RECOGNITION OF EMOTION FROM MARATHI SPEECH USING MFCC AND DWT ALGORITHMS

Dipti D. Joshi, M.B. Zalte

(EXTC Department, K.J. Somaiya College of Engineering, University of Mumbai, India)
Diptijoshi3@gmail.com

***Abstract*- Field of emotional content recognition of speech signals has been gaining increasing interest during recent years. Several emotion recognition systems have been constructed by different researchers for recognition of human emotions in spoken utterances. In this paper we have proposed system to recognize emotions from Marathi Speech samples which is one of the Indian Language.**

The database for the speech emotion recognition system is the emotional speech samples and the features extracted from these speech samples are the energy, pitch, formants and Mel frequency cepstrum coefficient (MFCC). Discrete wavelet transform also used for feature vector extraction. The SVM classifier used to differentiate emotions such as anger, happiness, sadness and neutral state. The classification performance is based on extracted features is discussed for given emotions.

***Index Terms*- DWT, MFCC, Speech Emotion recognition, SVM**

I. INTRODUCTION

Speech is a complex signal which contains information about the message, speaker, language and emotions. Speech is produced from a time varying vocal tract system excited by a time varying excitation source. Emotion on other side is an individual mental state that arises spontaneously rather than through conscious effort. [8] There are various kinds of emotions which are present in a speech. The basic difficulty is to cover the gap between the information which is captured by a microphone and the corresponding emotion, and to model the specific association. This gap can be bridge by narrowing down various emotions in few, like anger, happiness, sadness, and neutral. Emotions are produced in the speech from the nervous system consciously, or unconsciously. Emotional speech recognition is a system which basically identifies the emotional as well as physical state of human being from his or her voice [1]. Emotion recognition is gaining attention due to the widespread applications into various domains: detecting frustration, disappointment, surprise/amusement etc.

There are many approaches towards automatic recognition of emotion in speech by using different feature vectors. A proper choice of feature vectors is one of the most important tasks. The feature vectors can be distinguished into the following four groups: continuous (e.g., energy and pitch), qualitative (e.g., voice quality), spectral (e.g., MFCC), and features based on the Teager energy operator (e.g., TEO autocorrelation envelope area [2]). For classification of speech, methodologies followed are: HMM, GMM, ANN, k-NN, and several others as well as their combination which maintain the advantages of each classification technique. After studying the related literature it can be identified that the feature set which is mostly employed is comprised of pitch, MFCCs, and HNR. Additionally, the HMM technique is widely used by the researchers due to its effectiveness. Feature extraction by temporal structure of the low level descriptors or large portion of the audio signal is taken could be helpful for both the modeling and classification processes.

In this paper we propose a speech emotion recognition system using multi-algorithm approach. 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 linear cosine transform of a log power spectrum on a nonlinear Mel Scale of frequency. 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 4^{th} level decomposition using db4 wavelets. Similarity between the extracted features and a set of reference features is calculated by SVM classifier.

Success of any Speech Emotion recognition system depends on naturalness of database used. Here we have created a Marathi corpus database to evaluate a proposed system.

The rest of the paper is organized as follows: Basic speech emotion recognition system is given in section II, the methodology of the study is provided in section III, which is followed by result and discussion in section IV, and finally concluding remarks are given in section V.

II. SPEECH EMOTION RECOGNITION SYSTEM

Speech emotion recognition aims to automatically identify the emotional state of a human being from his or her voice. It is based on in-depth analysis of the generation mechanism of speech signal, extracting some features which contain emotional information from the speaker's voice, and taking appropriate pattern recognition methods to identify emotional states. [3]

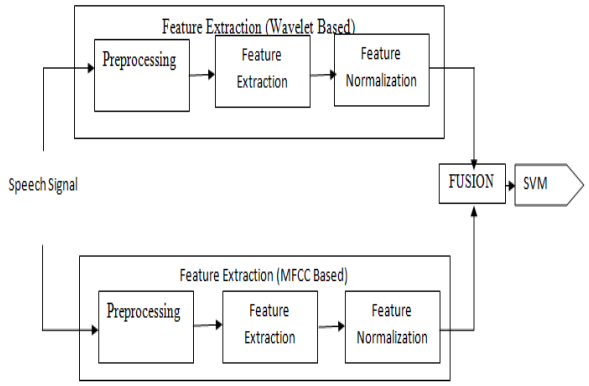

Fig.1 Proposed Speech Emotion Recognition System

Like typical pattern recognition systems, speech emotion recognition system contains four main modules: Emotional speech input, feature extraction, feature labeling, classification, and output emotional class. [4] Multi- algorithm approach is proposed for emotion recognition from speech signal. The MFCC and Discrete Wavelet Transform based algorithms proposed to be used to extract emotional information from speech signal. Fig.1 shows proposed system.

III. METHODOLOGY

A. Feature Extraction of MFCCs

The first purpose to explore the spectral features by using the Mel-frequency cepstral coefficients (MFCCs) is that they have been widely employed in speech recognition due to superior performance when compared to other features. [1] The Mel-frequency cepstrum is a representation of 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. For each speech frame of 30 ms, a set of Mel-frequency cepstrum coefficients was computed. Fig. 2 shows the MFCC feature extraction process containing following steps:

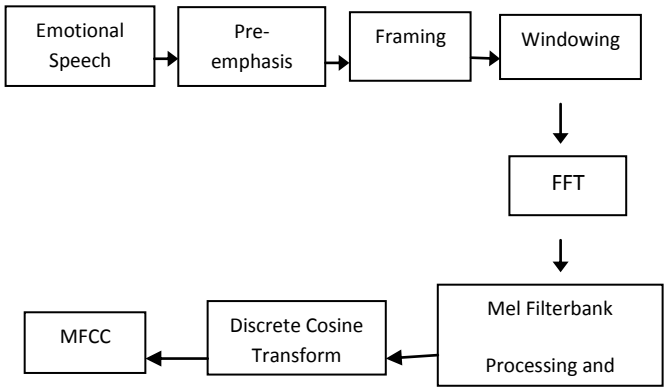

Fig.2 Mel- Frequency Cepstral Coefficients

Step 1: Pre–emphasis

This step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

$$Y[n] = X[n] * 0.95\ X\ [n-1] \quad \ldots\ldots (1)$$

Step2 Framing: It is a process of segmenting the speech samples obtained from the analog to digital conversion (ADC), into the small frames with the time length within the range of 20-40 ms. Framing enables the non stationary speech signal to be segmented into quasi-stationary frames, and enables Fourier Transformation of the speech signal. It is because, speech signal is known to exhibit quasi-stationary behaviour within the short time period of 20-40 ms.

Step3: Windowing: Windowing step is meant to window each individual frame, in order to minimize the signal discontinuities at the beginning and the end of each frame. Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. The Hamming window equation is given as:

If the Hamming window is defined as W (n), $0 \le n \le N-1$ where,

N = number of samples in each frame, Y[n] = Output signal, X (n) = input signal

W (n) = Hamming window, then the result of windowing signal is shown below:

$$Y(n) = X(n)\ \ X\ \ W(n)$$

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) 0 \le n \le N-1 \quad \ldots.. (2)$$

Step4: FFT: FFT converts each frame of N samples from the time domain into the frequency domain. The Fourier Transform is to convert the convolution of the input pulse and the vocal tract impulse response in the time domain. This statement supports the equation below:

$$Y(w) = FFT[h(t) * X(t)] = H(w) * X(w) \quad \text{.....(3)}$$

Step5: Mel Filter bank and Frequency wrapping: The mel filter bank consists of overlapping triangular filters with the cut off frequencies determined by the center frequencies of the two adjacent filters. The filters have linearly spaced centre frequencies and fixed bandwidth on the mel scale. Then, each filter output is the sum of its filtered spectral components. Following equation is used to compute the Mel for given frequency f in HZ:

$$F(Mel) = [2595 * \log 10 [1+F] 700] \quad \text{.....(4)}$$

Step6: Discrete Cosine Transform: It is used to orthogonalise the filter energy vectors. Because of this orthogonalization step, the information of the filter energy vector is compacted into the first number of components and shortens the vector to number of components.

B. Feature Extraction By Using Discrete Wavelet Transforms:

The wavelet transform theory provides an alternative tool for short time analysis of quasi stationary signal such as Speech as opposed to traditional transforms like FFT. DWT is the most promising mathematical transformation which provides both the time – frequency information of the signal and is computed by successive low pass filtering and high pass filtering to construct a multi resolution time frequency plane [7].

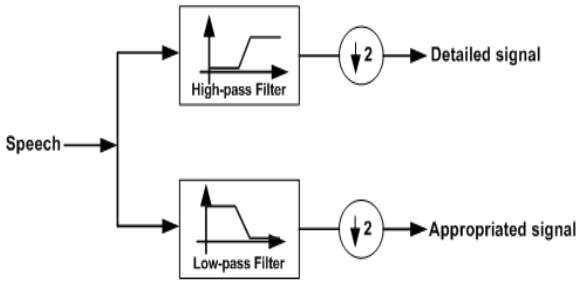

Fig 3 Discrete Wavelet Transform

In DWT a discrete signal x[k] is filtered by using a high pass filter and a low pass filter, which will separate the signals to high frequency and low frequency components. To reduce the number of samples in the resultant output we apply a down sampling factor of ↓2. The Discrete Wavelet Transform is defined by the following equation.

$$W(j,k) = \sum_j \sum_k X(k) 2^{-j/2} \psi(2^{-j} n - k) \quad \text{..... (5)}$$

Where Ψ(t) is the basic analyzing function called th e mother wavelet. The digital filtering technique can be expressed by the following equations:

$$Y_{high}[K] = \sum nX[n]\, g[2k-1] \quad \text{..... (6)}$$

$$Y_{low}[k] = \sum nX[n]\, h[2k-1] \quad \text{..... (7)}$$

Where Y high and Y low are the outputs of the high pass and low pass filters. The speech signals can be decomposed in sub bands by using the Discrete Wavelet Transform (DWT) 4 iterations. For each sub band, the mean energy was calculated. The original signal can be represented by the sum of coefficients in every sub band, which is cD4, cD3, cD2, cD1. Feature vectors are obtained from the detailed coefficients applying common statistics i.e. standard deviation, mean etc.

C. Classifier

Support Vector Machines (SVMs) are a relatively new learning algorithms introduced by Vapnik (1995). Support vector machines are built by mapping the training patterns into a higher dimensional feature space where the points can be separated by using a hyperplane. In SVM approach, the main aim of an SVM classifier is obtaining a function f(x), which determines the decision boundary or hyperplane. This hyper-plane optimally separates two classes of input data points. This hyperplane is shown in Fig 4 Where M is margin, which is the distance from the hyperplane to the closest point for both classes of data points.[5]

In SVM, the data points can be separated two types: linearly separable and non-linearly separable. For a linearly separable data points, a training set of instance-label pairs (Xk , Yk), k =1,2,3, ... , t where $X_k \in Rn$ and $Y_k \in \{+1, -1\}$, the data points can be classified as :

$$\begin{aligned} \langle w \cdot x_k \rangle + b_0 &\geqslant 1, \quad \forall y_k = 1, \\ \langle w \cdot x_k \rangle + b_0 &\leqslant -1, \quad \forall y_k = -1, \end{aligned} \quad \text{..... (9)}$$

Where (W . X_k) shows the inner product of W and X_k. The inequalities in Eq. (2) can be combined as in

$$y_k[\langle w \cdot x_k \rangle + b] - 1 \geqslant 0, \quad \forall k = 1, \ldots, t \quad \text{.. (10)}$$

The SVM classifier places the decision boundary by using maximal margin among all possible hyper planes. For maximize the M, ||w|| has to minimize subject to conditions given as:

$$\min \frac{\|w\|^2}{2} \quad \text{subject to} \quad \forall_k, y_k(\langle w \cdot x_k \rangle + b) \geqslant 1 \quad \text{..... (11)}$$

This optimization problem is named as quadratic optimization problem. For solving of this problem, the Lagrange function is used. Here, aim of this optimization is to find the appropriate Lagrange multipliers (ak)

$$L(w,b,a) = \frac{1}{2}w^{T}.w - \sum_{k=1}^{G} a_k[y_k(\langle w \cdot x_k \rangle + b) - 1], \quad \forall_k, a_k \geqslant 1, \qquad \ldots\ldots (12)$$

Where w represents a vector that defines the boundary, X_k are input data points, b represents a scalar threshold value. a_k is Lagrange multiplier and it must be $a_k \geq 0$. For appropriate a_k, L function should be minimized with respect to the w and b variables. At the same time, this function should be maximized with respect to the non-negative dual variable a_k. Once the a_k Lagrange multiplier is calculated the appropriate w and b parameters of optimal hyperplane are estimated with respect to a_k. The acquired optimal hyperplane f(x) can be given as below:

$$f(x) = \sum_{k=1}^{t} y_k a_k \langle x_k . x \rangle + b \qquad \ldots\ldots\ldots (13)$$

If Xk input data point has a non-zero Lagrange multiplier (ak), this Xk is called support vector. Using of the data points outside support vectors is not necessity for calculating the f(x).[5]

IV. EXPERIMENTS AND RESULT

Database Creation:

Our data is from the Emotional Prosody Speech corpus. This corpus contains Marathi continuous sentences uttered in 6 emotions i.e. happiness, anger, neutral, fear, sadness, boredom produced by 10 speakers (5 female, 5 male). A large corpus of sentences is recorded for the analysis of various emotions. A set of 10 sentences uttered three times given to each speaker for all the six emotions.

Recording: Recording is done using an Electrat microphone in a partially sound treated room with "PRAAT" software with sampling rate 16 kHz/16 bit and distance between mouth and microphone was adjusted nearly 30 cm.

Listening Test: Files of continuous sentences of one speaker were presented to 10 naive listeners to evaluate the emotions within six categories: neutral, happiness, anger, sadness, fear and boredom and the same process were repeated for all 10 speakers. All the listeners were educated and of age group of 18 to 28 years. Only those sentences, whose emotions were identified by at least 80% of all the listeners, were selected for this study.

From these emotional speech samples 30 samples of each emotion separated for training and 10 samples selected for testing purpose.

Feature extracted using two algorithms that is MFCC and DWT. Steps followed in feature extraction are nothing but Pre-processing, Segmentation, Coefficient extraction, Feature vector generation, Features normalization and classification. Marathi emotional speech database contains six emotional classes and each emotional class contains 40 emotional samples. Experiments were performed individually on each feature extraction algorithm by considering training and testing samples.

Mel frequency cepstral coefficients (MFCCs) computed by using MATLAB 7.0 from speech signal given in vector S and sampled at FS (Hz). The speech signal is first preemphasised using a first order FIR filter with preemphasis coefficient ALPHA. The preemphasised speech signal is subjected to the short-time Fourier transform analysis with frame durations of TW (ms), frame shifts of TS (ms) and analysis window function given as a function handle in WINDOW. This is followed by magnitude spectrum computation followed by filterbank design with M triangular filters uniformly spaced on the mel scale between lower and upper frequency limits given in R (Hz). The filterbank is applied to the magnitude spectrum values to produce filterbank energies (FBEs) (M per frame). Log-compressed FBEs are then decorrelated using the discrete cosine transform to produce cepstral coefficients. Fig indicates Log Mel filterbank energies.

Final step applies sinusoidal lifter to produce liftered MFCCs also returns FBEs and windowed frames, with feature vectors and frames as columns. Output obtained after MFCC extraction algorithm is column wise feature vectors.

The wavelet coefficients were formed by 4th 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. Each extracted feature is stored in a database along with its class label. Though the SVM is binary classifier it can be also used for classifying multiple classes. Each feature is associated with its class label e.g. angry, happy, sad, neutral, fear and boredom. It shows the variations that occur when emotion changes. By using MFCC and DWT algorithm feature is extracted from which we can observe how changes occur in pitch, frequency and other features when emotion changes. By using SVM classifying algorithm we can classify different emotions as described in section III.

The performance result of the system is shown in table1 and corresponding graph is illustrated in figure 4.

Emotion	Emotion Recognition %					
	Anger	Boredom	Fear	Happy	Sad	Neutral
Angry	75	0	0	0	0	25
Boredom	0	84	0	0	16	0
Happy	16	0	16	67	0	0
Sad	0	0	0	11	89	0
Neutral	0	12	12	12	0	64

Table1: % Recognition of SVM classifier

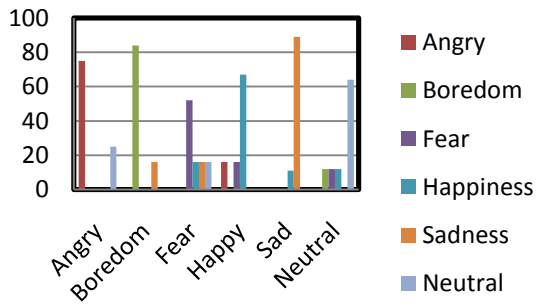

Fig.4 Performance graph of % Recognition of SVM classifier

CONCLUSION

In this paper MFCC and Discrete Wavelet Transform based algorithms proposed to extract emotional information from Marathi Emotional speech Database. The MFCC feature extraction procedure and wavelet based feature carried out and the obtained results noted down. Similarity between the extracted features and a set of reference features is calculated by using SVM classifier. The performance for SVM model is evaluated by % recognition rate and by corresponding graph which gives classification of emotions.

The experimental results shows the performance of system for Marathi emotional speech database which is promising.

REFERENCES

[1] Jia Rong, Gang Li , Yi-Ping Phoebe Chen "Acoustic feature selection for automatic emotion recognition from speech", Elsevier , Information Processing and Management volume 45 (2009)

[2] M. E. Ayadi, M. S. Kamel, F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases", Pattern Recognition 44, PP.572-587, 2011.

[3] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. "Emotional speech: towards a new generation of databases." Speech Communication, 40:33–60, 2003.

[4] Simina Emerich, Eugen Lupu, "Improving speech emotion recognition using frequency And time domain acoustic features", Proceeding of SPAMEC 2011, cluj- Napoca, Romania.

[5] P.Shen, Z. Changjun, X. Chen, "Automatic Speech Emotion Recognition Using Support Vector Machine", International Conference On Electronic And Mechanical Engineering And Information Technology, 2011

[6] Stavros Ntalampiras and Nikos Fakotakis, 'Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition', IEEE transactions on affective computing, vol. 3, no. 1, january-march 2012

[7] S. Mallat, "A wavelet tour of signal processing", NewYork, Academic Press, 1999

[8] Lan McLoughlin, "Applied Speech And Audio Processing", Cambridge University Press, 2009.

◈◈◈