Emotion Detection From Speech Using Mfcc & Gmm

K. J. Patil
student
SSBT'S College of Engineering
Jalgaon, M.S., India

P. H. Zope
Asst. Professor
SSBT'S College of Engineering
Jalgaon, M.S., India

S. R. Suralkar
H.O.D
SSBT'S College of Engineering
Jalgaon, M.S., India

Abstract

In these years, literature about automatic emotion recognition is growing dramatically due to the development of techniques in computer vision, speech analysis and machine learning. However, automatic recognition on emotions occurring on natural communication setting is a largely unexplored and challenging problem.

Speech processing is emerged as one of the important application area of digital signal processing. Various fields for research in speech processing are emotion detection from speech, speech recognition, speaker recognition, speech synthesis, speech coding etc. The objective of automatic emotion detection is to extract, characterize and recognize the information of speaker's emotions. Feature extraction is the first step for speaker recognition. Many algorithms are suggested/developed by the researchers for feature extraction. In this report, the Mel Frequency Cepstrum Coefficient (MFCC) feature has been used for designing an automatic emotion detection system. Some modifications to the existing technique of MFCC for feature extraction are also suggested to improve the emotion detection efficiency.

This report presents an approach to emotion recognition from speech signals. In this report, the framework to extract features from the speech signal that can be used for the detection of emotional state of the speaker. An essential step in the generation of expressive speech synthesis is the automatic detection and classification of emotions most likely to be present in speech input.

Keywords— speech recognition , MFCC ,GMM.

1.INTRODUCTION

. EMOTION plays a crucial role in day-to-day interpersonal human interactions. Recent findings have suggested that emotion is integral to our rational and intelligent decisions. It helps us to relate with each other by expressing our feelings and providing feedback. This important aspect of human interaction needs to be considered in the design of human–machine interfaces (HMIs). To build interfaces that are more in tune with the users' needs and preferences, it is essential to study how emotion modulates and enhances the verbal and nonverbal channels in human communication.

The tight coupling between emotional expression and human behavior is well documented, even if logical reasoning skills remain intact, the brain becomes incapable of making appropriate decisions when its emotion-controlling centers are damaged. Hence, the vital importance of cogent emotion analysis in most affective computing applications, ranging from natural language interfaces to e-learning environments,educational or entertainment games, opinion mining and sentiment analysis, humor recognition, and security informatics. For example, emotion detection is an essential tool for monitoring the presence of hateful or violent rhetoric. When it comes to man–machine communication, it is thus highly desirable to take account of emotional states as an integral part of human–computer interaction, at both input and output levels. If a spoken dialog system could reliably determine that a user is upset or annoyed, for instance, it could switch to a potentially more adequate mode of interaction. Likewise, expressive speech synthesis is expected to play a pivotal role in the widespread deployment and acceptance of future natural language interfaces.

Detecting emotion from speech can be viewed as a classification task. It consists of assigning, out of a fixed set, an emotion category e.g., joy, anger, boredom, sadness, fear, frustration, annoyance, satisfaction & neutral, to a speech utterance. This report presents an approach to emotion recognition from speech signals. In this report, the framework to extract features from the speech signal that can be used for the detection of emotional state of the speaker is discussed.

2.System design

The system consists of four major parts :-

I . Speech Acquisition
II. Feature Extraction
III. Machine Learning
IV. Information Fusion

For the purpose of feature extraction, spectral analysis algorithm such as Mel-frequency Cepstral Coefficients, MFCCs will be used. For prosody analysis, the statistics of pitch and energy will be used to determine prosodic features. To determine the emotion, Information fusion algorithm will be designed. For the fusion algorithm, Spectral Analysis GMM model will be applied to determine the probability density function. A k-NN will be used for the prosody feature analysis in the fusion algorithm [1].

But it is found that, emotion recognition algorithm that use prosodic features are not sufficiently accurate. However, phonetic feature have less information for discriminating emotions. Actually there is more independent component in the phonetic features of speech than in prosodic features of speech, then the accuracy of emotion recognition can be improved by increasing the number of independent phonetic features. Therefore, we propose an emotion recognition algorithm that focuses more on phonetic features of speech.

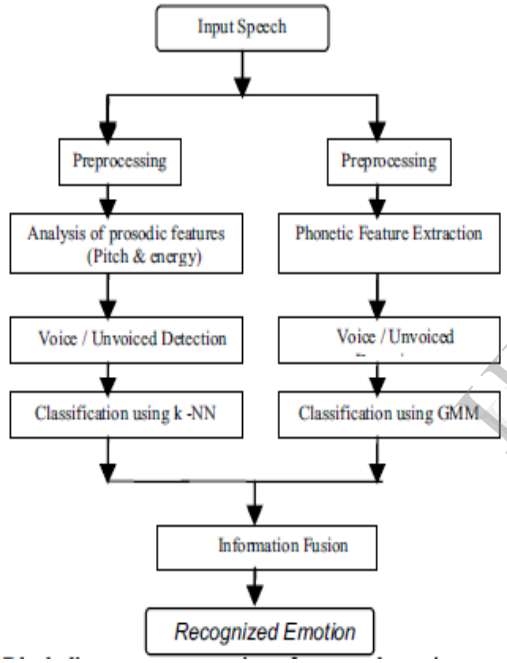

Fig 1. Typical emotion recognition system[1]

.

The extraction and selection of the best parametric representation of acoustic signals is an important task in the design of any speech recognition system; it significantly affects the recognition performance. A compact representation would be provided by a set of mel-frequency cepstrum coefficients (MFCC), which are the results of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale. The MFCCs are proved more efficient [1] [3]. Therefore, here we are using MFCC for spectral feature extraction.

The acoustic features will be modeled by Gaussian mixture models, GMMs, on the frame level. Survey indicates that using GMM on the frame level is a feasible technique for emotion classification. Also Gaussian modeling is among the best methods to distinguish emotional classes in a space spanned by the following phonetic parameters: pitch, pitch range, average pitch, all measured across the entire utterance after end pointing (i.e. pause/speech boundary detection) [1]. Therefore, GMM algorithm is best for spectral feature classification

2.1 Voice Recognition Algorithms

A voice analysis is done after taking an input through microphone from a user. The design of the system involves manipulation of the input audio signal. At different levels, different operations are performed on the input signal such as Pre-emphasis, Framing, Windowing, Mel Cepstrum analysis and Recognition (Matching) of the spoken word.

The voice algorithms consist of two distinguished phases. The first one is training sessions, whilst, the second one is referred to as operation session or testing phase as described in figure 1 [2][7]

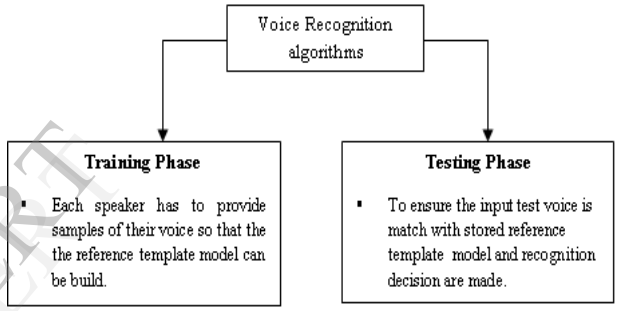

Fig 2. Phases of reconition system[2]

2.2 Database of Emotion codebooks

Like any other recognition systems, emotion recognition systems also involve two phases namely, training and testing. Training is the process of familiarizing the system with the emotions characteristics of the speakers. Testing is the actual recognition task. The block diagram of training phase is shown in Fig.2.1. Feature vectors representing the emotion characteristics of the speaker are extracted from the training utterances and are used for building the reference models. During testing, similar feature vectors are extracted from the test utterance, and the degree of their match with the reference is obtained using some matching technique. The level of match is used to arrive at the decision. [3]

3. Feature extraction

3.1Phonetic feature extraction

This feature extraction can be implemented in many ways, but a very common, is to use Mel-based Cepstral Coefficients. Mel Frequency Cepstral Coefficients (MFCC) is the most widely used spectral representation of the speech signal in many applications, such as speech recognition and speaker recognition. These are based on an (fast) Fourier transform, followed by a non-linear warp of the frequency axis, the logarithm of the power spectrum, and the evaluation of the first N coefficients of this log warped power spectrum in terms of cosine basis functions.

So it is said that emotion recognition algorithm that use prosodic features are not sufficiently accurate. However, phonetic feature have less information for discriminating emotions. Actually there is more independent component in the phonetic features of speech than in prosodic features of speech. E.g. 12-16 dimensional Mel-Frequency Cepstral Coefficient (MFCCs) have been used as the effective phonetic features for speech recognition. If even a small amount of useful information is kept in the phonetic feature, the accuracy of emotion recognition can be improved by increasing the number of independent phonetic features.

Therefore, an emotion recognition algorithm that focuses more on precise classification of the MFCCs. To realize such a precise classification, we will give the emotion label to each frame using multi-template MFCC clustering. The algorithm is simple enough to realize immediate response even in a low-end computer, as well as the higher accuracy than the conventional method.

3.2 Feature Extraction using MFCC

MFCC is based on the human peripheral auditory system. The human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency *t* measured in Hz, a subjective pitch is measured on a scale called the 'Mel Scale' .The Mel frequency scale is a linear frequency spacing below 1000 Hz and logarithmic spacing above 1kHz.As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels.

First the voice data is divided into frame. Each frame is windowed using Hamming window. Second the analysis frame is converted to the frequency domain using a short time Fourier Transform. Third a certain number of sub-band energies are calculated using a Mel filter bank, which is a non linear- scale filter bank that imitates a human's aural system. Fourth, the logarithm of the sub-band energies is calculated. Finally, the MFCC is computed by an inverse Fourier Transform.

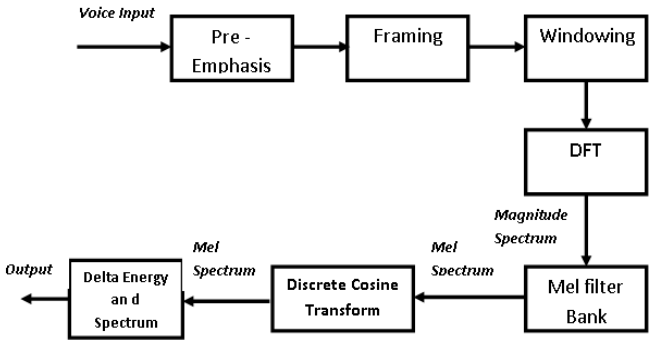

Fig. 3. MFCC Block Diagram [2]

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.

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

Lets consider a = 0.95, which make 95% of any one sample is presumed to originate from previous sample.

Step 2**:** Framing
The process of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 msec. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M (M<N). Typical values used are M = 100 and N= 256.

Step 3: Hamming windowing
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 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:

$$Y(n) = X(n) \times W(n) \quad (2)$$

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) 0 \le n \le N-1 \quad (3)$$

Step 4**:** Fast Fourier Transform
To convert each frame of N samples from time domain into frequency domain. The Fourier Transform is to convert the convolution of the glottal pulse U[n] and the vocal tract

impulse response H[n] in the time domain. This statement supports the equation below:

$$Y(w) = FFT\ [h(t) * X(t)] = H(w) * X(w) \quad (4)$$

If X (w), H (w) and Y (w) are the Fourier Transform of X (t), H (t) and Y (t) respectively.

Step 5**:** Mel Filter Bank Processing
The frequencies range in FFT spectrum is very wide and voice signal
does not follow the linear scale. The bank of filters according to Mel
scale as shown in figure 4 is then performed

.

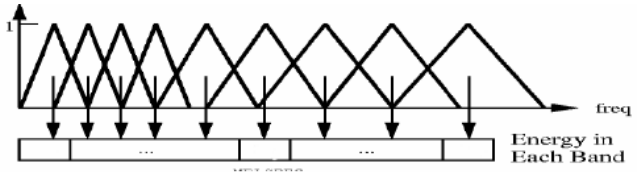

Fig. 4. Mel scale filter bank, from (young et al,1997)[2]

This figure shows a set of triangular filters that are used to computea weighted sum of filter spectral components so that the output ofprocess approximates to a Mel scale. Each filter's magnitude frequencyresponse is triangular in shape and equal to unity at thecentre frequency and decrease linearly to zero at centre frequency oftwo adjacent filters [7, 8]. Then, each filter output is the sum of itsfiltered spectral components. After that the following equation isused to compute the Mel for given frequency f in HZ:

$$F(Mel) = [2595 * \log 10[1 + f]700] \quad (5)$$

Step 6: Discrete Cosine Transform
This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector

Step 7: Delta Energy and Delta Spectrum
The voice signal and the frames changes, such as the slope of a formant at its transitions. Therefore, there is a need to add features related to the change in cepstral features over time . 13 delta or velocity features (12 cepstral features plus energy), and 39 features a double delta or acceleration feature are added. The energy in a frame for a signal x in a window from time sample t1 to time sample t2, is represented at the equation below:

$$Energy = \sum X^2[t] \quad (6)$$

Each of the 13 delta features represents the change between frames in the equation 8 corresponding cepstral or energy feature, while each of the 39 double delta features represents the change between frames in the corresponding delta features.

$$d(t) = \frac{c(t+1) - c(t-1)}{2} \quad (7)$$

4. CLASSIFICATION

4.1Frame Level Classification

In this section we will attempt to describe the method for classifying analysis frames. Each emotion is expressed by a codebook, and each codeword is represented as a vector in the feature space. When we have an input feature vector, we calculate the distance between the input and all the code words. Finally, the emotional label of the nearest codeword becomes the classification result of the analysis frame

4.2 Gaussian Mixture Model (GMM)

Figure 7 shows the HMM with one emitting state. A speech starts from a start state, and stays at an emitting state for a while and finally ends at an end state. While staying in the emitting state, several observations (features) which follow a Gaussian mixture model (GMM) probability are generated. The feature vectors extracted from the speech can be described using this model. The feature vectors follow the Gaussian mixture model (GMM) probability in the emitting state and each person has a unique probability model.

4.2.1 Illustration of GMM

At present, Gaussian mixture model (GMM) often to be used to the speaker recognition, this model has the good ability of recognition. In this work, the Gaussian mixture model (GMM) is adopted to represent the distribution of the features. Under the assumption that the feature vector sequence $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ is an independent identical distribution (i.i.d) sequence, the estimated distribution of the *D*-dimensional feature vector x is a weighted sum of *M* component

A GMM is a weighted sum of M component densities and is given by the form

$$p(X / \lambda) = \sum_{i-1}^{N} c_i.b_i(x).....(1)$$

Where x is a dimensional random vector,
bi(x), i = 1,. . ., N, is the component densities and
ci, i = 1,. . .,N, is the mixture weights.
Gaussian function of the form

$$b_i(x)=1/\{(2\pi)^{d/2} |\sum_i|^{1/2}\}\exp\{-1/2(x - \mu_i)^r \sum_i^{-1} (x - \mu_i)\} .....(2)$$

with mean vector _i and covariance matrix _ *i*
The mixture weights satisfy the constraint that:

$$\sum_{i=1}^{N} c_{i=1}.....(3)$$

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation: $\hat{\lambda} = \{ci, \mu_i \Sigma_i\}$, _
In speaker recognition system, each speaker is represented by such a GMM and is referred to by this model
For a sequence of T test vectors X = x1, x2. . . xn, the standard approach is to calculate the GMM likelihood in the log domain as:

$$L(X | \lambda) = \sum_{i=1}^{r} \log(x_i | \lambda_i).....(4)$$

The emotion-specific GMM parameters are estimated by the EM algorithm using training data uttered by the corresponding speaker using the HTK toolkit.
Gaussian densities *Ni*(x), each parameterized by a mean vector *μi* and covariance matrix K*i*; the mixture density for the model ¤m is defined

$$f(x | \Lambda_m) = \sum_{i=1}^{M} p_i.N_i(x)$$

We will use the expectation maximization (EM) algorithm for the mixtures to get maximum likelihood as explained below

Given a collection of training feature vectors, maximum likelihood model parameters will be estimated using an iterative expectation–maximization (EM) algorithm]. The EM algorithm iteratively refines the GMM parameters to monotonically increase the likelihood of the estimated model for the observed feature vectors. Generally, five iterations are sufficient for parameter convergence. The EM equations for training a GMM can be found in the reference papers
After parameter estimation, we will determine which category the test emotional speech belongs to. By computing the likelihood of all emotional speech models and finding the model which has a maximum likelihood value, we can categorize the test sample of speech. The likelihood of speaker λ is

$$likelihood = \sum_{t=1}^{T} \log[p(x_t | \lambda)],$$

where *T* is the number of frames and *t x* is the feature vector from the *t* -th frame The probability of *t x* given the speaker model _ is

$$. p(x_t | \lambda) = \sum_{i=1}^{M} w_i^{\lambda} p_i^{\lambda}(x)$$

$$p_i^{\lambda}(x) = \frac{1}{(2\pi)^{D/2} |\Sigma_i^{\lambda}|^{1/2}} \exp\left[-\frac{1}{2}(x_t - \mu_i^{\lambda})^T (\Sigma_i^{\lambda})^{-1}(x_t - \mu_i^{\lambda})\right]$$

$$w_i^{\lambda}, \Sigma_i^{\lambda}, \mu_i^{\lambda}$$

denote the weight, thecovariance matrix and the mean vector of the *i* -thGaussian of the speaker model _ , respectively.

We choose the model λ for the test speech $x = \{x_1, x_2, ..., x_T\}$ by $\arg\max_{\lambda} = \sum_{t=1}^{T} \log[p(x_t | \lambda)]$

4.3 Machine Learning Algorithm

In this work, we will utilize the statistics of pitch and energy as prosodic features. We will extract the pitch and energy contours for a given segment, and calculate their statistics to construct feature vector. The statistics used are mean, standard deviation, maximum, minimum, median, and jitter. Only one feature vector is generated per utterance.
In this work, the k nearest neighborhood algorithm (*k*-NN) is chosen to model the prosodic features. It is a simple and nonparametric machine learning algorithm, which classifies the input based on the prototypes in the training data.
The posteriori probability that given a feature vector x belongs to class *m* is

$$P(\Lambda_m | x) = P_m(x) = N_m/k$$

Where, *Nm* denotes the number of prototypes which belong to the class *m* among the *k* nearest prototypes

Since *k*-NN is based on Euclidean distance, we will normalize each component of prosody feature in the training data so that it has zero mean and unit standard deviation.

4.3.1 K –Nearest Neighborhood

The goal of this clustering method is to simply separate the data based on the assumed similarities between various classes. Thus, the classes can be differentiated from one another by searching for similarities between the data provided.

A distance is assigned between all points in a dataset. Distance is defined as the Euclidean distance between two points or:

$$d = \sqrt{\sum_{i=0}^{i=n} (x_i - y_i)^2}$$

From these distances, a distance matrix is constructed between all possible pairings of points (x, y).
Each data point within the data set has a class label in the set, C= {c1, ……, cn}.

The data points', k-closest neighbors (k being the number of neighbors) are then found by analyzing the distance matrix. The k-closest data points are then analyzed to determine which class label is the most common among the set. The most common class label is then assigned to the data point being analyzed.

In the case where two or more class labels occur an equal number of times for a specific data point within the dataset, the KNN test is run on K-1 (one less neighbor) of the data point in question. This is a recursive process. If there is again a tie between classes, KNN is run on K-2. This continues in the instance of a tie until K=1. When K=1 there is only one class represented in the outcome and thus there can be no tie. These resulting class labels are used to classify each data point in the data set.

4.3.2 k-NN -- The Nearest Neighbor

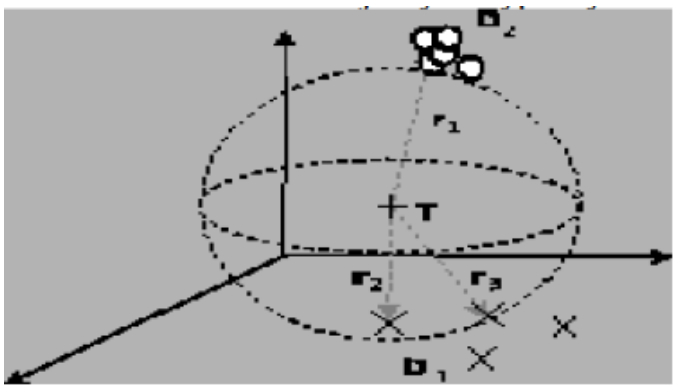

Fig 5 .vector classification using k-NN[1]

Classification using k-NN, k = 3. The test sample T is being classified as x, because in the hyper cycle surrounding T are 2 elements from x and only one from o.

Let us assume we have sets **Di**, these represent **c** classes, **k>0** and test sample **x**. We want to classify **x** as a member of one of classes **Di**,, k-NN does this very simply.
- find k closest vectors to test vector x, let these are r1,r2, ……., rn
- make a hyper cycle Cn around x with radius r, r = max i=1,…,k| ri|

$$\text{classify } x \text{ as: } Di \ : Si \subseteq Di \text{ where } Si = \text{argmax } Ri$$
$$|Ri| \ \& \ Ri = \{ri : \| ri - x \| \leq r \wedge ri \in Di\}$$

The last step says: classify input vector x as the member of the class which has the majority in hyper cycle Cn.

In kNN, prior to testing a sample against the data, all samples of that person was should be removed to ensure that no match would occur due to similarity of voice rather than emotion. However, when tested, this will have only slight effect on the results

3) Information Fusion Algorithm:

Starting with a simple binary classification theory, a classifier function C(¢) which takes a input speech s will yield a result as to which hypothesis the speech belongs based on a threshold comparison, i.e.,

$$C(s) = \begin{cases} H_0 & ;\xi \geq \theta \\ H_1 & ;\xi < \theta \end{cases},$$

Where

$$\xi = S_0 - S_1$$

and *Sm* represents likelihood that the speech s belongs to *Hm*.
In this work, we set the hypothesis as:
*2 H*0: the input speech is of one emotional status
*2 H*1: the input speech is of another emotional status.
The decision arising from the spectral and prosodic feature classifiers need to be combined in order to have a unique and more accurate classification. Many algorithms have been proposed to deal with multiple modalities. One of the most simple and popular methods is a weighted sum of likelihoods from different modalities with a weighing factor that will be empirically determined.

5. RESULT

Here in referred paper they obtained some 1500 features, which partly consist of frequently used features but also introduce new experimentally designed features into the analysis. All features were calculated on a 10ms frame shift rate. Table 1 shows the different feature information sources

and the number of features calculated from them. Many methods are developed for feature extraction but the table below signifies that MFCC gives better accuracy than any other method.

Table 1. Information sources, number of features calculated, and Average Accuracy. [5*]*

Features	Number of Features	Average Accuracy
ZCR,elongation, duration, correlation	10	61.5%
Intensity	171	68.9%
MFCC	576	71.1%
Loudness	171	67.6%
Formants	216	65.4%
Spectrum	135	63.6%
Pitch	236	62.6%
Linguistic features	11	49.9%
Inverse filtering	33	64.3%

6. CONCLUSION

Automatic detection of emotions will be evaluated using standard Mel-frequency Cepstral Coefficients, MFCCs. These acoustic features will be modeled by Gaussian mixture models (GMMs), on the frame level. Survey indicates that using GMM on the frame level is a feasible technique for emotion classification. Also Gaussian modeling is among the best methods to distinguish emotional classes by the following phonetic parameters: pitch, pitch range, average pitch, all measured across the entire utterance.

As a result of changes in shape of human vocal tract during generation of different emotions, resonance frequencies of vocal tract, formants, also changes. Using this phenomenon, we can extract voice features of each emotion and we can implement an emotion detection system.

References

[1] S. D. Shirbahadurkar, A. P. Meshram, Ashwini Kohok & Smita Jadhav, "*An Overview and Preparation for Recognition of Emotion from Speech Signal with Multi Modal Fusion*" IEEE Proceedings, Vol.5., 2010.

[2] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "*Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques*", Journal Of Computing, Volume 2, Issue 3, ISSN 2151-9617, , March 2010.

[3] Vibha Tiwari, "*MFCC and its applications in speaker recognition*", International Journal on Emerging Technologies, ISSN : 0975-8364, 2010.

[4] Mahdi Shaneh, and Azizollah Taheri, "*Voice Command Recognition System Based on MFCC and VQ Algorithms*", World Academy of Science, Engineering and Technology, 2009.

[5] Florian Metze, Tim Polzehl and Michael Wagner, "*Fusion of Acoustic and Linguistic Speech Features for Emotion Detection*", IEEE International Conference on Semantic Computing, 2009.

[6] Ashish Jain,Hohn Harris,*Speaker identification using MFCC and HMM based techniques*,university Of Florida,April 25,2004.

[7] Cheong Soo Yee and abdul Manan ahmad, *Malay Language Text Independent Speaker Vertification using NN ‐ MLP classsifier with MFCC, 2008* international Conference on Electronic Design.

[8] P. Lockwood, J. Boudy, *Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the Projection, for Robust*
Speech Recognition in Cars, Speech Communication, 1992.

[9] A. Rosenberg, C. ‐ H. Lee, F. Soong, *Cepstral Channel Normalization*
Techniques for HMM ‐ Based Speaker Verification, 1994.

[10] Dr Philip Jackson, *Features extraction 1.ppt*,, University of Surrey,
guilford GU2 & 7XH.

[11] Zaidi Razak,Noor Jamilah Ibrahim, emran mohd tamil,mohd Yamani Idna Idris, Mohd yaakob Yusoff,*Quranic verse recition feature extraction using mel frequency ceostral coefficient (MFCC)*,Universiti Malaya.

[12] http://www.cse.unsw.edu.au/~waleed/phd/html/node38.html, downloaded on 3rd March 2010
.

[13] Jamal Price, sophomore student, *Design an automatic speech recognition system using maltab*, University of Maryland Estern Shore Princess Anne.

[14] Ahmad Kamarul,Ariff Bin Ibrahim, *Biomedical engineering labiratory*
student pack,UTM Jjohor

[15] E.C. Gordon,*Signal and Linear System Analysis*.John Wiley & Sons Ltd., New York, USA,1998.

[16] Stan Salvador and Pjilip Chan,*FastDTW: Toward Accurate Dynamic*
Time Warping in Linear time space,Florida Institute of Technology,Melbourne.