Emotion Specific Feature Representation by using Mel-Frequency Cepstral Coefficients

Ch.V.Kiranmayi[1], Biswaranjan Barik[2]
[1]PG Scholar, [2]Associate Professor
Department of ECE, Godavari Institute of Engineering and Technology, Rajahmundry, Andhra Pradesh, India.
(E-mail:[1] Kiranmayee427@gmail.com, [2]barikbiswa65@gmail.com*)*

***Abstract*—** Emotion Recognition from speech is a new field of research, it has number of applications. One of the applications is Human machine interaction (HCI). This research work focused on the development of various state-of-the-art AER techniques to recognize emotions in Telugu. It also compared the performance of these techniques and designed new enhanced ones for the better operation of the emotion recognition system. The effect of DWT, cepstral coefficients in the detection of emotions is performed and also a comparative analysis of cepstrum, Mel-frequency Cepstral Coefficients (MFCC), pitch on emotion classification is done. The classification task is done by using artificial neural network's back propagation algorithm. The Developed method shown improved recognition rates of identifying four emotions from Telugu database .Misclassification rate is reduced.

Keywords— AER, DWT, MFCC, Cepstrum and Pitch, ANN. Introduction

I. INTRODUCTION

Speech emotion recognition is the process of identifying emotions from the speech input signals. It is very difficult to predict human emotions quantitatively. Though facial expressions and gestures are the best ways to figure out one's emotions, it becomes difficult to identify them as the age of a person increases, because people learn to control their expressions with age and experience [1]. For instance, it can be used to auto remote telephone service center for discovery of dissatisfaction of customers or remote teaching to timely recognize emotions of students for proper treatment for the purpose of improving teaching quality [6]. It can also be used to the criminal scout for auto detection of psychological state of criminal suspects and auxiliary lie detection [7]. It is observed that SER provide efficient communication between human being and a machine in Human and Machine Interaction (HMI) applications. Earlier research work [1] have used MFCC and enlarged MFCC for feature extraction and failed to identify the emotion i.e. sad and also emotion recognition rate is very poor. To improve the same speech signal is preprocessed before extracting emotion specific features.

II. EMOTION CORPUS

Telugu database is considered for training and testing the Artificial Neural Network. One male and one female speaker's voices are recorded with four different emotions namely angry, happy, neutral and sad.

III. DISCUSSION

This section explains process involved in the planned method.
The total practice is carried in two steps. First step involve extraction of features using DWT, MFCC, Cepstrum. Second step is classification process. Artificial Neural Network is used as pattern recognizer and a classifier.

IV. PROPOSED SYSTEM BLOCK DIAGRAM

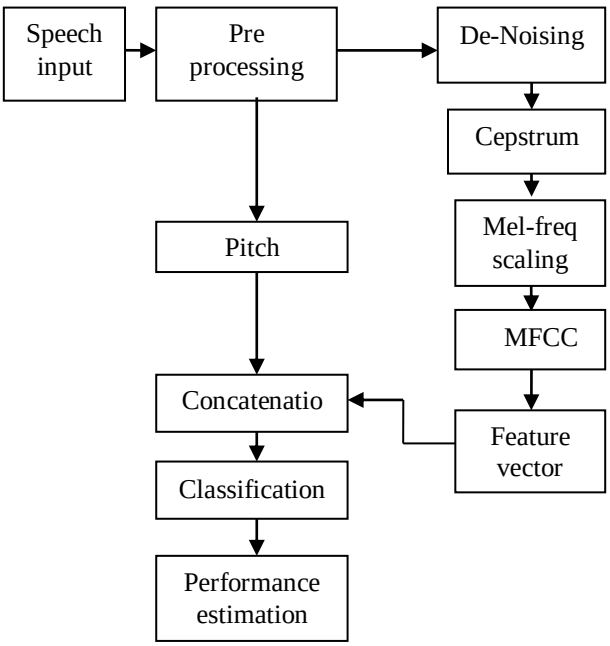

Fig.1.Speech emotion recognition block diagram.

Referring to the diagram above, it is clear that the input speech will pass through two major stages in order to get the speaker identity, they are:

1- Feature extraction.
2- Classification and Feature matching.

A. *Pre-processing*

Pre-processing of speech signals is the initial and crucial step in development of efficient and robust emotion recognition system after creating the database. Moreover, due to the variations in a speech signal, two visually similar waveforms may not produce perceptually similar sounds.To enhance the accuracy of the extraction process, speech signals are normally pre-processed before features are extracted.This pre-emphasis is done by using a energy equation.

B. *De-noising*

When voice signals are recorded, different types of degradation components like background noise, noise introduced by environment and recording hardware as well as reverberation and disturbances may interfere with the required speech contents. This affects the quality and intelligibility of the speech signals and this in turn causes degradation in the performance of the emotion recognition system. Discrete Wavelet Transform plays a crucial role while evaluation of emotion performance. Here DWT is used to eliminate noise in speech signal. The type of wavelet used is 'db4' because of its high accuracy.

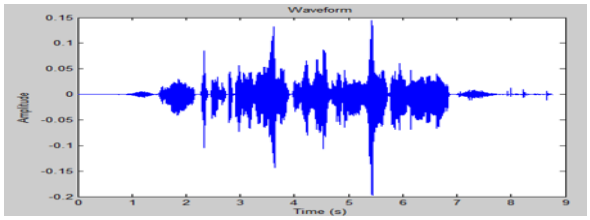

Fig.2 signal before noise elimination.

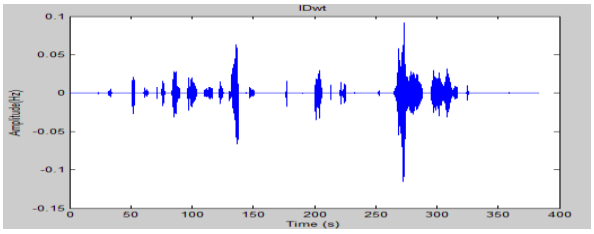

Fig.3.De-noised speech signal.

C. *Cepstrum*

Cepstrum is obtained by taking the inverse transform of the logarithm of Fourier transform of the signal [8]. Cepstrum separates excitation from Speech signal which is convenient for encoding. Excitation frequency in speech processing carries information regarding speaker's mood.

MFCC: Mel frequency cepstral coefficients are obtained by applying Mel-scaled filter banks to the cepstral coefficients obtained in the previous step. Discrete cosine transform is used to convert the MFCC coefficients into time domain.

Pitch: Pitch is the fundamental frequency or the lowest frequency of the sound signal. It is one of the useful features for prosody analysis of the speech signal and is considered as an important clue for recognizing the speaker's emotions.

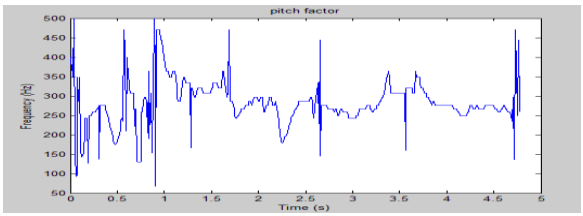

Fig.4. Pitch of the speech signal

Concatenation: All the extracted features are concatenated and applied as input to the classifier.

D. *Classifier*

Artificial Neural Network is used as classifier especially back propagation algorithm is used for the purpose of recognition process.

V. RESULTS

Total 157 emotion samples are obtained from speakers and out of 157, 141 have been considered as training data and remaining 16 (4 samples for each emotion) considered as testing data.

To identify individual emotion and its performance with single evaluation will not give accurate results. so here we considered 1 Iteration to calculate accuracy with different feature sets. DWT+Pitch+CEPS will give 89.71for 1 iteration. DWT+CEPS+MFCC+Pitch has 95.38. This combination will give higher performance when compared with other feature combinations. Here in this DWT will be considered as emotion-Specific feature and MFCC is considered as Spectral feature and Pitch will be considered as Prosody feature.

The following figures shows the results of emotion 'angry' :

a) Angry

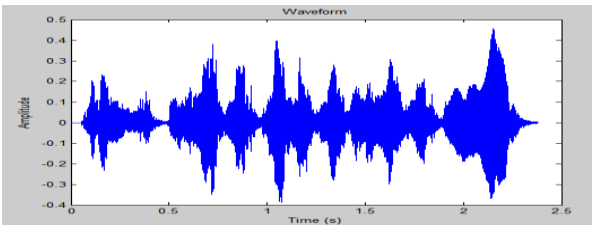

Fig.5. Speech signal of emotion angry before noise elimination.

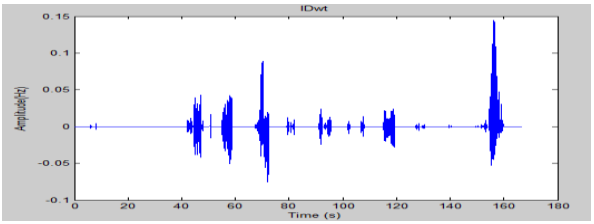

Fig.6.Denoised speech signal

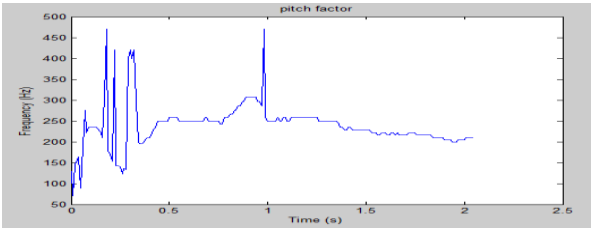

Fig.7. Pitch of the speech signal

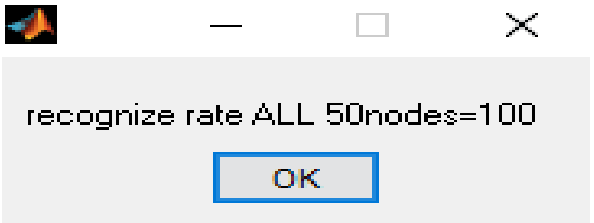

Fig.8. Recognition rate with 50 ANN nodes

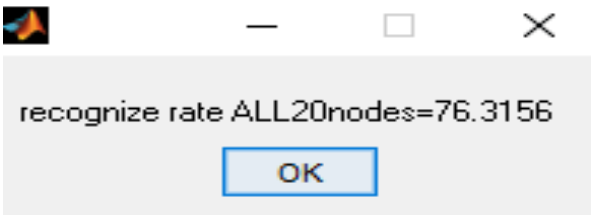

Fig.9. Recognition rate with 20 ANN nodes

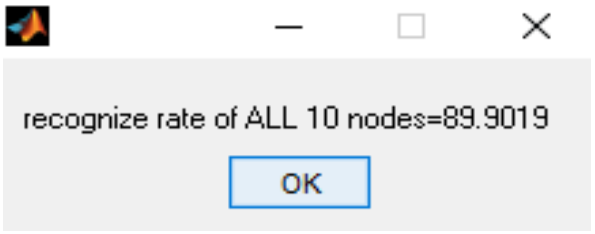

Fig.10. Recognition rate with 10 ANN nodes

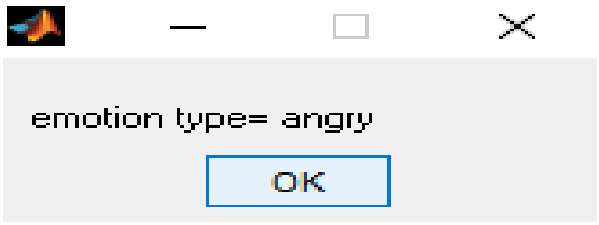

Fig.11. Type of recognized emotion

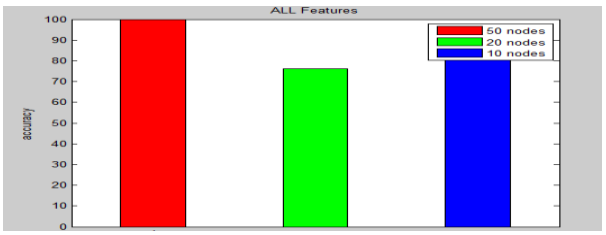

Fig.12. Comparison of accuracies with 50, 20, 10 ANN nodes

VI. Tables

Table 1: Emotion recognition rate

50 nodes (%)	20 nodes (%)	10 nodes (%)
95.38	94.75	41.625

From table 1 we considered with different nodes in ANN network i.e 50, 20 and 10. By using 50 nodes the efficiency rate is improved and misclassification is reduced by a factor of 10.

Table 2: Over all emotion recognition rate

Over ALL Feature Set with 1 Iterations	**Accuracy (%)**
DWT	87.17
Pitch	63.29
Cepstrum	66.34
MFCC	84.44
DWT+Pitch	95.76
DWT+CEPS	84.74
DWT+MFCC	85.07
DWT+MFCC+CEPS	86.785
DWT+MFCC+Pitch	88.33
DWT+Pitch+CEPS	89.711
DWT+CEPS+MFCC+Pitch	95.38

VIII. CONCLUSION

The first remarkable conclusion about these results is the fact that in all cases the low level feature seems to be useful for emotion recognition. Instantaneous features provide better performance than syllabic ones for both energy and pitch, and pitch features work also better than energy ones. As a result, the best partial combination is instantaneous pitch. Being the worst the syllabic contour of energy. Nevertheless, the best combination at all is the complete 11 features set, which reports the highest accuracy independently of the number of states. All the sets of features show a noticeable improvement when bigger models are used, although there is certain saturation for MFCC of more than 32 states. Misclassification is reduced by 10% when considering 4 emotions and we used 50 target nodes instead of 10 and 20 to reduce misclassification. Here different combinations of features to identify the corresponding emotion and these features are referred as Emotion-Specific features. Here we considered DWT, Cepstrum, MFCC and Pitch are used to extract the features information. After feature extraction we do classification, the type of classifier is artificial neural network. And we evaluate its performance by its mean value. Compared with existing system there will be a improvement of 10% with 4 emotions. Future scope may be considered with different features with different classifiers and with varied databases.

IX. REFERENCES

[1]. S Lalitha, D Geyasruti, R Narayanan, M Shravani (2015) Emotion Detection Using MFCC and Cepstrum Features. In: Procedia Computer Science 70 (2015) 29-35.

[2]. FirozShah.A, Vimal Krishnan V.R., RajiSukumar.A, AthulyaJayakumar, BabuAnto.P "Speaker Independent Automatic Emotion Recognition from Speech:-A Comparison of MFCCs and Discrete Wavelet Transforms" 2009 International Conference on Advances in Recent Technologies in Communication and Computing, pp:528-531,2009.

[3]. S.Kim, P.Georgiou, S.Lee, S.Narayanan "Real-time emotion detection system using speech: Multi-modal fusion of different timescale features", Proceedings of IEEE Multimedia Signal Processing Workshop, Chania, Greece, 2007.

[4]. K.V.Krishna Kishore, P. Krishna Satish "Emotion Recognition in Speech using MFCC and Wavelet Features",3rd IEEE International Advance Computing Conference.

[5]. Firoz Shah.A, Vimal Krishnan V.R, Raji Sukumar.A, Athulya Jayakumar, Babu Anto.P "Speaker Independent Automatic Emotion Recognition from Speech:-A Comparison of MFCCs and Discrete Wavelet Transforms" 2009 International Conference on Advances in Recent Technologies in Communication and Computing, pp:528 531,2009.

[6]. Batliner A, Fischer K, Huber R, et al. How to Find Trouble in Communication [J]. Speech Communication, 2003, 40,(1-2):pp.117~143;

[7]. Cowie R, Douglas-Cowie E, Tsapatsoulis N, et al. Emotion Recognition in Human-Computer Interaction[J].IEEESignalProcessingmagazine,Vol.18,No.1,2001 pp.32~80;

[8]. J. SirishaDevi ,Dr. Srinivas Yarramalle,Siva Prasad Nandyala "Speaker Emotion Recognition Based on Speech Features and Classification Techniques" I.J. Image, Graphics and Signal Processing, 2014, No:7, pp: 61-77 ,June 2014.

[9]. Siqing Wu, Tiago H.Falk, Wai-Yip Chan "Automatic speech emotion recognition using modulation spectral features". In: Speech communication 53 (2011) 768-785.

[10]. Michael Grimm, KristianKroschel, Emily Mower, Srikanth Narayanan"Primitives-based evaluation and estimation of emotions in speech". In: Speech communication 49 (2007) 787-800.