Multi-algorithm Fusion for Speech Emotion Recognition

Gyanendra K Verma, U. S. Tiwary and Shaishav Agrawal

Indian Institute of Informaiton Technology, Allahabad, India
Allahabad, India - 211012
{gyanendra, ust}@iiita.ac.in, shaishav.engr@gmail.com

Abstract. In this paper, we have proposed a speech emotion recognition system based on multi-algorithm fusion. Mel Frequency Cepstral Coefficients (MFCC) and Discrete Wavelet Transform (DWT), the two prominent algorithms for speech analysis, have been used to extract emotion information from speech signal. MFCC, a representation of the short-term power spectrum of a sound is a classical approach to analyze speech signal whilst the DWT, a multiresolution approach mainly approximate the frequency information along with time information. Feature level fusion of algorithms has been performed after extraction of features by acoustic analysis of speech emotion signal. The final emotion state was determined by classification using Support Vector Machine. Popular Berlin emotion database is used for evaluation of the proposed system. The results achieved are very promising as the proposed fusion algorithm performed well compared to individual algorithms.

Keywords: Multi-algorithm Fusion, MFCC, DWT, Speech Emotion Recognition

1 Introduction

Natural human-human interactions are made in two ways: verbal which includes speaking, singing and tone of voice, and Non-verbal which involves facial expression, body language, sign language, touch, eye contact etc. Emotion can be present in both types of communication i.e. verbal and non-verbal. Emotion helps to improve the level of interaction. Emotion is expressed in several ways: facial expressions, body gesture and speech are few of them. Recognizing emotion from human speech is an interesting and challenging problem for researchers working in the field of human computer interaction. Emotion recognition is one of the prime factors of Human-computer Interaction (HCI).

Fundamental frequency is an important voice feature for emotion recognition from speech signal [1] and can be easily extracted. Vocal expressions are conveyed by prosodic features, which include the fundamental frequency, intensity and rhythm of the voice [2]. Other important attributes that can be used for emotion recognition are pitch intensity and spectral measures. [3]. Some attributes are very proficient in particular emotion state, such as pitch is good to capture anger and fear whereas intensity is good for happiness. The success of the emotion recognition system depends upon the features incorporating different emotion states of the speaker.

We have already proposed [3] a novel approach for emotion recognition based on relative amplitude of speech signal. Many researchers have used wavelet based approach [4, 5] for emotion recognition [6, 7] used MFCC features for emotion recognition. Liqin Fu et. al. [8] utilized temporal features i.e. pitch, amplitude energy, energy frequency value, and zero cross ratio and formant frequency.

We propose a multi-algorithm approach for emotion recognition from speech signal. The MFCC and Discrete Wavelet Transform based algorithms have been successfully used to extract emotional information from speech signal. MFCC is a classical approach to analyze speech signal. It represents the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency [9]. In the other approach approximation and detail coefficients were calculated by decomposing the input speech signal using DWT. The wavelet features for each input speech signal are obtained from 4th and 6th level decomposition using db4 wavelets. The similarity between the extracted features and a set of reference features is calculated by K-NN and SVM classifiers. We have used Berlin emotion database to evaluate the proposed system. The results obtained from fusion are better as compared to the separate performance reported in the literature [4, 5, 6, 7].

The rest of the paper is organized as follows: MFCC and wavelet transform algorithms are described in section II. Multi-algorithm fusion is explained in section III. Finally the experiment and the result analysis are given in section V and concluding remarks are given in section VI.

2 Our Approach

We have been used wavelet transform and Mel Frequency Cepstral Coefficient (MFCC) algorithms to recognize emotion from speech signal. The emotion recognition system based on multi-algorithm fusion is given in figure1.

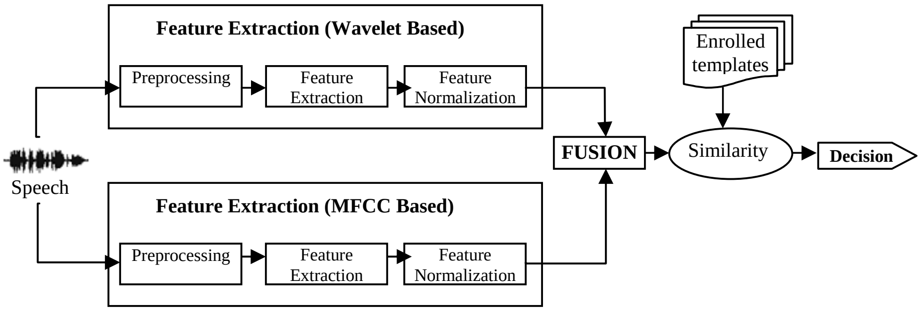

Fig. 1. Multi-algorithm fusion emotion recognition system

2.1 Wavelet Transform

Wavelet transform provides a compact representation that depicts the energy distribution of the signal in time and frequency domain [10]. We have used discrete wavelet transform to decomposes the signal into multilevel successive frequency bands utilizing two sets of function called scaling function $\boldsymbol{\phi}$ and wavelet functions ($\boldsymbol{\psi}$) associated with low pass and high pass filters respectively [11]. Information captured by wavelet transform depends on properties of wavelet function family like Daubechis, Symlet, Biorthogonal, Coiflet etc. and properties (waveform) of the target signal. Information extracted by wavelet transforms using different family of wavelet function need not be same. It is required to choose or evaluate the wavelet function that provides more useful information for particular application.

In Discrete wavelet decomposition of signal, the output of high pass filter and low pass filter can be represented mathematically by equation 1 and 2.

$$Y_{high}[k] = \sum X[n]g[2k-1] \tag{1}$$

$$Y_{low}[k] = \sum X[n]h[2k-1] \tag{2}$$

Where Y_{high} and Y_{low} are the outputs of the high band pass and low band pass filters respectively.

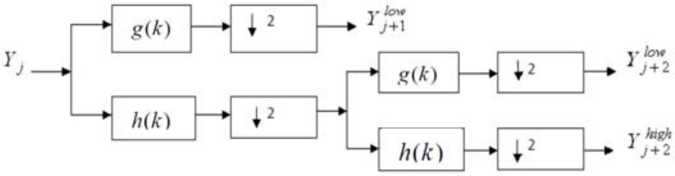

Fig. 2. Schematic of Discrete Wavelet decomposition of a speech signal.

2.2 Mel Frequency Cepstral Coefficients (MFCC) Algorithm

MFCC’s are based on the known variation of the human ear’s critical bandwidths with frequency. The technique is based on two types of filters, namely linearly spaced filter and logarithmically spaced filters. The phonetically important characteristics of speech can be captured by representing the signal at the Mel frequency scale. This scale has a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Normal speech waveform may vary from time to time depending on the physical condition of speakers' vocal cord. MFFCs are less susceptible to these variations [12].

The Algorithm

The process of calculating MFCC consists of the following steps:

- Framing: In this step the speech signal segmented into N samples with 25% overlapping frames.
- Windowing: this step is applied for spectral analysis of the speech signal. We have used Hamming window, given in equation 3.

$$W(n) = 0.54 - 0.46 \times \cos(2 \times \frac{n}{N-1}) \qquad (3)$$

Where 0 <= n <= N-1

- Fast Fourier Transform (FFT): FFT was applied on each frame to obtain the spectral information from the time domain signal.
- Mel-frequency: For a given frequency, the Mel scale is being calculated by equation 4.

$$Mel(f) = 2595 \times \log_{10}(1 + \frac{f}{700}) \qquad (4)$$

- Cepstrum: Finally the discrete cosine transform (DCT) was applied to the signal in order to obtain MFCC coefficients.

2.3 Fusion of Algorithms

Information fusion from different streams is a major challenge, specially an adaptive and optimal fusion of information. Information fusion can be at three levels: (i) signal level (ii) feature level (iii) decision level. In this study we performed feature level information fusion of both algorithms. The fusion of algorithms is illustrated in fig 3.

We get two feature vectors $F_{mfcc} = (f_1, f_2, ... f_m)$ and $F_{wav} = (f_1, f_2, ... f_n)$ by applying mfcc algorithm and wavelet transform respectively. The feature vectors having vector length *n* and *m*. Let $Fi = (f_1, f_2, ... f_{m+n})$ be the fused feature vector of feature vectors F_{mfcc} and F_{wav}. Fi is obtained by augmenting the normalized feature vectors F_{mfcc} and F_{wav}, then performing the feature selection on the concatenated vector formed.

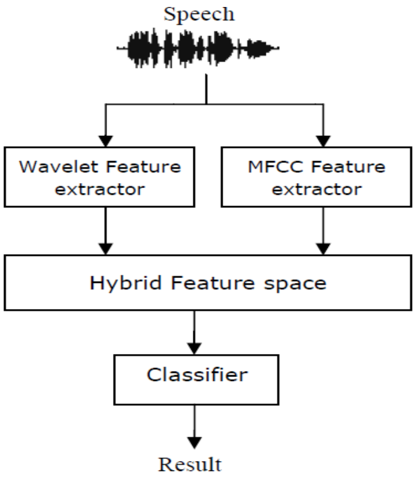

Fig. 3. Feature level fusion of algorithms

3 Feature Extraction

3.1 Feature Extraction from Wavelet Transform

The speech signal is passed into successive high pass and low pass filter in order to extract wavelet coefficients. Selection of suitable wavelet coefficients and the number of levels of decomposition is important. Daubechis wavelet family provides good results for non-stationary one dimensional signal analysis [13]. Therefore we have used Daubechis wavelet in this study. The feature vectors obtained at six level wavelet decomposition provides compact representation of the signal. The coefficients occur in the whole bandwidth from low to high. The original signal can be represented by the sum of coefficients in every sub band, which is cD6, cD5, cD4, cD3, cD2, cD1. Feature vectors are obtained from the detailed coefficients applying common statistics i.e. standard deviation, mean etc.

3.2 Feature Extraction from MFCC

For the short period of time the characteristics of the speech signal are fairly stationary, therefore the short-time spectral analysis is the most common way to characterize the speech signal [14]. The input emotion speech signal is segmented into a number of frames. Windowing operation is performed to capture the static property of the signal. Hamming window with 20 ms size and 25% overlapping has been used here. Then Fast Fourier transform is applied to produce the spectral characteristic of the speech signal. For the given frequency the mel frequency was calculated by equation 4. Finally the log Mel spectrum is converted back to time domain in order to get mel frequency cepstral coefficients (MFCC).

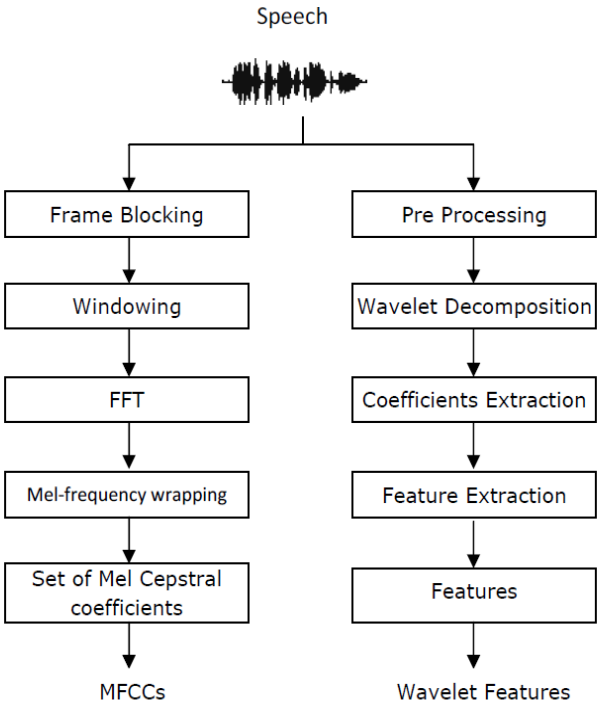

Fig. 4. Feature extraction process using MFCC and DWT

4 Feature Matching

K-NN algorithm has been used for feature matching between feature set (reference features) and query feature. The training set S contains $\boldsymbol{l}$ points $\{f_1,......, f_l\}$, $f_{xl} \in R^n$ and their corresponding class labels $\{y_1,.......y_N\}$, $\mathrm{y_i} \in c$, $c=\{1,.....N_c\}$ where N_c is the number of different classes.

The distance between two vectors X and Y having length n can be calculated by Euclidean distance, Manhattan distance etc.

Euclidean distance is defined as follow:

$$d(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \qquad (5)$$

5 Experiments and Results

The Berlin emotional dataset, recorded by speech workgroup headed by Prof. Dr. W. Sendlmeieris was used for experimental purpose [15]. The database accommodates spoken sentences of 10 actors (5 males and 5 females) in seven states of emotional references: Anger, Boredom, Disgust, Fear, Joy, Sad and Neutral. The actors have spoken 10 predefined sentences with each emotion. They are sampled at 16 KHz sampling frequency. We have rearranged the database by renaming the folder of each emotion state from one to seven for seven emotion classes. Then we also renamed the emotion files from one to thirty. All the experiments were performed on MatLab 7.0 platform.

The steps involved are Pre-processing, segmentation, coefficient extractions, feature vector generation, features normalization and classification. The feature vectors obtained from multi-algorithms has been normalized using min-max normalization for classification purpose. The normalization was applied into both feature vectors obtained from MFCC and wavelet transform. We have seven emotion classes and each class contains thirty emotion files. Twenty five emotion files were used for training and rest files were used for testing. The experiments were also performed individually by taking wavelet and mfcc features. A feature level fusion was also performed as discussed in section III. Each emotion samples of same emotion state was assigned same class i.e. twenty-five emotion samples of same emotion states were assigned the same class. In this way the whole dataset is grouped into seven classes for seven emotion states. Euclidean distance was used to calculate the distances among feature vectors.

The wavelet coefficients were formed by 6th level wavelet decomposition of emotion samples. General statistics was applied on wavelet coefficients in order to form feature vectors. Feature vectors are generated by applying common statistics described in section III. The performance results of different algorithms are shown in table1 and corresponding graph is illustrated in figure 5.

TABLE 1: CLASSIFICATION RESULTS

Sl.	Emotion	Wavelet	MFCC	Fusion
1	Anger Vs Boredom	0.10	0.15	0.10
2	Boredom Vs Disgust	0.30	0.25	0.20
3	Disgust Vs Fear	0.25	0.35	0.25
4	Fear Vs Happy	0.20	0.25	0.15
5	Happy Vs Neutral	0.25	0.34	0.20
6	Neutral Vs Sad	0.30	0.25	0.20

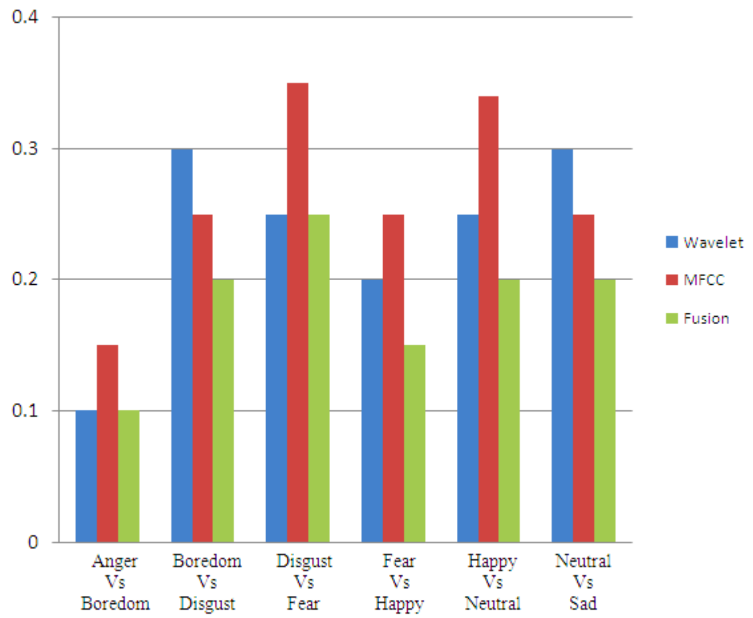

Fig. 5. Performance graph of different algorithms with fusion

6 Conclusion

In this study, we have proposed speech emotion recognition system based on MFCC and Wavelet Transform. Features were extracted using the proposed algorithm and evaluated using Berlin emotion speech database. A Comparative study is also performed here. The experimental results are very promising. A feature level fusion of algorithms was performed in this study.

References

1. Cohn, J.F., katz, G.S.: Bimodal expressions of emotion by face and voice. In: Workshop on face/gesture recognition and their applications, the sixth ACM international multimedia conference, Bristol, England (1998)
2. Fasel, B., Luettin, J.: Automatic facial expression analysis: A survey. In: Pattern Recognition, vol. 36, pp. 259–275 (2003)

3. Kudiri, K. M., Verma, G. K., Gohel, B.: Relative Amplitude based Features for Emotion Detection from Speech. In: 3rd IEEE Int. Conf. on Signal and Image Processing, pp. 301-304 (2010)
4. Rizon, Mohamed.: Discrete Wavelet Transform Based Classification of Human Emotions Using Electroencephalogram Signals. In: American Journal of Applied Sciences 7 (7): pp. 865-872 (2010)
5. Shah, Firoz et al.: Discrete Wavelet Transforms and Artificial Neural Networks for Speech Emotion Recognition. In: International Journal of Computer Theory and Engineering, Vol. 2, No. 3, 1793-8201, (2010)
6. Kwon, Oh-Wook.: Emotion Recognition by Speech Signals. In: EUROSPEECH-2003, Geneva, (2003)
7. Xia Mao.: Speech Emotion Recognition based on a Hybrid of HMM/ANN. In: Proceedings of the 7th WSEAS International Conference on Applied Informatics and Communications, Athens, Greece, August 24-26, (2007)
8. Liqin, Fu et al.: Relative Speech Emotion Recognition Based Artificial Neural Network. In: IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application, (2008)
9. http://en.wikipedia.org/wiki/Mel-frequency_cepstrum
10. Dutta, Tridibesh: Dynamic Time Warping Based Approach to Text Dependent Speaker Identification Using Spectrograms. In: Congress on Image and Signal Processing, Vol. 2, pp. 354-360 (2008)
11. Tzanetakis, George, Essl, Georg and Perry Cook.: Audio Analysis using the Discrete Wavelet Transform. In: Proc. Conf. in Acoustics and Music Theory Applications, Skiathos, Greece (2001)
12. Lindasalwa, M., Begam, M., Elamvazuthi, I.:Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. In: Jour. of Computing, vol. 2, Issu 3, pp. 138-143 (2010)
13. Toh, A.M., Togneri, R., Northolt, S.: Spectral entropy as speech features for speech recognition. In: The Proceedings of PEECS, Perth, pp. 22–25 (2005)
14. Kan, Phak Len Eh, Allen, Tim, and Quigley, F,: A GMM-Based Speaker Identification System on FPGA. In: 6th international symposium on Reconfigurable Computing: Architectures, Tools and Applications. Bangkok, Thailand march 2010, LNCS (2010)
15. Burkhardt, F, Paeschke, A.: A database of German emotional speech. In: Interspeech, (Lisbon, Portugal), pp. 1517-1520 (2005)