtures and number of states and models are trained using HTK (Young, et al., 2001). The model parameters are re-estimated using Baum Welch re-estimation algorithm. Now for all the emotions, models are developed. During the recognition phase, when unknown emotion is given, MFCC features are extracted from them and they are compared with the HMM models developed for all the emotions by considering the dictionary, word network and task grammar, the corresponding emotion is identified.

RESULTS AND DISCUSSION

For speaker, independent and text independent emotion recognition, emotional utterances from first 5 speakers are used for training and utterances from other 5 speakers are used for testing. A total of 241 utterances are considered for training and 290 utterances are considered for testing. In the testing phase, the test speeches are converted into series of acoustical vectors, i.e. feature vectors. These vectors are compared with the models already developed for each language. Then **HVite** tool compares the speech file against the HMM networks and produces a corresponding transcription.

Experimental analysis based on HTK: Speaker and Text independent ERS system is evaluated using the data for utterances in specific emotions. For creating a training model, 241 utterances spoken by 5 speakers in different emotions are considered and for each signal. Then, for each frame, MFCC features are computed. Initially, proto type HMM models are created for different states and mixtures. The no. of states represents the no. of phonemes in a word and no. of mixtures corresponds to variability among the speeches. Similar ly, in our work it is not possible to initialize the no. of states corresponds to no. of phonemes, since lengthy sentences are considered. So, that, in our work, number of mixture is initially taken as 3 and number of states is taken as 3. The proto type models are generated for 3m3s (3 states and 3 mixtures).

Then, for the input utterances corresponding to all the emotions HMM parameters are re estimated using Baum Welch algorithm. During testing phase, 290 utterances of all emotions are considered and MFCC features are extracted for the test speech. Out of the 290 utterances, 198 are correctly recognized and the overall recognition accuracy obtained is 68.27%.

We can express the recognition of emotions using the confusion matrix. It is table which gives the performance of an algorithm visually. It shows exactly how the system is confusing with other classes. In this matrix each row represents actual emotion and each column represents, no. of times the particular emotion is recognised by the system.

The total no. of misidentified emotion is 92 out of 290 which is due to confusion by other emotions. This is clearly depicted in table 1. The performance of the system for 3m3s is depicted in figure 6.

Table- 1: Performance analysis of speaker independent emotion recognition for 3m3s

Emotions	Total utterances	Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
Anger	67	65	-	-	-	2	-	-
Boredom	46	-	25	2	1	-	12	6
Disgust	30	2	16	8	1	2	-	1
Fear	40	8	1	-	25	5	-	1
Happiness	39	20	-	1	1	17	-	-
Neutral	36	1	1	1	2	-	30	1
Sadness	32	-	3	1	-	-	-	28

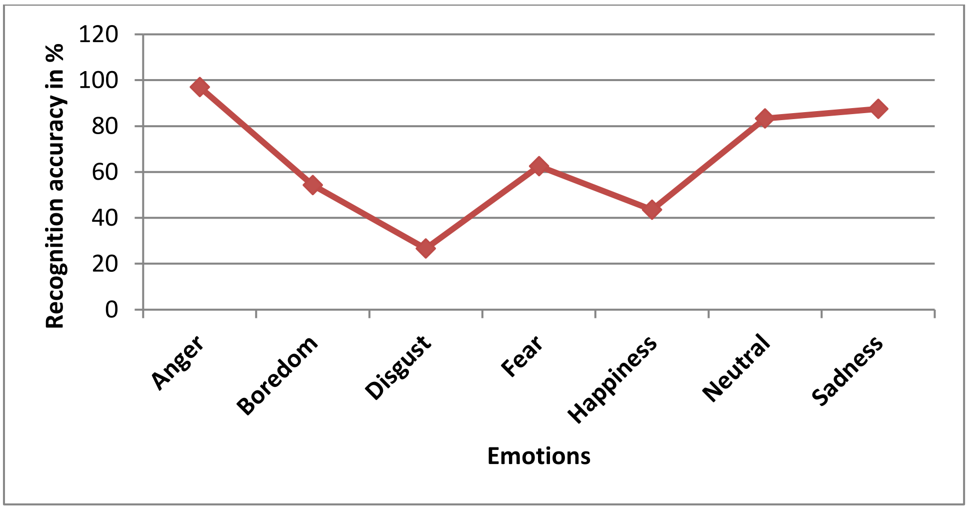


Fig.-6: Performance of the ERS system for 3 mixtures and 3 states

To analyze the system performance, the no. of mixtures is increased from 3 to 5 and the states are fixed at 5 and again the system is validated for all the emotions. Table 2 shows the confusion matrix of ERS system for 3m5s and figure 7 represents the corresponding recognition accuracy for different emotions.

Table 2 – Performance analysis of speaker independent emotion recognition for 3m5s

Emotions	Total utterances	Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
Anger	67	64	-	-	1	2	-	-
Boredom	46	-	36	-	-	1	3	6
Disgust	30	1	16	5	3	3	1	1
Fear	40	2	1	-	21	7	-	1
Happiness	39	18	-	-	1	20	-	-
Neutral	36	-	11	-	4	1	20	-
Sadness	32	-	2	1	-	-	-	29

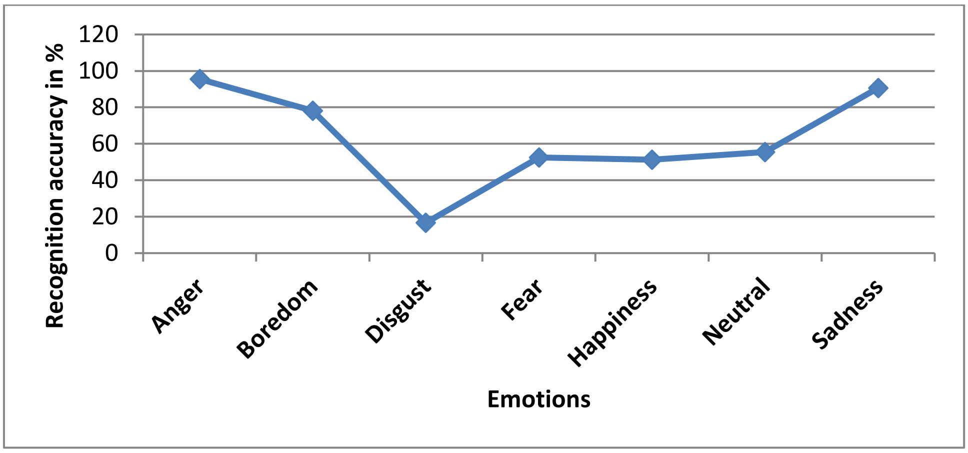


Fig.-7: Performance of the ERS system for 3 mixtures and 5 states

The total no. of misidentified emotion is 95 out of 290 which are due to confusion by other emotions. This is clearly illustrated in table 3. The overall recognition accuracy is 67.2%. From the results, for 3m3s and 3m5s, there is only slight change in the recognition accuracy.

Now the states are randomly chosen between 2 to 13 and mixture is taken as 3. In order to compare the recognition accuracy, the same 290 utterances of all emotions are considered. The overall recognition accuracy is 61.0%. It is 6.2% less than the previous results. The table 3 and figure 8 clearly indicates the system performance.

Table-3: Performance analysis of speaker independent emotion recognition for 3m13s

Emotions	Total utterances	Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
Anger	67	67	-	15	-	-	-	-
Boredom	46	2	1	21	-	-	17	5
Disgust	30	4	15	15	-	6	2	1
Fear	40	21	1	10	5	-	3	1
Happiness	39	31	-	1	-	7	-	-
Neutral	36	2	-	-	-	-	34	-
Sadness	32	-	1	1	2	-	-	28

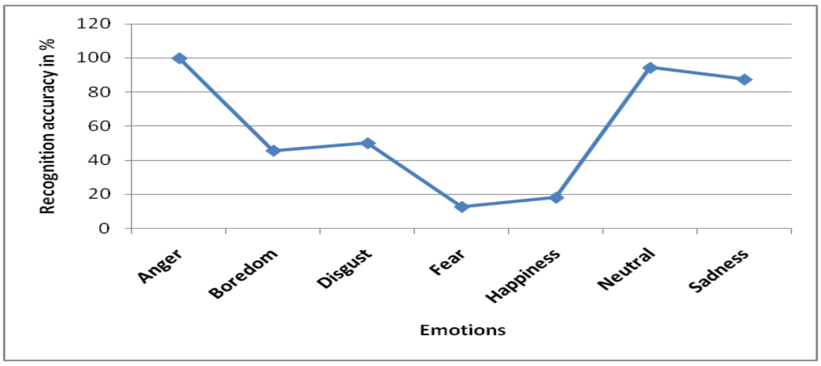


Fig.-8: Performance of the system for 3 mixtures with maximum of 13 states

To check the performance of the ERS system for higher value of mixture values it is increased from 3 to 10 and the states are randomly chosen between 2 to 13 similar to the previous case. The overall recognition accuracy obtained is 56.89%. Again, the accuracy is reduced from 61% to 56.89% which reveals that by increasing the mixture value there is no significant increase in the accuracy. This is clearly indicated in the Table 4 and figure 9.

Table- 4: Performance analysis of speaker independent emotion recognition for 10m13s

Emotions	Total utterances	Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
Anger	67	67	-	-	-	-	-	-
Boredom	46	5	22	1	-	-	14	5
Disgust	30	12	2	-	-	10	4	1
Fear	40	22	1	-	6	8	2	1
Happiness	39	35	-	-	-	4	-	-
Neutral	36	2	-	-	-	-	34	-
Sadness	32	-	-	-	1	-	-	31

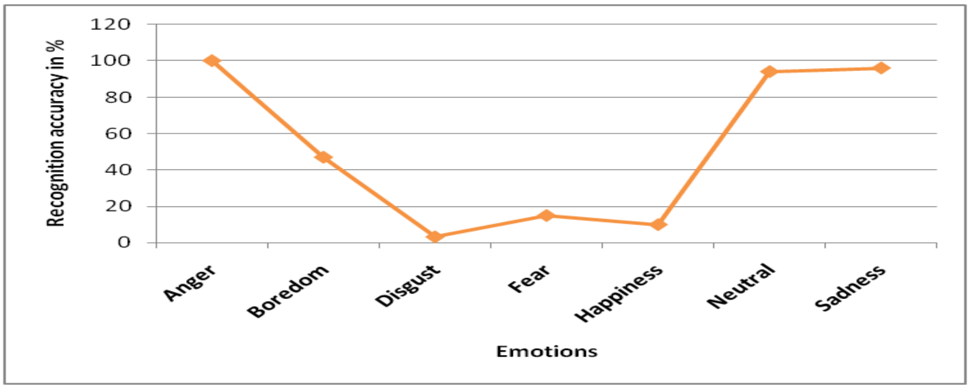


Fig.-9: Performance of the system for 10 mixtures with maximum of 13 states

From Table 2-5, it is evident that the speaker independent and text independent emotion recognition produces a maximum recognition accuracy of 68.27% with 3m3s nearly equal to the 67.2% for 3m5s. If we increase the mixture doesn't improve the performance. Table 6 shows the performance analysis on the system by computing the accuracy with respect to speaker dependent and text independent emotions i.e for training, the samples from all the speakers are considered. Since the test speaker is also included in the training the ERS system produces 81.7%. i.e out of 258 emotional utterances, 211 emotions are correctly recognized. This is depicted in table 6 and figure 10 clearly. With higher no. of training samples, we can improve the accuracy further.

Table- 6: Performance analysis of speaker dependent emotion recognition system for 3m5s

Emotions	Total utterances	Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
Anger	63	53	-	2	2	7	-	-
Boredom	43	-	34	1	-	-	2	3
Disgust	23	-	-	16	1	1	1	1
Fear	47	1	-	-	23	5	4	1
Happiness	50	8	-	2	-	26	-	-
Neutral	60	-	3	-	-	-	33	-
Sadness	34	-	-	2	-	-	-	26

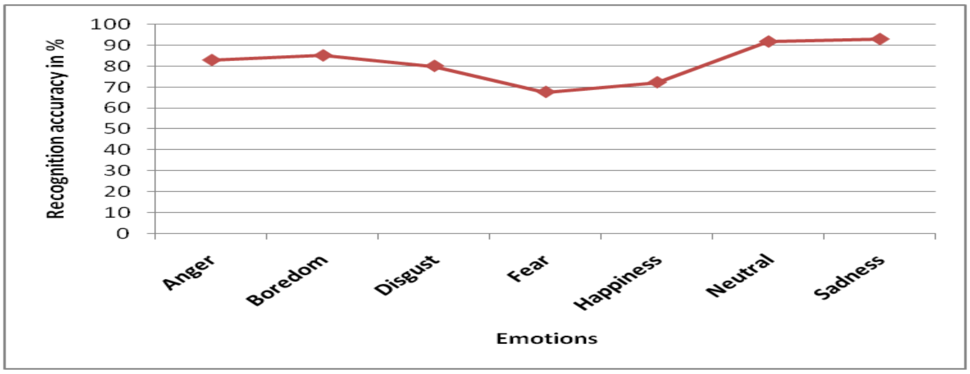


Fig. -10: Performance of the speaker dependent ERS system for 3 mixtures and 5 states

CONCLUSIONS

In this paper, the use of different value of mixtures and states for developing the emotion models using HMM implemented by HTK is proposed. The performance of ERS system is also assessed for the speaker independent and speaker dependent environment. With respect to speaker independent emotion recognition, from all the results it is clear that, the system gives the high accuracy for 3 emotions anger, neutral and sadness. Similarly, better accuracy is obtained for boredom, fear and happiness. But for the emotion disgust, accuracy is very low due to lesser no. of samples. From the results, it is understood that the overall accuracy for the speaker dependent system is higher than that of the speaker independent system. 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 the emotional speech revealing the characteristics of the speech and speaker. The same system can also be verified for different features and modeling techniques with the same database. By improving the size of the database, we can improve the recognition accuracy for the ERS system.

REFERENCES

Ankur, S., N. Panwar and S. Panwar, Emotion Recognition from Speech. *International Journal of Emerging Technology and Advanced Engineering* 3(2): 187-191 (2013).

Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S.Kollias, W.Fellenz and J.G.Taylor, Emotion recognition in human-computer interaction, IEEE Signal Processing Magazine 18(1): 32–80 (2001).

Donn, M., R. Wang and L.C. De Silva, Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication* 49: 98–112 (2007).

Emerichand, S. and E. Lupu, Improving speech emotion recognition using frequency and time domain acoustic features, Proceedings of the Signal Processing and Applied Mathematics for Electronics and Communications (SPAMEC'11), Cluj Napoca, Romania, Pp, 85-88 (2011).

Lee, C.-C., E.Mower, C.Busso, S.Lee and S. Narayanan, Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication* 53: 1162–1171 (2011).

Murty, K.S.R. and B. Yegnanarayana, Combining evidence from residual phase and MFCC features for speaker recognition. IEEE Signal Processing Letters 13(1): 52–55 (2006).

Nwe, T.L., S.W. Foo and L.C. De Silva, Speech emotion recognition using hidden Markov models. *Speech Communication* 41: 603–623 (2003).

Rabiner, L. and B.H. Juang, *Fundamentals of speech recognition*, Prentice Hall, NJ (1993).

Rabiner, L.R., A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE 77(2): 257-286 (1989).

Rao, K.S., T.P. Kumar, K. Anusha, B. Leela, I. Bhavana nd S.V.S.K. Gowtham, Emotion Recognition from Speech. *International Journal of Computer Science and Information Technologies* 3(2): 3603-3607 (2012).

Shuller, B., G. Rigoll and M. Lang, Hidden Markov model based speech emotion recognition, Proceedings of the International Conference on Multimedia and Expo (ICME '03) 1: 401–404 (2003).

Shahin, I., Speaker Identification in Emotional Environments. *Iranian Journal of Electrical and Computer Engineering* 8(1): 41-46 (2009).

Shashidhar, G.K., K. Sharma and K.S. Rao, Speaker Recognition in Emotional Environment. Communications in Computer and Information Science 305: 117-124 (2012).

Umesh, S., L. Cohen and D. Nelson, Fitting the mel scale, Proc. of ICASSP, IEEE 1: 217-220 (1999).

Vogt, T. and E. Andr, Improving Automatic Emotion Recognition from Speech via Gender Differentiation, *Proc. Language Resources and Evaluation Conference* (2006).

Wua, S., H.F.B. Tiago and W.-Y. Chan, Automatic speech emotion recognition using modulation spectral features. *Speech Communication* 53: 768–785 (2011).

Young, S., G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, The HTK Book, Cambridge University Engineering department (2001).