This full-text paper was peer-reviewed and accepted to be presented at the IEEE WiSPNET 2017 conference.

Speech Based Human Emotion Recognition Using MFCC

M.S. Likitha,[1] Sri Raksha R. Gupta,[2] K. Hasitha[3] and A. Upendra Raju[4]
Dept. of Electronics, Mount Carmel College, Autonomous, Bangalore
Email: [1]likitha11292@gmail.com [2]raksha.gupta1993@gmail.com [3]hasitha.koka@gmail.com [4]upendraraju@gmail.com

***Abstract*—Speech is a complex signal consisting of various information, such as information about the message to be communicated, speaker, language, region, emotions etc. Speech Processing is one of the important branches of digital signal processing and finds applications in Human computer interfaces, Telecommunication, Assistive technologies, Audio mining, Security and so on. Speech emotion recognition is important to have a natural interaction between human being and machine. In speech emotion recognition, emotional state of a speaker is extracted from his or her speech. The acoustic characteristic of the speech signal is Feature. Feature extraction is the process that extracts a small amount of data from the speech signal that can later be used to represent each speaker. Many feature extraction methods are available and Mel Frequency Cepstral Coefficient (MFCC) is the commonly used method. In this paper, speaker emotions are recognized using the data extracted from the speaker voice signal. Mel Frequency Cepstral Coefficient (MFCC) technique is used to recognize emotion of a speaker from their voice. The designed system was validated for Happy, sad and anger emotions and the efficiency was found to be about 80%.**

***Index Terms*—Speech processing, Emotion recognition, Feature extraction, MFCC, FFT.**

I. Introduction

Emotion recognition from a human speech is an attractive field of speech signal processing. It is drawing more attention in the applications where emotion recognition eases the speaker identification and mental status, such as in criminal investigation, intelligent assistance [1], detecting frustration, disappointment, surprise/amusement [2], health care and medicine [3] and a better Human Computer Interface [4]. Extracting the emotional state of a speaker from his/her speech is called speech emotion recognition. The speech emotion recognition involves analysis of the speech signal to identify the appropriate emotion based on training its features such as pitch, formant and phoneme. For feature extraction and testing of a speech signal a good number of algorithms have been formulated. Few of them are Artificial neural networks (ANN), linear prediction cepstrum coefficients (LPCC), Mel Frequency cepstrum coefficients (MFCC), combination of Linear Prediction coefficients and Mel Cepstrum coefficients (LPCMCC), the Support Vector Machine (SVM); combination of HMM and SVM etc [5]. Our proposed work is based on feature extraction using MFCC and decision making using standard deviation. Flow chart for the proposed emotion recognition system is as shown in Fig. 1. Organization of this paper is as follows: Section II describes the feature extraction using MFCC, Section III illustrates decision making using mean and standard deviation, Section IV gives results-discussions and conclusions are made in Section V.

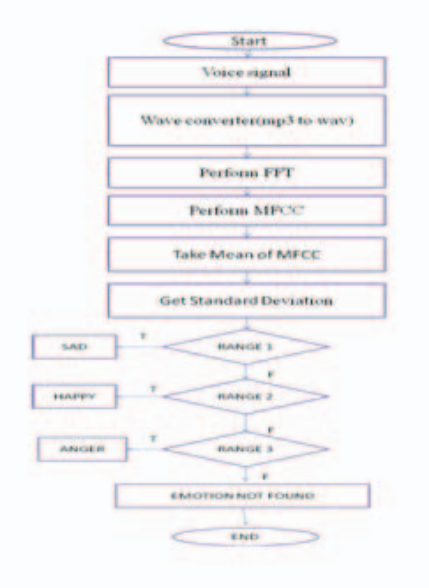

Fig. 1. Flow chart for the proposed system.

II. Feature Extraction Using MFCC

The acoustic characteristic of the speech signal is Feature. A small amount of data from the speech signal is extracted to analyse the signal without disturbing its acoustic properties. This extracted signal is used for training and testing phases. It comprises of the following steps:

A. Frame Blocking

Processing of speech signals is done in short time intervals called frames with sizes generally between 20 and 40 ms [7]. Overlapping of frames is done to smoothen the transitions between frames by a predefined size. The first frame consists of the first $N = 256$ (typical value) samples. The second frame begins at $M = 100$ (typical value) samples after the first frame, and overlaps it by $N - M$ samples and so on. This process continues until the entire speech signal is accounted.

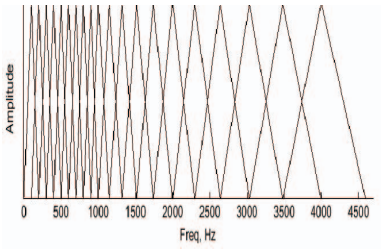

Fig. 2. Filter bank in Mel frequency scale.

B. Windowing

After frame blocking, the signal is pass through a windowing function to minimize discontinuities and spectral distortion at the extremes of each frame. The result of windowing a signal $x(n)$ is given by [2],

$$Y(n) = x(n)w(n), \quad 0 \leq n \leq N-1 \tag{1}$$

where $w(n)$ is the window function, $0 \leq n \leq N-1$, and N is the number of samples in each frame.

In this paper Hamming window is used, which has the form [2]:

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1. \tag{2}$$

C. Fast Fourier Transform

FFT is performed on the windowing signal to convert it into frequency domain. The discrete Fourier Transform (DFT) over a discrete signal $x(n)$ of N samples, converts each frame of N samples into the frequency domain from its time domain. The FFT is the fast algorithm to implement DFT, which is defined as

$$X_k = \sum_{n=0}^{N-1} x(n)e^{\frac{-j2\pi kn}{N}}, \quad k = 0, 1, 2, 3, \ldots, N-1. \tag{3}$$

The result after this step is often referred to as spectrum.

D. Mel Frequency Warping

The Mel-frequency scale is linearly spaced for frequencies below 1000 Hz and logarithmically spaced above 1000 Hz. A pitch of 1000 Hz tone, which is equivalent to 1000 mels when it is above the perceptual hearing threshold level by 40 dB. The following approximate formula can be used to compute the mels for a given frequency f in Hz [10]:

$$mel(f) = 2595 * \log 10\left(1 + \frac{f}{700}\right). \tag{4}$$

The mel filter bank structure has overlapped triangular filters. The cut off frequency of each filter can be computed using center frequencies of two adjacent filters, as shown in Fig. 2. Center frequencies are linearly spaced on mel scale with fixed bandwidth. The output of each filter is the sum of its filtered spectral components. Filter bank mel scale is shown in Fig. 2.

E. Mel Frequency Cepstrum Coefficient (MFCC)

MFCCs are coefficients that represent audio based on perception with their frequency bands logarithmically positioned and mimics the human vocal response.

III. Decision Making

A. Mean of MFCC

Mean for a data set is termed as arithmetic mean. In this paper, mean of MFCCs is taken to reduce the huge set of values which are obtained from MFCC. When 'x' $= \{x1, x2, x3, \ldots, xn\}$ represents MFCC values and total number of MFCCs is n then the arithmetic mean is taken as,

$$x = \frac{x1 + x2 + x3 + \cdots + xn}{n}. \tag{5}$$

B. Standard Deviation

The standard deviation (σ) is a measure of variation or dispersion of a set of data from their mean. A smaller standard deviation indicates that the data to be very close to their mean and a high standard deviation indicates that the data are spread out from their mean and among themselves [8]. In this paper standard deviation of 60 speakers for different emotions, namely sad, happy and anger, are experimentally computed and range of standard deviations for these emotions are optimized as follows [9]: Sad: 0.1700 to 0.2199; Happy: 0.2200 to 0.2899; Angry: 0.2900 to 0.3999.

IV. Results and Discussions

A data base has been generated with voices of 60 people with different emotions. Speaker's speech signal was read using the function 'wavread' in MATLAB [11]. The speech signal further made to undergo framing, after which they were passed through 'Hamming window'. Fast Fourier Transform was performed on the signal. This was followed by converting to Mel scale, after which the Mel Frequency Cepstral Coefficients were obtained. Mean value of MFCC was found and the standard deviation for the mean value was found, and this value was passed through "if else" statement, where the obtained standard deviation of that particular emotion is compared with the optimized values of standard deviation for different emotions, and the corresponding emotion were displayed.

Response of the designed system for different emotions is as follows:

A. Sad emotion

See Figs. 3, 4, 5 and 6.

B. Happy emotion

See Figs. 7, 8, 9 and 10.

C. Angry emotion

The experiment was repeated for various levels of input noises in terms of SNR. It was observed that overall efficiency proved to remain 80% even at lower SNR values (see Figs. 11, 12, 13 and 14).

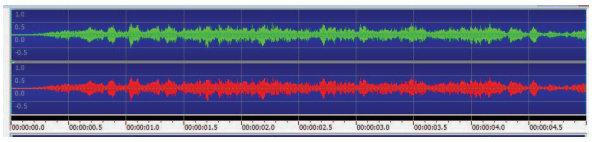

Fig. 3. Input signal for sad emotion using Goldwav converter.

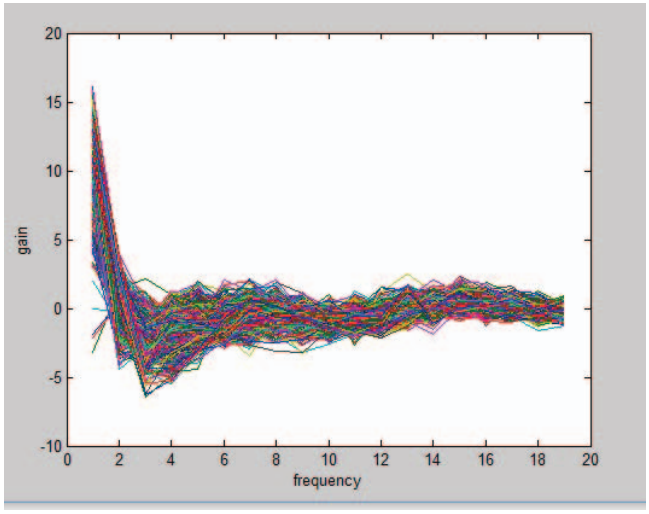

Fig. 4. MFCC for sad emotion.

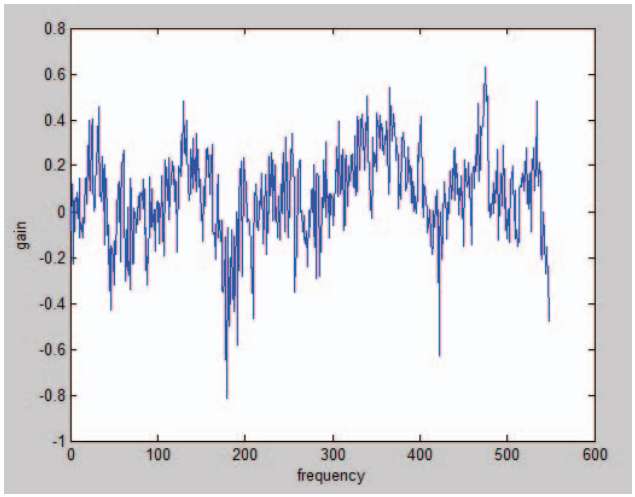

Fig. 5. Mean for sad emotion.

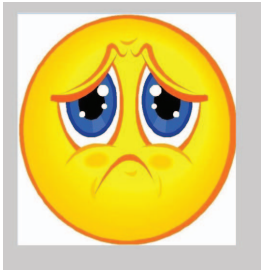

Fig. 6. Output for sad emotion.

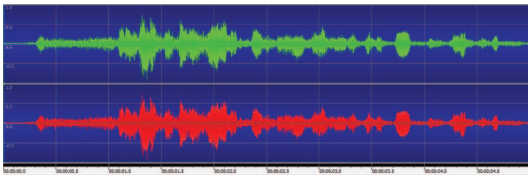

Fig. 7. Input signal for happy emotion using Goldwav converter.

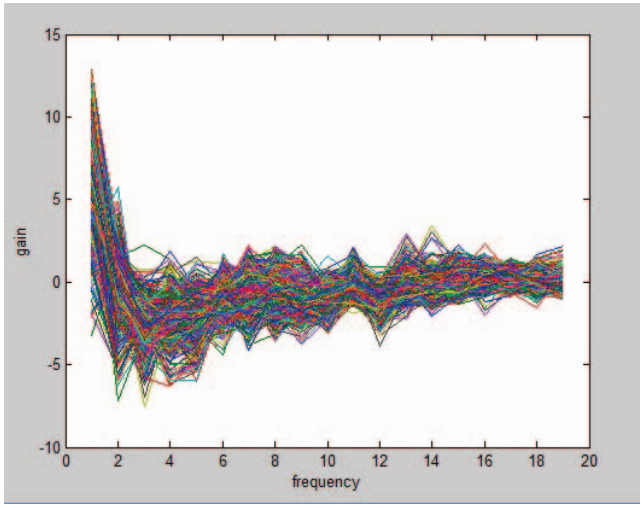

Fig. 8. MFCC for happy emotion.

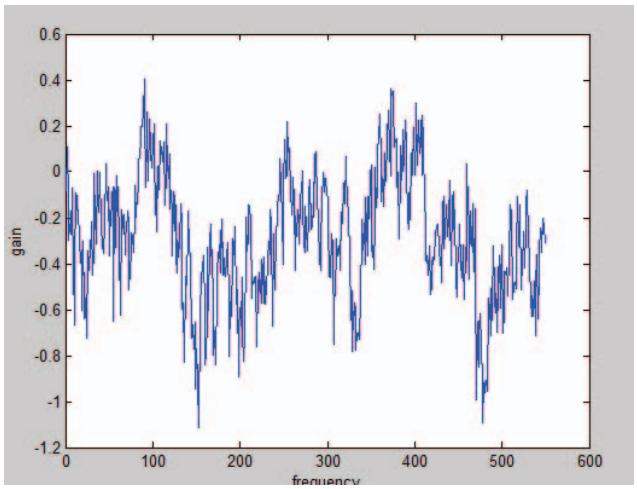

Fig. 9. Mean for happy emotion.

Fig. 10. Output for happy emotion.

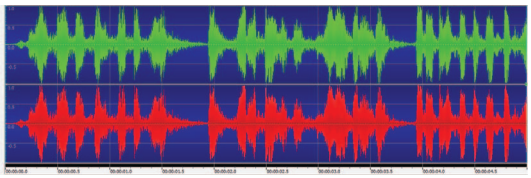

Fig. 11. Input signal for angry emotion using Goldwav converter.

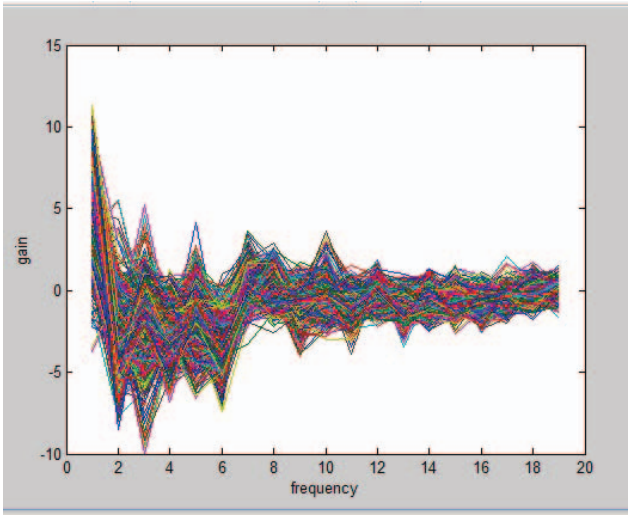

Fig. 12. MFCC for angry emotion.

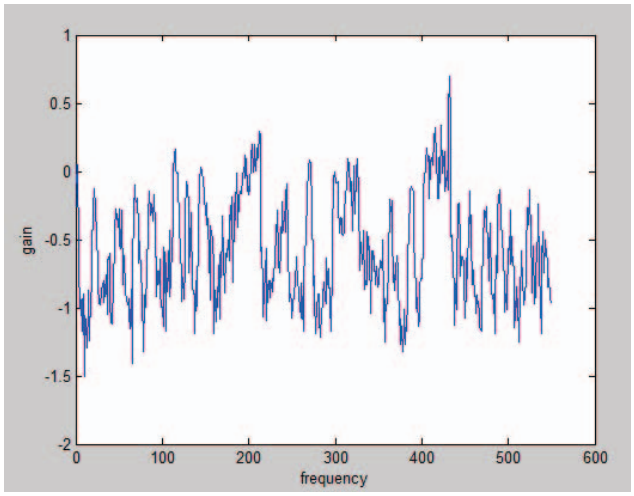

Fig. 13. Mean for angry emotion.

Fig. 14. Output for angry emotion.

TABLE I
EXPERIMENTAL RESULTS OF THE EMOTION RECOGNITION SYSTEM.

Serial no.	Standard deviation	Expected emotion	Emotion obtained
1.	0.2287	Happy	Happy
2.	0.2286	Happy	Happy
3.	0.3065	Happy	Angry
4.	0.2793	Happy	Happy
5.	0.2710	Happy	Happy
6.	0.3155	Angry	Angry
7.	0.3312	Angry	Angry
8.	0.3898	Angry	Angry
9.	0.3233	Angry	Angry
10.	0.2067	Angry	Sad
11.	0.1806	Sad	Sad
12.	0.1906	Sad	Sad
13.	0.1951	Sad	Sad
14.	0.2095	Sad	Sad
15.	0.2468	Happy	Happy

V. CONCLUSIONS

This method of speech emotion recognition has proven to be 80% efficient, as shown in Table I. This efficiency in performance continued even in noisy environment. Hence this system can serve as noise robust emotion recognition system. Such efficiency in noisy environment extends the scope of the work wherein emotion recognition systems can be utilized in military.

REFERENCES

[1] Kunxia Wang, Ning An, Bing Nan Li, Yanyong Zhang, and Lian Li, "Speech emotion recognition using fourier parameters," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 69–75, January–March 2015.

[2] Dipti D. Joshi and M. B. Zalte, Recognition of emotion from marathi speech using MFCC and DWT algorithms," *International Journal of Advanced Computer Engineering and Communication Technology (IJACECT)*, Issue-2, 2013, vol. 2, no. 2, pp. 59–63, 2013.

[3] R. Subhashree and G.N. Rathna, "Speech emotion recognition: performance analysis based on fused algorithms and GMM modelling," *Indian Journal of Science and Technology*, vol. 9, no. 11, pp. 1–8, Mar. 2016. DOI: 10.17485/ijst/2016/v9i11/88460.

[4] A. Milton, S. Sharmy Roy, and S. Tamil Selvi, "SVM scheme for speech emotion recognition using MFCC feature," *International Journal of Computer Applications (0975–8887)*, vol. 69, no. 9, pp. 34–39, May 2013.

[5] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray, Survey on speech emotion recognition: Features, classification schemes, and databases,

[6] Ozan Mut and Mehmet Göktürk, "Improved weighted matching for speaker recognition," in *Proceedings of World Academy of Science, Engineering and Technology*, Apr. 2005, vol. 5, pp. 229–231.

[7] Lindasalwa Muda, Mumtaj Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," *Journal of Computing*, vol. 2, no. 3, pp. 138–143, Mar. 2010.

[8] S. Demircan and H. Kahramanlı, "Feature extraction from speech data for emotion recognition," *Journal of Advances in Computer Networks*, vol. 2, no. 1, pp. 28–30, Mar. 2014.

[9] Shashidhar G. Koolagudi, SudhamayMaity, Vuppala Anil Kumar, SaswatChakrabarti, and K. SreenivasaRao, "IITKGP-SESC:Speech Database for emotion analysis," S. Ranka et al.(Eds.): *IC3 2009, CCIS 40*, pp. 485–492, 2009. Berlin Heidelberg: Springer-Verlag, 2009.

[10] Shilna Sasheendran, M. J. Spoorthi, A. Upendra Raju, B. C. Shalini, and V. Uma, "Speaker identification and verification using MFCC and VQ methods," *International Journal of Scientific Engineering and Technology Research*, vol. 3, no. 1, pp. 163–170, Jan. 2014.

[11] www.mathworks.com.