Towards Detecting Emotions from Real Time Speech

Dipti H. Kale
VIVA Institute of Technology
Shirgaon, Virar(E)

Chitra N. Takle
VIVA Institute of Technology
Shirgaon, Virar(E)

Shoeb S. Shaikh
VIVA Institute of Technology
Shirgaon, Virar(E)

Reshma Chaudhari
VIVA Institute of Technology
Shirgaon, Virar(E)

ABSTRACT
Fundamental factor in communication of humans is nothing but Emotions. It would be ideal to have human emotions automatically recognized by machines, mainly for improving human machine interaction. Most of work in field of emotion recognition is done using recorded or offline database. Very selective research work is carried in real time high performance emotion recognition. In application of human computer interaction Real-time high performance emotion recognition is necessary for analyzing and responding to the user's emotions while he or she is interacting with an application. The proper choices of features and classifiers are important for a real-time high performance emotion recognition system. In this paper real time emotion recognition system is proposed, which extracts the emotions from real time speech based on extracting prosody, quality and dynamic features, classification of emotions using Multidimensional SVM and testing real time speech samples with training databases with emotional speech in 'Native Marathi' language has been presented

Keywords
MFCC, Prosody features, Quality features, Speech Emotion Recognition, Support Vector Machine

1. INTRODUCTION
Everyday activities of human being such as communication, learning and decision-making are impacted by Emotions. The ways of expressing Emotions are mainly speech, facial expressions, gestures and other non-verbal clues. Speech emotion recognition systems deals with analyzing the vocal behavior of a person while focus on the non-verbal aspects of speech. Discovering which features are indicative of emotional states and extracting them can be a difficult task. Emotions which are observed in uttered speech also reflected in mental and physiological state of a person. In processing of the generated speech, different features can be estimated, which can be utilized to learn the relationship between features and emotions [1]. Once relationship between generated speech and the emotion contents is learned, one can calculate the features and then automatically recognize the emotions present in speech. The most important challenge in Speech emotion recognition is the identification of speech features (prosodic, spectral and voice quality) contributing to the emotional behavior. Many features for emotion recognition from speech have been explored, but there is still no agreement on a fixed emotional state and some

quantifiable parameters of speech [1]. As per the literature survey most of the researchers use standard databases for SER systems so it is needed to work on the variation in emotion recognition in the real time speech. The classifier performance provides the efficiency to SER [8][9]. Therefore a system needed to develop which will extract powerful features and classification should also be accurate.

In this paper, real time speech emotion recognition system is proposed which uses combination of prosody features (i.e. pitch, energy, Zero crossing rate)[3], quality features (i.e. Formant Frequencies, Spectral features etc.)[3][5], derived features ((i.e.) Mel-Frequency Cepstral Coefficient (MFCC) [7], for robust automatic recognition of speaker's emotional states. Multilevel SVM classifier is used for identification of six discrete emotional states namely angry, fear, happy, neutral, sad and surprise in ' native Marathi Language'. The overall experimental results can be demonstrated using MATLAB simulation.

2. THE NATURE OF EMOTION IN SPEECH
An agreement on set of the important emotions to be classified by an automatic emotion recognizer is important issue of Speech emotion recognition systems. A typical set contains 300 emotional states. However, classifying such a large number of emotions is very difficult. According to pallet theory, any color is a combination of some basic colors. Similarly any emotion can be decomposed into primary emotions. Primary emotions are Anger, Disgust, Fear, Joy, Sadness, and Surprise. [1][2] These emotions are the most obvious and distinct emotions in our life. A highly qualitative correlation between emotion and some speech features is presented in Table 1.

According to Table 1 prosody and voice quality are most important features which distinguishes between emotions according to human perception. In particular, pitch and intensity seem to be correlated to activation. It can be observed that high pitch and intensity values imply high, low pitch and intensity values low activation. The automatic recognition of emotion seems straight-forward when looking at Table 1 However, when examining closer different studies on acoustic correlates of emotions of multiple authors, often contradicting results can be found. This is partly due to different variants of certain emotions such as hot and cold anger, but as well to the intrinsically great variability of emotional expressions. Thus, there is just no direct mapping between acoustics and emotions.

Table 1. General correlates to emotion in speech.

	Anger	Happiness	Sadness	Fear	Disgust
Speech Rate	Slightly faster	Faster or Slower	Slightly Slower	Much faster	Very much slower
Pitch average	Very much higher	Much higher	Slightly slower	Very much higher	Very much slower
Pitch range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	lower
Voice quality	Breathy, chest tone	Breathy, blaring	Resonant	Irregular voicing	Grumbled chest tone
Pitch change	Abrupt on stressed syllables	Smooth upward inflections	Downward inflections	Normal	Wide downward terminal inflections
Articulation	Tense	Normal	Slurring	Precise	Normal

3. REAL TIME EMOTION RECOGNITION SYSTEM

The speech emotion recognition system contains five main modules emotional speech input, Pre-processing, feature extraction, feature normalization, classification, and recognized emotional output with respect to training and testing phase [4].

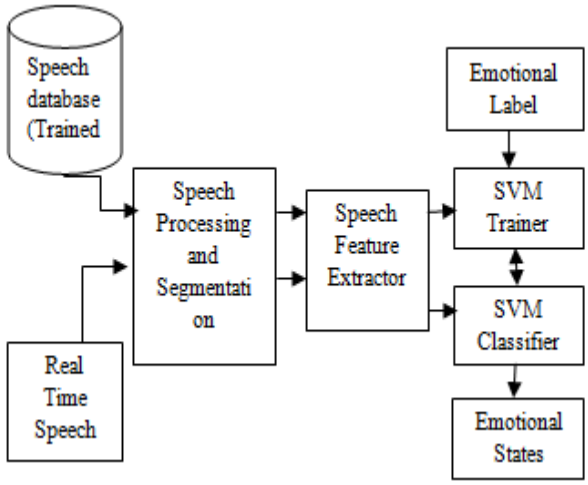

Fig 1: Emotion Recognition System.

3.1 Speech Acquisition

Emotional Speech Database in native language 'Marathi' is used as an input to the speech emotion recognition system. The performance of the speech emotion recognition system is high for the database having natural speech samples for training. The database which is an input to the real time speech emotion recognition system contains the real world emotions for testing while the acted emotions for training the classifier. It is more practical to use database that is collected from the real life situations [15] For real time speech emotion recognition system speech samples for training set formed by speakers other than testing phase speakers.

3.2 Feature Extraction

Feature extraction is the process by which the measurements of the given input can be taken to differentiate among emotional classes. In the field of Speech processing there are no established analytical methods that can reliably determine the emotion carried by the speech signal. A possible approach in this paper as seen in research is performing a trial to apply different and known signal processing methods, and to combine their results in such a way that there is a possibility for their pointing in the right direction - towards the emotion "hidden" in the signal. In this paper feature can be extracted which is combination of prosody features (i.e. pitch, energy, Zero crossing rate), quality features (i.e. Formant Frequencies, Spectral features etc.), derived features ((i.e.) Mel-Frequency Cepstral Coefficient (MFCC)), along with prosodic features like Pitch and Formants. Fig 2 shows the overall model for feature extraction that has been used for both training of classifier and testing the unknown speech samples.

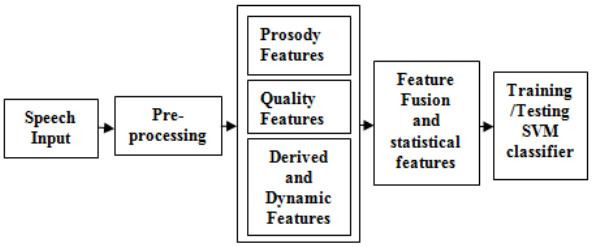

Fig 2: Steps for feature extraction

3.2.1 Pre-processing

The speech samples which are going to be processed for emotion recognition should go through a pre-processing step that removes the noise and other irrelevant components of speech corpus for better perception of speech data. The preprocessing step involves three major steps such as pre - emphasis, framing and windowing. In the pre-emphasis step FIR filter used. The filter impulse response is given by

$$H(z) = 1 + a\,z^{-1}$$, where a = -0.937

The filtered speech signal is then divided into frames of 25ms with an overlap of 10ms. A hamming window is applied to each signal frame to reduce signal discontinuity and thus avoid spectral leakage. Then speech emotion related features are extracted from the pre-processed speech data.

3.2.2 Prosody Features

Pitch, or fundamental frequency, is the acoustic correlative of the perceived tone height. The automatic estimation of pitch is

a non-trivial task. After prepressing of speech signal windowed signal is considered and FFT of it is obtained. For obtained signal the absolute values of the signal are calculated and then the logarithm of the signal is obtained. The signal is then transformed into Cepstral domain by taking its IFFT. The very first signal peak represents the pitch frequency.

Features Estimated from pitch contour: Standard Deviation Of Pitch, Variance Of Pitch, Average Peak, Total Peak Count, Peaks At 20 Hz Away (Twenty Count), Peaks At 50 Hz Away (Fifty Count), Peaks At 70 Hz Away Seventy Count, Maximum Difference, Difference Between Local Minima.

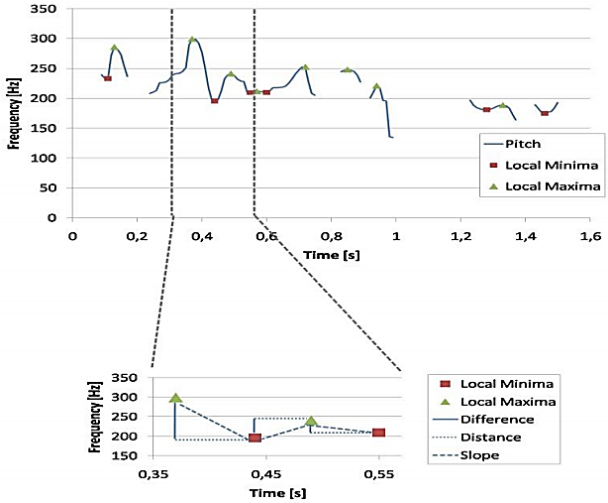

Fig 3: Feature Extraction from Speech

Zero-crossing rate is a key feature for identification of percussive sounds and information retrieval, and gives information related to change in frequency components of speech. The zero-crossing rate is the rate of sign-changes along a signal. Speech signal is of the varying nature and is stationary for short interval of time. Energy related features will show the variation of energy in the speech signal associated with a short-term region. To form feature vector, zero crossing rate and energy of feature can be extracted.

3.2.3 Quality Features

Formants play an important role as feature and can be termed as the spectral peaks of the sound spectrum of voice. They are measured as the amplitude peaks in the frequency spectrum of the sound. The first three formant frequencies can be taken as relevant features in the complete feature vector.

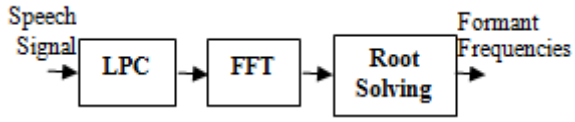

Fig 4: Formant Frequencies Extraction

Vocal tract resonances or the peaks in the sound envelope are nothing but formant frequencies. Linear Prediction Coefficients Based Formants Estimation Technique is proposed here for formant estimation. In this the vocal tract is considered as a linear filter with resonances and resonance frequencies of the vocal tract are called formant frequencies [2]. Graphically, when we plot vocal tract response the peaks of the vocal tract response of speech signal indicates to its formant frequencies. If we consider vocal tract as a time-invariant, all-pole linear system, then it results in conjugate pair of poles. Each of the conjugate pair of poles will correspond to a formant frequency or resonance frequency [2].

3.2.4 Dynamic Features

For various frequencies human ears perception is different. Mel-frequency cepstral coefficients (MFCCs) are the coefficients which identifies the variation of the human ear's critical bandwidth with frequency [8-10]. MFCC uses two types of filter. First filter is spaced linearly for the frequencies below 1000Hz whereas the second filter is logarithmic which is spaced above 1000Hz.The Mel-frequency cepstrum based on a non-linear Mel scale of frequency where short-term power spectrum of a sound is calculated. Linear cosine transform of a log power spectrum is derived which gives the coefficients known as MFCC. In this paper each speech frame of 30 ms, is considered for which a set of Mel-frequency cepstrum coefficients is computed. Fig.5 shows the MFCC feature extraction process.

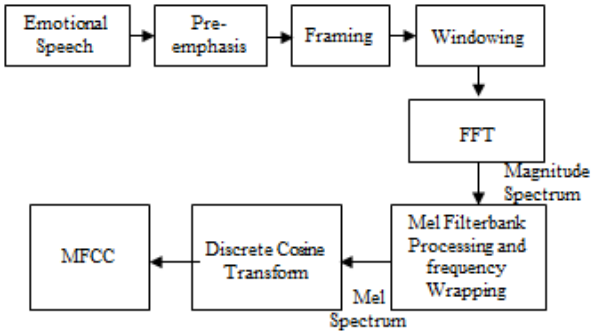

Fig 5: MFCC Feature Extraction from Speech

As described in previous feature extraction methods Preprocessing, Framing and windowing steps are performed. Next steps to be followed are as follows,

For evaluating frequency spectrum of windowed speech signal Fast Fourier Transform (FFT) algorithm is used. The Fourier Transform is to convert the convolution of the input pulse and the vocal tract impulse response in the time domain to frequency domain. This statement supports the equation below:

$$Y(w) = FFT[h(t) * X(t)] = H(w) * X(w) \quad \text{.....(11)}$$

The Mel filter bank [8] consists of 28 overlapping filters with triangular magnitude response. The cut off frequencies of these filters are determined by the center frequencies of the two adjacent filters. The filters have linearly spaced for the frequencies below 1000Hz and logarithmic spacing for frequencies above 1000 Hz. Filters have fixed bandwidth on the Mel scale. The frequencies range in FFT spectrum is very wide. Voice signal does not follow the linear scale as in FFT of signal. The bank of filters according to Mel scale is then performed. To approximate filter response on Mel scale triangular filters are used. Triangular filters computes the weighted sum of filter spectral components. [7, 8]. Then, each filter output is the sum of its filtered spectral components. After that the following equation is used to compute the Mel for given frequency f in HZ:

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

The logarithm of filtered components is calculated after frequency wrapping as the logarithm has the effect of changing multiplication into addition. Thus log step converts

the multiplication of the magnitude in the Fourier transform into addition.

Discrete Cosine Transform is used to orthogonalize the filter energy vectors after the log step. The orthogonalization step, compacts the information of the filter energy vector is into the first number of components and shortens the vector to number of components.

4. CLASSIFICATION USING SVM

The input speech signal was divided into frames and all the features were calculated for each frame. Now, In order to draw one conclusion from all the features of several frames of the input signal, we need to consider some kind of statistics. Statistical features [16] like Mean, Standard Deviation, Max and Range were considered for each feature over all the frames, and a single feature vector was formed including all the statistical parameters, representing the input signal. Then, the normalized statistical feature vector was provided to the Support Vector Machine (SVM) classifier for training or testing. A single SVM is a binary classifier which can classify 2- category data set. For this, first the classifier is manually trained with the pre-defined categories, and the equation for the hyper-plane is derived from the training data set. When the testing data comes to the classifier it uses the training module for the classification of the unknown data. But, automatic emotion recognition deals with multiple classes. Two common methods used to solve multiple classification problems like emotion recognition are (i)one-versus-all [17], and (ii)one-versus-one [18]. Fig.6 demonstrates these two methods of multilevel SVM [19], [20] classification for two different classes. In the former, one SVM is built for each category, which distinguishes this category from the rest. In the latter, one SVM is built to distinguish between every pair of categories. The final classification decision is made according to the results of all the SVMs with the majority rule. In the one-versus-all method, the category of the testing data is determined by the classifier based on the winner-takes-all strategy. In the one-versus-one body method, every classifier assigns the utterance to one of the two emotion categories, then the vote for the assigned category is increased by one vote, and the emotion class is the one with most votes based on a max-wins voting strategy. This paper uses one versus all SVM classification method to recognize the emotional states.

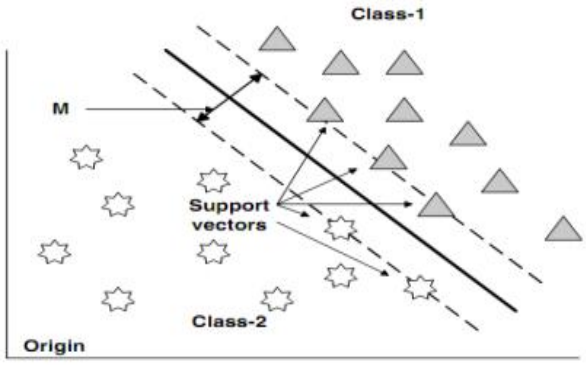

Fig 6: Hyperplane Decision Boundary.

5. CONCLUSION

The In this paper for recognition of emotion from real time speech the method is proposed which use combination of prosody, quality and dynamic features with the help of SVM classifier. The database to be used for this work is from one of the regional language 'Marathi' speech corpus. As the Real time Emotion Recognition System utilizes method of selecting optimized parameters it will reduces the time complexity compared with common method and also maintains the recognition accuracy rate at the same time.

In future for recognition of emotion, a real time model for this application can be developed.

6. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the paper.

7. REFERENCES

[1] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, "Emotion recognition from speech using global and local prosodic features," Int. J. Speech Technol., vol. 16, no. 2, pp. 143–160, Aug. 2012.

[2] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," Speech Commun., vol. 48, no. 9, pp. 1162–1181, Sep. 2006.

[3] I. Luengo and E. Navas, "Automatic Emotion Recognition using Prosodic Parameters" pp. 493–496, 2005.

[4] C. M. Lee, S. Member, S. S. Narayanan, and S. Member, "Toward Detecting Emotions in Spoken Dialogs," vol. 13, no. 2, pp. 293–303, 2005.

[5] M. Borchert and a. Dusterhoft, "Emotions in speech - experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments," 2005 Int. Conf. Nat. Lang. Process. Knowl. Eng., vol. 00, pp. 147–151, 2005.

[6] S. Wu, H. Tiago"Automatic Recognition Of Speech Emotion Using Long-Term Spectro-Temporal Features," 2009.

[7] A. B. Kandali, S. Member, A. Routray, and T. K. Basu, "Emotion recognition from Assamese speeches using MFCC features and GMM classifier."

[8] Y. Pan, P. Shen, and L. Shen, "Speech Emotion Recognition Using Support Vector Machine," vol. 6, no. 2, pp. 101–108, 2012.

[9] M. Dumas, "Emotional Expression Recognition using Support Vector Machines."

[10] P. Shen and X. Chen, "Automatic Speech Emotion Recognition Using Support Vector Machine," pp. 621–625, 2011.

[11] D. Fradkin and I. Muchnik, "Support Vector Machines for Classification," vol. 0000, pp. 1–9, 2006.

[12] A. Hassan and R. I. Damper, "Multi-class and hierarchical SVMs for emotion recognition."

[13] N. Yang, R. Muraleedharan, J. Kohl, I. Demirkol, and W. Heinzelman, "Speech-based Emotion Classification using Multiclass SVM with Hybrid Kernel and Thresholding Fusion" pp. 455–460, 2012