Speech Emotion Verification Using Gabor Dictionary & SVM

Minu Mariam Mathew[1], Mahesh B.S.[2], Shanavz K.T.[3]

[1]*College of Engineering, Kallooppara*

[2]*Assistant Professor,* [3]*Associate Professor*

Abstract - Speech verification uses speech recognition to verify the correctness of the pronounced speech. The proposed system consists of two parts – feature extraction & emotion verification. In the first part, for each frame, important atoms are selected from the Gabor dictionary by using matching pursuit algorithm. The scale, frequency and magnitude of the atoms are extracted to construct a non uniform scale – frequency map, which supports auditory discriminability by the analysis of critical bands. In addition to SFM values, Pitch, Mel Frequency Cepstral Coefficients(MFCC),LPC coefficients and spectrum is also extracted from the speech signal. In the second part, the features extracted are used to train the SVM classifier and thus, classifies the emotion and emotion is verified.

Keywords—Gabor dictionary, Pitch, MFCC, LPCC, Spectrum, Scale Frequency Map(SFM), Matching Pursuit Algorithm, SVM.

I. INTRODUCTION (10pt, Caps, normal)

Emotion is a person's state of feeling in the sense of an affect. Emotions are complex. According to some theories, they are a state of feeling that results in physical and psychological changes that influence our behavior. The physiology of emotion is closely linked to arousal of the nervous system with various states and strengths of arousal relating, apparently, to particular emotions. Emotion is also linked to behavioral tendency. According to other theories, emotions are not causal forces but simply syndromes of components, which might include motivation, feeling, behavior, and physiological changes, but no one of these components is the emotion.

Speech is the fundamental and intuitive communication channel of human interaction. Emotional changes are reflected in the speaking rate, intonation, stress, and many other linguistic features. Speech verification uses speech recognition to verify the correctness of the pronounced speech. Speech verification does not try to decode unknown speech from a huge search space, but instead, knowing the expected speech to be pronounced, it attempts to verify the correctness of the utterance's pronunciation, cadence, pitch, and stress.

One of the primary steps for building a good system for emotion recognition from speech is to extract the discriminative features for determining a set of the important emotions to be classified by an automatic emotion recognizer. Identification of emotions can be done using three factors - the content of the speech, facial expressions of the speaker or by the features extracted from the emotional speech. This paper is confined to the verification of emotion by making use of the features extracted from the speech. To this end researchers have used the statistics of the different attributes of speech for a good representation of the signal. Basically, these attributes have been broadly categorized as contextual and non-context based attributes. Speech emotion verification is useful for applications requiring natural man-machine interaction, such as web movies and computer tutorial applications, where the response to the user depends on the detected emotions.

II. PREVIOUS WORK

1. A framework for automatic human emotion classification using emotion profiles

This paper describes an emotion classification paradigm, based on emotion profiles (EPs). This paradigm is an approach to interpret the emotional content of naturalistic human expression by providing multiple probabilistic class labels, rather than a single hard label. EPs provide an assessment of the emotion content of an utterance in terms of a set of simple categorical emotions: anger; happiness; neutrality; and sadness [3]. The first step is the feature extaction were pitch features are extracted and features were normalized over each speaker using z-normalization The next step is feature selection and creation of emotional profile. In this method Support Vector Machine(SVM)is used for classification. An emotional utterance is assigned to an emotion class based on the representation of the emotions within the EP[5].

2. Improved emotion recognition with a novel speaker-independent feature

In this paper, a novel speaker-independent feature, the ratio of a spectral flatness measure to a spectral center

(RSS), with a small variation in speakers when constructing a speaker-independent system is proposed. Gender and emotion are hierarchically classified by using the proposed feature (RSS), pitch, energy, and the mel frequency cepstral coefficients. First, gender is detected by pitch, energy, and the MFCCs; then, the pitch, energy, and the MFCCs are used to separate anger and joy (group 1) from neutrality and sadness (group 2). Finally, group 1 is classified into anger and joy, and group 2 into neutrality and sadness using the proposed feature, the RSS. After gender detection, all classifications are performed by a classifier learned using the appropriate gender database[8].

3. Speech emotion recognition in 3d space

This paper presents three approaches (robust regression, support vector regression, and locally linear reconstruction) for emotion primitive estimation in 3D space (valence/activation/dominance), and two approaches (average fusion and locally weighted fusion) to fuse the three elementary estimators for better overall recognition accuracy. The first step is the data acquisition and emotion primitive evaluation. Next step is the feature extraction and feature ranking and selection. Finally the emotion is verified by taking the average of output of all the elementary estimators[7].

4. Selection of classifier is a major issue

K-Nearest Neighbors (KNNs), Artificial Neural Networks (ANNs), Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), and SVMs. Although these classifiers have been broadly used in pattern recognition, they still have some drawbacks. For instance, KNNs do not have a discriminant learning process and are susceptible to outliers. Unlike KNNs, ANNs typically require a large number of training data. When addressing the aforementioned problem of emotion variance, collecting sufficient balanced training samples is difficult.

Regarding the HMM, despite the advantage of time-series alignment, the Markovian assumption does not map well to many real-world domains. Moreover, the HMM is trained by using only the data of a certain category, potentially causing biased indiscriminate learning. Research has revealed that GMMs suffered from the dimensionality problem, and the number of mixture models has to be recursively set many times to achieve a satisfactory result. SVMs do not have the aforementioned shortcomings and are effective with limited training data but, when compared with sparse representation it is challenging [1].

5.Previous work

The system can be divided into two stages—feature extraction and verification. In the first stage, two types of acoustic features are analyzed. One is the scale-frequency map, and the other is the prosodic feature set. First, for each sound frame, important atoms from the Gabor dictionary are selected by using the Matching Pursuit algorithm. Each atom in the dictionary takes the form of a Gabor function, which includes frequency, scale, phase, and position information. Following atom selection, the magnitude of the atoms with the same frequency and scale is accumulated and averaged, yielding a scale-frequency map (SFM). SFMs provide discriminability and have a fine resolution at critical bands that are more appropriate than uniform-distributed frequencies for modeling human auditory perception and less susceptible to noise [1].

To increase further robustness against emotion variance, sparse representation is used to transform scale-frequency maps into sparse coefficients. After feature extraction,the first is the proposed sparse representation verification, based on Gaussian-modeled residual errors, which can accommodate emotion variance in the voices of speakers. The second approach involves the use of an indicator, the EAI, for measuring the degree of blended emotions and emotion variances. A higher index indicates that the proposed sparse representation verification approach generates a score with higher confidence. The indicator uses the same sparse coefficients as the sparse representation verification [1].

III. PROPOSED METHODOLOGY

This proposed model aims to verify emotion by using SVM as classifier.

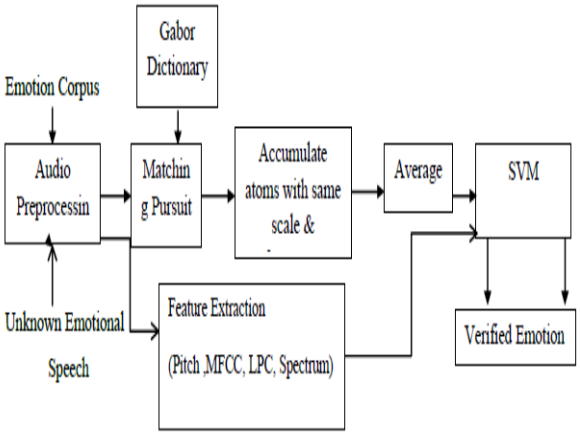

Fig3.1 Block Diagram of the proposed System

Fig3.1 shows the overall block diagram of the proposed method. First audio is preprocessed. Next step is to create gabor dictionary from which important atom is selected by using matching pursuit algorithm. Accumulate atoms which is having same scale and frequency and then taking average to give the scale frequency map(SFM). In addition

to SFM values, pitch, MFCC, LPCC, spectrum features are also extracted from the input speech. These features are used to train the SVM classifier and which classifies the emotion and hence verify the emotion.

1.Framing

In speech processing it is often advantageous to divide the signal into frames to achieve stationarity. Normally a speech signal is not stationary, but seen from a short-time point of view it is. When the signal is framed it is necessary to consider how to treat the edges of the frame. This result from the harmonics the edges add. Therefore it is expedient to use a window to tone down the edges. As a consequence the samples will not be assigned the same weight in the computations and for this reason it is prudent to use an overlap.

An audio signal is constantly changing, so we assume that on short time scales the audio signal doesn't change much. This is why we frame the signal into 20-40ms frames. If the frame is much shorter we don't have enough samples to get a reliable spectral estimate, if it is longer the signal changes too much throughout the frame, and the FFT will end up smearing the contents of the frame.

2. Feature Extraction

This section introduces the extraction of the scale- frequency maps, which consists of two parts, i.e., matching pursuit and map generation.

2.1 Scale Frequency Map

The Gabor representation is formed by a bandlimited signal of finite duration, thus making it more suitable for time–frequency localized signals. The Gabor representation was to be optimal in the sense of minimizing the joint two-dimensional uncertainty in the combined spatial-frequency space. Gabor atoms result in the lowest reconstruction error, as compared with the Haar or the Fourier transforms using the same number of coefficients. Due to the nonhomogeneous nature of environmental sounds, using features with these Gabor properties would benefit a classification system [14].Gabor functions are sine-modulated Gaussian functions that are scaled and translated, providing joint time–frequency localization Mathematically, the discrete Gabor time–frequency atom [2] is written as

$$G_{\rho,\mathrm{u},\mathrm{f},\theta}(t)=\frac{K_{\rho,\mathrm{u},\mathrm{f},\theta}}{\sqrt{\rho}}e^{-\Pi(t-u)^2}\rho^{-2}\cos\left[2\Pi f(t-u)+\theta\right] \quad (1)$$

where 'u' represents the central temporal position of the Gabor function, 'ρ' denotes the scale, which controls the width of the Gabor function, f refers to the frequency, 'θ' denotes the phase, 't' represents the indices of the sampling points of an input signal, and 'K' denotes the normalization factor such that $\left\|G_{p,u,f,\theta}\right\|$ =1.The scale ρ=$\{2^j \mid j=1,2,..8\}$,the central temporal position and u= 0,64,128,192 and the frequency f = { 50, 150, 250, 350, 450, 570, 700, 840, 1000, 1170, 1370, 1600, 1850, 2150, 2500, 2900, 3400, 4000, 4800, 5800, 7000},the phase θ =0, and the indices of the sampling points t=0- 255 [1].

The properties of the signal components are explicitly given by the scale, frequency, time and phase indices of the selected atoms. It computes the best nonlinear approximation to a signal in a complete, redundant dictionary. A typical MP procedure consists of two phases: Selection and decomposition phases. The MP procedure iteratively performs these phases until it reaches a stopping criterion. In the selection phase, the MP process chooses an atom from a dictionary, examining the close similarity between the atom and the input signal [14]. Assume D that denotes the dictionary containing a collection of atoms. Then, D can be represented as

$$\mathrm{D}=\sigma_\gamma \mid \gamma \in \Gamma \quad (2)$$

where γ is the parameter vector of atom σ, and Γ is the set of the parameter vectors. Next, the MP process computes the similarity between each atom and the input signal based on the inner product operation. The atom is determined based on the following equation:

$$\sigma^*(n)=\arg\max\left\{\left|\left\langle\sigma,R_s(n-1)\right\rangle\right|\right\} \quad (3)$$

where $\sigma^*(n)$ is the atom with the largest absolute value of inner products after the nth iteration, $R_s(n)$ is the residual signal, $\langle . \rangle$ means the inner product operation, and $|.|$ denotes the absolute operation. After selecting the best atom, the process then subtracts from the previous residual signal and gets a new residue.

The decomposition phase is

$$R_s(n)=R_s(n-1)-\left\langle\sigma^*(n),R_s(n-1)\right\rangle\sigma^*(n) \quad (4)$$

$R_s(n)$ is equivalent to the original signal s when n equals zero. Since the selection phase favors the atom with the most similarity in each iteration, the energy of the residual signal reaches the minimum in the decomposition phase. In other words, the reconstruction error is minimal when the system rebuilds the signal using the selected atoms [1].

After the Gabor dictionary is generated, an input signal can be decomposed into small units by using the MP algorithm. To enhance decomposition, the input signal is first divided into frames, each of which is further represented by the top best N atoms based on MP selection. Each atom can be re-written in the form of Gabor function, which contains scale and frequency information. The characteristic of an atom can be converted into a scale-frequency map by extracting its scale and frequency information. After each frame is represented by atoms, the system extracts the information (i.e., frequency, scale, phase, and position information.) from each atom [1]. To obtain the values in the scale-frequency map, the system accumulates the magnitude of each of the N atoms with the same frequency and scale.

2.2 Pitch

The sound that comes through vocal tract starts from the larynx where vocal cords are situated and ends at mouth. The vibration of the vocal cords and the shape of the vocal tract are controlled by nerves from brain. The sound, which we produce, could be categorized into voiced and unvoiced sounds. During the production of unvoiced sounds the vocal cords do not vibrate and stay open whereas during voiced sounds they vibrate and produce what is known as glottal pulse. A pulse is a summation of a sinusoidal wave of fundamental frequency and its harmonics (Amplitude decreases as frequency increases). The fundamental frequency of glottal pulse is known as the pitch.

2.3 Spectrum

The power spectrum of a speech signal describes the frequency content of the signal over time. The first step towards computing the power spectrum of the speech signal is to perform a Discrete Fourier Transform (DFT). A DFT computes the frequency information of the equivalent time domain signal. Since a speech signal contains only real point values, we can use a real-point Fast Fourier Transform (FFT) for increased efficiency. The resulting output contains both the magnitude and phase information of the original time domain signal.

2.4 MFCC

Framing is done to extract the stationary features. 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. Fast Fourier Transform (FFT) algorithm is ideally used for evaluating the frequency spectrum of speech. FFT converts each frame of N samples from the time domain into the frequency domain. The mel filter bank consists of overlapping triangular filters with the cut-off frequencies determined by the center frequencies of the two adjacent filters [4].

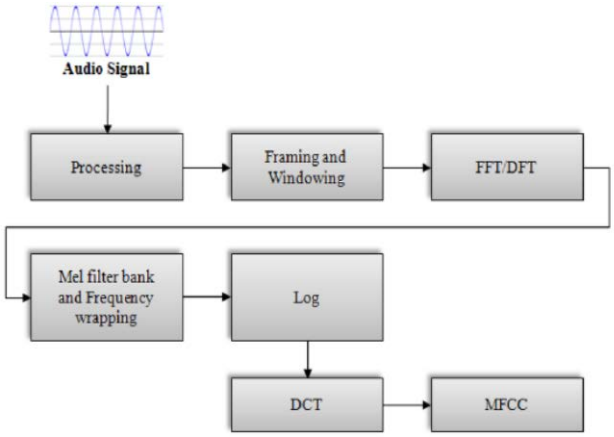

Fig3.2 Block diagram of MFCC

The filters have linearly spaced centre frequencies and fixed bandwidth on the mel scale. Human ear don't hear loudness on a linear scale. Hence log is taken. Log is used as it allows us to use cepstral mean subtraction which is a channel normalization technique. The logarithmic Mel-Scaled filter bank is applied to the Fourier transformed frame. This scale is approximately linear up to 1 kHz, and logarithmic at greater frequencies [12]. The relation between frequency of speech and Mel scale can be established as:

Frequency (Mel Scaled) = [2595log (1+f (Hz)/700] (5)

Since the filter banks are overlapping, energies are correlated. DCT decorrelates the energies. Only higher coefficients of about 12 are kept as higher DCT coefficient represent faster changes in filter band energies which may degrade the performance of the system [4].

2.5 LPCC

LPC (Linear Predictive Coding) analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz [3]. The process of removing the formants is called inverse filtering and the remaining signal is called the residue.

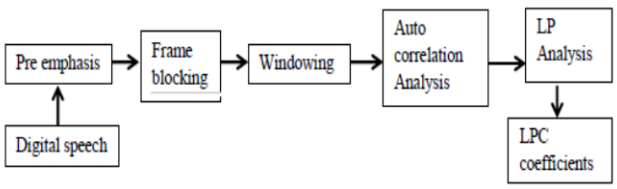

Fig3.3 Block diagram of LPCC

The basic idea behind LPC coding is that each sample can be approximated as a linear combination of a few past samples. The linear prediction method provides a robust, reliable, and accurate method for estimating the parameters. This could be used to give a unique set of predictor coefficients. These predictor coefficients are estimated every frame, which is normally 20 ms long. The predictor coefficients are represented by ak. Another important parameter is the gain (G). The transfer function of the time varying digital filter is given by H(z) = G/(1-Σakz-k). In reality the actual predictor coefficients are never used in recognition, since they typical show high variance. The predictor coefficient is transformed to a more robust set of parameters known as cepstral coefficients [13].

3. Verification

In the verification part we use Support Vector Machine (SVM) as classifier. SVM is trained by the features extracted from the speech and which classifies the emotion and hence verify the emotion.

3.1 Support Vector Machine

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. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non- probabilistic binary linear classifier.

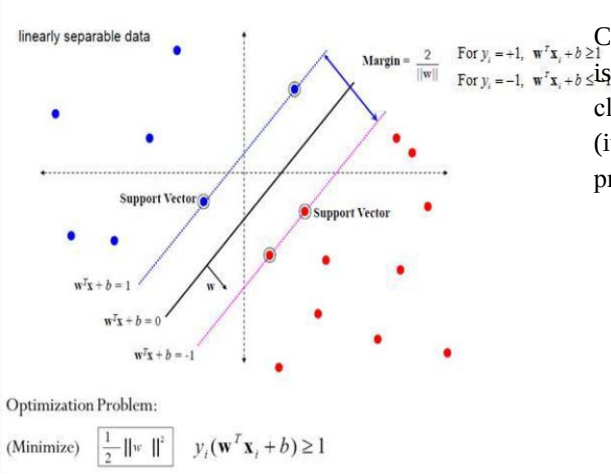

Fig3.4 SVM Optimization

An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces[3].

Geometrically, the distance between these two hyperplanes is 2/‖w‖ , so to maximize the distance between the planes we want to minimize ‖w‖. As we also have to prevent data points from falling into the margin, we add the following constraint: for each i either

$w.x - b \geq -1$ for xi of the first class

or

$w.x - b \leq -1$ for xi of the second.

This can be rewritten as:

$yi(w.x - b) \geq -1$, for all $1 \leq i \leq n$

3.2 Multiclass SVM

Multiclass SVM aims to assign labels to instances by using support vector machines, where the labels are drawn from a finite set of several elements. The dominant approach for doing so is to reduce the single multiclass problem into multiple binary classification problems. Common methods for such reduction include building binary classifiers which distinguish between

(i) one of the labels and the rest (one-versus-all)

(ii) between every pair of classes (one-versus-one).

Classification of new instances for the one-versus-all case is done by a winner-takes-all strategy, in which the classifier with the highest output function assigns the class (it is important that the output functions be calibrated to produce comparable scores).

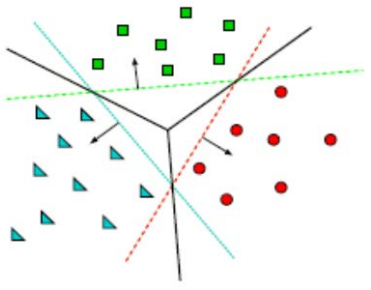

Fig3.5 One-versus-all

For the one-versus-one approach, classification is done by a max-wins voting strategy, in which every classifier assigns the instance to one of the two classes, then the vote for the assigned class is increased by one vote, and finally the class with the most votes determines the instance classification [6].

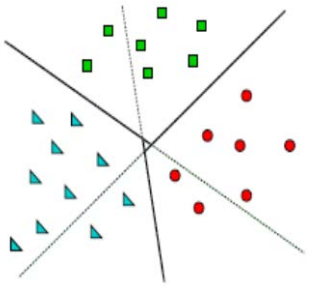

Fig3.6 One-versus-one

IV. SIMULATION/EXPERIMENTAL RESULTS

In this work, two databases are used Berlin Emotional database with seven emotions happy, sad, anger, disgust, neutral, boredom and fear(The database was created by the Technical University of Berlin) and Toronto Emotional Speech set with five emotions happy, sad, anger, neutral and fear (The database was created by University of Toronto, Psychology Department, 2010).

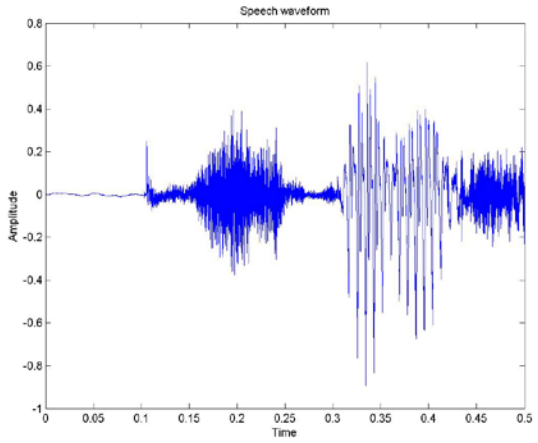

Fig4.1 Test Signal

The overall accuracy is calculated as the ratio of the total number of emotion correctly verified to the total number of files tested.

Table 1: Accuracy of each emotion for Berlin Emotional Speech

EMOTION	TOTAL NO.OF FILES TESTED	CORRECTLY VERIFIED
HAPPY	10	10
SAD	10	10
ANGER	10	9
NEUTRAL	10	10
DISGUST	10	9
BOREDOM	10	8
FEAR	10	10

Table 2: Accuracy of each emotion for Toronto Emotional Speech set

EMOTION	TOTAL NO.OF FILES TESTED	CORRECTLY VERIFIED
HAPPY	10	9
SAD	10	10
ANGER	10	10
NEUTRAL	10	9
FEAR	10	9

Overall Accuracy = (Total no:of emotions correctly verified)/(Total no:of files)

(i) Proposed Method

Overall Accuracy for Berlin Emotional Speech

= 66/70 = 94.28%

Overall Accuracy for Toronto Emotional Speech Set

= 47/50 = 94%

(ii) Previous Work

Overall Accuracy for Berlin Emotional Speech

= 26/35 = 74.28%

V. CONCLUSION & FUTURE SCOPE

Caregivers and doctors can analyze the status logs that are generated by the method and evaluate the mental health of subjects or patients. Speech emotion verification is useful for applications requiring natural man-machine interaction, such as web movies and computer tutorial applications, where the response to the user depends on the detected emotions. To create a distinguishable feature when the system verifies emotions, the scale-frequency map derived from critical bands is calculated by using the Matching Pursuit algorithm. The system also extracts features like pitch, MFCC, LPC, spectrum . Multi class

SVM is used to verify the emotion. The overall accuracy is 94.28% for Berlin Emotional database and 94% for Toronto Emotional Speech set and when compared with sparse representation, where there is a need of manual thresholding function, classifying the emotions using SVM is more accurate.

REFERENCES

[1] Jia-Ching Wang, Yu-Hao Chin, Bo-Wei Chen,Chang-Hong Linand Chung-Hsien Wu, "speech emotion verification using emotion variance Modeling and discriminant scale-frequency maps," IEEE/ACM Trans.Audio, Speech, and Language processing, vol.23, no. 10, october 2015 .

[2] J. C. Wang, C. H. Lin, B. W. Chen, and M. K. Tsai, "Gabor-based nonuniform scale-frequency map for environmental sound classification in home automation," IEEE Trans. Autom. Sci. Eng., vol. 11, no. 2, pp. 607–613, Apr. 2014.

[3] Yixiong Pan, PeipeiShen and LipingShen "Speech Emotion Recognition Using Support Vector Machine" Department of Computer Technology International Journal of Smart Home Vol. 6, No. 2, April, 2012.

[4] Bhoomika Panda , DebanandaPadhi, Kshamamayee Dash, Prof. SanghamitraMohanty, "Use of SVM Classifier & MFCC in Speech Emotion Recognition System". Volume 2, Issue 3, March 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.

[5] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 5, pp. 1057–1070, Jul. 2011.

[6] Aamir Khan,Muhammad Farhan, Asar Ali "Speech Recognition: Increasing Efficiency Support Vector Machine". International Journal of Computer Applications (0975 – 8887)Volume 35– No.7, December 2011.

[7] D. Wu, T. D. Parsons, E. Mower, and S. Narayanan, "Speech emotion estimation in 3D space," in Proc. IEEE Int. Conf. Multimedia Expo, Singapore, Jul. 19–23, 2010, pp. 737–742.

[8] E. H. Kim, K. H. Hyun, S. H. Kim, and Y. K. Kwak, "Improved emotion recognition with a novel speaker-independent feature," IEEE/ASME Trans. Mechatronics, vol. 14, no. 3, pp. 317–325, Jun. 2009.

[9] Xia Mao, Lijiang Chen, Liqin Fu, "Multi-Level Speech Emotion Recognition based on HMM and ANN" 2009 World Congress on Computer Science and Information Engineering.

[10] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time-frequency audio features," IEEE Trans. Audio, Speech, Lang. Process., vol. 17, no. 6, pp. 1142–1158, Aug. 2009.

[11] P. Dunker, S. Nowak, A. Begau, and C. Lanz, "Content-based mood classification for photos and music: A generic multi-modal classification framework and evaluation approach," in Proc. 1st ACM Int. Conf. Multimedia Inf. Retrieval, Vancouver, BC, Canada, Oct. 30–31, 2008, pp. 97–104.

[12] H. Ting, Y. Yingchun, and W. Zhaohui, ``Combining MFCC and pitch toenhance the performance of the gender recognition," in Proc. 8th Int. Conf.Signal Process., vol. 1. 2006.

[13] O. Pierre-Yves, ``The production and recognition of emotions in speech: Features andalgorithms," Int. J. Human-Comput. Stud., vol. 59, no. 1, pp. 157_183, 2003.

[14] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," IEEE Trans. Signal Process., vol. 41, no. 12, pp. 3397–3415, Dec. 1993.