# Smart Phone Based Detection of Emotions Using Real Time Speech

Dipti Kale
Assistant Professor, Department of Electronics and Telecommunications Engineering
Shree L.R. Tiwari College of Engineering
Mumbai, India
*dipti.kale@slrtce.in*

Kritika Shukla
Lecturer, Department of Electronics and Telecommunications Engineering
Shree L.R. Tiwari College of Engineering
Mumbai, India
*kritika2308@gmail.com*

Menka Singh
Assistant Professor, Department of Electronics and Telecommunications Engineering
Shree L.R. Tiwari College of Engineering
Mumbai, India
*menka.singh@slrtce.in*

***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 smart phone based real time emotion recognition system is proposed, which quantifies the sparseness in speech recorded on a smartphone and use it to obtain a highly accurate and sparse approximation of a widely used feature of speech called the Mel-Frequency Cepstral Coefficients (MFCC) efficiently. The new feature extracted is the sparse MFCC or sMFCC. Classification of emotions performed using Multidimensional SVM and testing real time speech samples with training databases with emotional speech.

***Keywords*—*Smartphone, Emotion recognition, sMFCC, Support vector machine***

*****

## I. INTRODUCTION

Everyday activities of human being such as communicating, 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 behaviour 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 [2]. 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 behaviour. 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 [2]. 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 [5][7]. Therefore a system needed to develop which will extract powerful features and classification should also be accurate.

These days numerous Voice driven applications like voice commands (e.g. to launch an application or call some contact), voice enabled search (e.g. Google's voice search), voice recognizing personal assistant (e.g. iPhone's SiRi), and voice-based biometrics supported by All major Smartphone platforms [1]. In order to offer fast, real-time services for these applications, fast acoustic feature extraction is required both in time-domain and frequency-domain. Time domain acoustic features are sufficient in a few applications but the frequency-domain features are a must for a robust and accurate encoding of acoustic signals.

There are two ways in which smartphone applications and platforms that extract acoustic features. In the first method it records the audio and sends it to a remote server over the Internetfor further processing. This method has several limitations such as the requirement for an uninterrupted Internet connectivity and high bandwidth, and the associatedexpense of sending alarge chunk of audio data overthe cellular network [3]. In the second method itperforms the entire signal processing task inside the phone. But the limitation of this approach is that in order for fast

and real-time feature extractions, they must limit the sampling rate to the minimum [4]. As we know higher the sampling rate the better is the quality of samples. But the problem is – there is no efficient algorithm that extracts frequency domain acoustic features inside the phone in real-time at such high sampling rates.

In this paper, a method is used which enables the extraction of frequency domain acoustic features inside a Smartphone in real-time, without requiring any support from a remote server even when the sampling rate is as high. In this paper sFFT algorithm is used to extract a highly accurate and sparse approximation of a widely used feature for speech, called the Mel Frequency Cepstral Coefficients (MFCC) on the phone. A recent work [3] coined sparse Fast Fourier Transform (sFFT) – which is a probabilistic algorithm for obtaining the Fourier Transformation of time-domain signals that are sparse in the frequency domain. The algorithm is faster than the fastest Fourier Transformation algorithm under certain conditions.

Here, smart phone based real time speechemotion recognition system is proposed which extracts spectral feature like MEL-FREQUENCY CepstralCoefficient MFCC for robust automatic recognitionof speaker's emotional states. Multilevel SVM classifier is used for identification of six discrete emotional states namely angry, fear, happy, neutral,and sad and surprise.

## II. SPEECH EMOTION RECOGNITION SYSTEM

Speech emotion recognition system is typical pattern 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 [6].

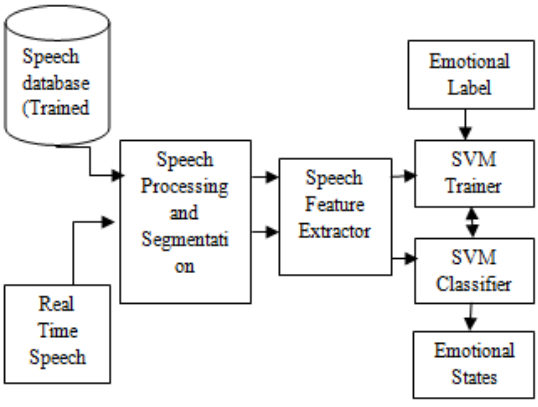


*Fig 1: Emotion Recognition System*.

### A. Speech Acquisition:

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 [2] For real time speech emotion recognition system speech samples for training set formed by speakers other than testing phase speakers. For the system proposed database collected from 10 Smartphone users whose emotional speech data recorded in emotions like Happy, Anger, Neutral, Sad, Fear and Surprise.Number of samples of each emotions are needed to record in Smartphone in normal room environment for this system.

### B. 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 most popularly used short-term, frequency domain acoustic feature extraction method MFCC to detect the emotion "hidden" in the signal.

The Mel-Frequency Cestrum Coefficients (MFCC) is one of the most popular short-term, frequency domain acoustic features of speech signals [6]. The MFCC have been widely used in speech analysis because of their compact representation (typically,each speech frame is represented by a 39-element vector), close resemblance to how human ear responds to different sound frequencies, and their less susceptibility to environmental noise. But system proposed here is to be meant for Smartphonetherefore the time to compute MFCC features is always longer than the duration of the recorded audio when the sampling rate is higher than 8 KHz. Therefore, at these higher rates, the application is not capable of real-time performance. For solution sparseness in speech can be exploited to compute a close approximation of MFCC feature vectors on a Smartphone in real-time.

### THE SPARSE MFCC ALGORITHM:

The idea of sparse MFCC algorithm is to compute a sparse approximation of MFCC features from a given frame of discrete time-domain signals xn of length n. The algorithm uses a modified version of sFFT as a subroutine. [2]

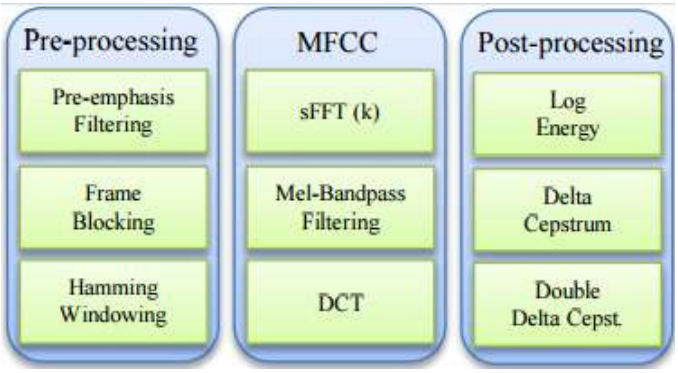


*Fig 2: sMFCC feature extraction process.*

a. Pre-processing:

The time-domain signals are first passed through a high-pass pre-emphasis filter to amplify the high-frequency formants that are suppressed in speech. We then segment the signals into frames of 64 ms with an overlap of 1/3 of the frame size. A hamming window is applied to each frame to ensure the continuity between the first and last points which is required for FFT.

b. MFCC:

The MFCC feature extraction starts with the estimation of power spectrum which is obtained by taking the square of the absolute values of the FFT coefficients. In sMFCC algorithm instead of FFT Block the process of sFFT is used in which we precompute the FFT, keep the FFT coefficients sorted in non-increasing order, and take only the largest k coefficients while making other coefficients zero – which computes sMFCC(k). Once the Fourier coefficients are obtained, we follow the standard procedure of MFCC [2]. We apply 20 triangular band-pass filters (called Mel-banking) to obtain 20 log energy terms, perform a DCT to compress them, and take the first 13 coefficients to constitute a 13-element sMFCC vector M.

c. Post-processing:

The 13-element sMFCC vector is augmented to include the delta and double delta cepstrums to add dynamic information into the feature vector, and thus we obtain a 39-element feature vector.

In sMFCC computing process first recorded signal divided into speech frames. Each frame goes through the MFCC feature extraction process which happens in real-time. Each spoken word produces a number of frames, and a 39-element MFCC feature vector is obtained for each frame. We take the mean and the standard deviation of each of the 39 MFCC coefficients over all frames to obtain a single 78-element feature vector which is used in the classification step.

C. Classification using SVM:

In the last decades, there were several machine learning methods that can use for classifier and recognition of human physical activities including Naïve Bayes, Support Vector Machines (SVMs), Threshold based and Markov chain [7]. Although there is not any study that can find out the best method for human physical activities classification, but SVMs have been successfully widely used in many research related to handwriting recognition and speech recognition. Therefore, in this study, SVMs method will be used to classify and recognize human activities. In order to find out the best hyperplane for data classification, SVMs search the hyperplane which has the largest margin. Figure 3 shows both two hyperplanes can be divided in two class. However, figure 1.b shows the larger margin between two classes than figure 1.a. The larger margin will help the classification in next modules easier and avoid mistakes as much as possible. Thus, in SVMs In particular, Anguita introduced the concept of Hardware-Friendly SVM [8]. The fixed point arithmetic is exploited in the feed-forward phase of SVM classifier. This model is extended for multiclass classification. Because this research achieved an average accuracy of 89% with small among of memory, it has good advantage when used in limited resources hardware devices like smartphone.

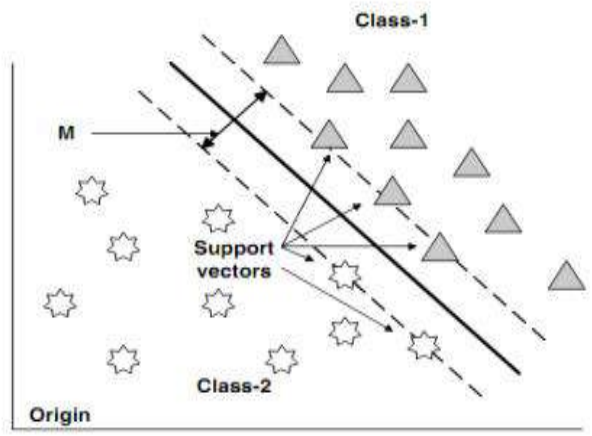


*Fig.3: Support Vector Machine*

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 [8] 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 [7], and (ii)one-versus-one [7]. Fig.3 demonstrates these two methods of multilevel SVM [7] 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.

## III. CONCLUTION AND FUTURE WORK

In this paper, a smartphone based system is proposed to recognize the emotion from speech. The proposed system includes data acquisition systems, features extraction, data processing, classification and human emotion recognition. The In this paper for smartphone based recognition of emotion from real time speech the method is proposed which use sparse MFCC feature extraction method with the help of SVM classifier. The database to be used for this work is recorded from smartphone in normal home environments. As the smartphone based Real time Emotion Recognition System utilizes method sMFCC and multiclass SVM 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 smartphone based real time model for this application can be developed.

## REFRENCES

[1] Z. Fang, Z. Guoliang, and S. Zhanjiang. “Comparison of different implementations of mfcc.” Journal of Computer Science and Technology, 16(6):582–589, nov 2001.

[2] H. Hassanieh, P. Indyk, D. Katabi, and E. Price. “Nearly optimal sparse fourier transform.” In 44th ACM Symposium on Theory of Computing 2012 (STOC ’12), NewYork, NY.

[3] H. Hassanieh, P. Indyk, D. Katabi, and E. Price. “Simple and practical algorithm for sparse fourier transform.” In ACM-SIAM Symposium on Discrete Algorithms 2012 (SODA ’12), Kyoto, Japan

[4] E. Miluzzo, N. D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. B. Eisenman, X. Zheng, and A. T. Campbell. “Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application.” In 6th ACM conference on Embedded network sensor systems 2008 (SenSys ’08), pages 337–350, Raleigh, NC, USA.

[5] L. Muda, M. Begam, and I. Elamvazuthi. “Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques.” Journal of Computing, 2(3):138–143, 2010

[6] B. Logan. “Mel frequency cepstral coefficients for music modeling.” In International Symposium on Music Information Retrieval 2000 (ISMIR ’10).

[7] A. Hassan and R. I. Damper, “Multi-class and hierarchical SVMs for emotion recognition.”

[8] 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

[9] J. Jensen, M. Christensen, M. Murthi, and S. Jensen. “Evaluation of mfcc estimation techniques for music similarity.” In European Signal Processing Conference 2006 (EUSIPCO ’06).