International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 04 | Apr-2018 www.irjet.net p-ISSN: 2395-0072

Emotion recognition using Speech Signal: A Review

Dhruvi desai

ME student, Communication System Engineering (E&C), Sarvajanik College of Engineering & Technology Surat, Gujarat, India.

Abstract *- Speech is one of the most natural communications between human beings. Humans also express their emotion via written or spoken languages. Speech emotion is an important role to express the feeling of one's expression. If one wants to understand the meaning of an utterance, it must be in proper emotion otherwise semantically utterance will go to the wrong direction and give wrong result. This Review paper includes different speech features consisting of fundamental frequency (f_0), energy signal, Zero Crossing Rate (ZCR), MFCC (Mel Frequency Cepstrum Coefficient), LPC (Linear Predictor Coefficient) along with the suitable classification schemes like DTW (Dynamic Time Wrapping), SVM (Support Vector Machine), GMM (Gaussian Mixture Model) and K-NN (K-Nearest Neighbour). Speech emotion recognition is particularly useful for applications which require natural man-machine interaction. Conclusion about the performance of speech emotion recognition system is discussed in the last section along with possible ways of improving speech emotion recognition systems.*

Keywords— Energy signal, ZCR, MFCC, LPC, DTW, ANN, K-NN,GMM

1. INTRODUCTION

Speech is the communication or expression of thoughts in spoken words. The speech signal is the fastest and the most natural method of communication between human. Speech emotion recognition is to identify emotion in spoken languages and convert it in to machine language. Speech emotion recognition is defined as extracting the emotional state of a speaker from his or her speech. It can be used to extract useful semantics from speech, and hence improves the performance of speech recognition systems.

The feature extraction like MFCC (Mel Frequency Cepstrum Coefficient) provide the highest accuracy on all databases provided using the linear kernel [6] and the spectral coefficients derived from LPC (Linear Predictive Coding) [7], classifier Technique like ANN (Artificial Neural Network)[1], GMM (Gaussian Mixture Model)[2], K-NN (K-Nearest Neighbor)[2], SVM (Support Vector Machine),Hidden Markov Model (HMM) and Vector Quatization(VQ)[14] etc. In fact, there has been no agreement on which classifier is the most suitable for emotion classification. In recent year, the extremely complex nature of human emotional states makes this problem more complicated in terms of feature selection and classification. The acoustic variability introduced by the existence of different sentences, speakers, speaking styles, and speaking rates adds another obstacle because these properties directly affect most of the common extracted speech prosody such as pitch, intensity and energy. Quality features like formant frequencies and spectra-temporal features have been made by several researchers. First human voice is converted into digital signal form to produce digital data representing each level of signal at every discrete time step. After that digitized speech samples are processed using combination of features like pre-processing to produce voice features. Then after these voice features can go through to select the matches the database and find classification between each reference database and test input file in order to minimize the resulting error between them on feature.

Speech emotion recognition is useful for application in car broad system which gives the information of the mental state of the driver and provide to the system to initiate his /her safety, E-tutoring applications would be more practical, if they can adapt themselves to listener's or student's emotional states. It is also useful to required natural men-machine interaction such as computer tutorial application where the response of this system to the user depends on the detected emotion. It may be also useful in call center application and mobile communication.

The rest of this paper is organized as follows: speech emotion recognition system in section-II, review in detail speech features extraction technique section-III, classifier technique section-IV and final conclusion in section-V.

2. Speech Emotion Recognition System

The basic block diagram of the speech emotion recognition system is illustrated in Fig.1 [15].

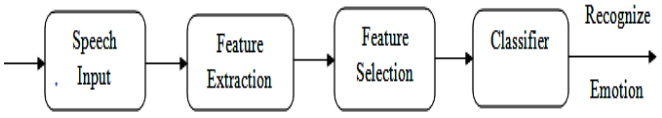

Fig 1: Block Diagram of Speech Emotion Recognition [15]

The main elements of the speech emotion recognition system are same as any typical pattern recognition system. It consists of the emotional speech as input, feature extraction, classification of the emotion using classifier and recognized emotion as the output.

The different emotion from the different speech signals generated by controlled environment. It uses a combination of features based on Short Time Energy (STE), start-point end-point detection, Mel Frequency Cepstral Coefficient (MFCC) is to detect the nearest recorded emotion from database.

2.1 Short Time Energy (STE)

The energy content of a set of samples is related by the sum of the square of the samples. It gives little information about the time-dependent properties of the speech signal. It is very useful to select the threshold for start-point & end-point detection. To calculate STE the speech signal is sampled using a rectangular window function of width ω samples, where $\omega \ll n$. Within each window, energy is computed as follows [13]:

$$e = \sum_{i=1}^{w} x_i^2 \qquad (1)$$

Where, e = Energy of a particular window

x_i = Indicating i^{th} sample

2.2 Zero Crossing Rate (ZCR)

A zero crossing is said to occur if successive samples have different algebraic signs. In other words we can say that ZCR of an audio signal is a consistent of the number of times the signal crosses the zero amplitude line by passage from a positive to negative or vice versa. It is very useful for detecting voiced and unvoiced part of a signal as well as for the start-point & end-point detection [12]. The audio signal is divided into temporal segments by the rectangular window function and zero crossing rate for each segment is computed as below, where $sgn(x_i)$ indicates the sign of the i^{th} sample and can have three possible values: +1, 0, −1 depending on whether the sample is positive, zero or negative.

$$z = \sum_{i=1}^{w} \frac{|sgn(x_i) - sgn(x_i - 1)|}{2} \qquad (2)$$

Where, ( ) = 1 ( ) ≥ 0

= −1 (x_i) < 0

2.3 Start-point & End-point Detection

Computation of these points is more beneficial as they are used to remove background noise and made voice signal better than previous signal. Start point of any voice signal provide the exact starting location of voice sample based on STE and ZCR values, so that all previous unwanted samples would be removed and new voice signal would be created. This same process is also applied to detect end points of any voice signal.

3. Features Extraction Technique

Feature extraction is a special form of dimensionality reduction. The main Purpose of feature extraction is to extract characteristics that are unique to each individual. In other words, they have represented that the feature extraction is a process of extracting best parametric representation of signals in order to produce a better performance of recognition emotion.

As survey shows that the MFCC gives good results as compare to other features for speech based emotion recognition system [11]. MFCC is the most axiomatic and popular feature extraction technique for speech emotion recognition. It approximates the human system response more closely than any other system because frequency bands are placed logarithmically here. The overall stepwise process of the MFCC is described and shown in Fig.2 [6].

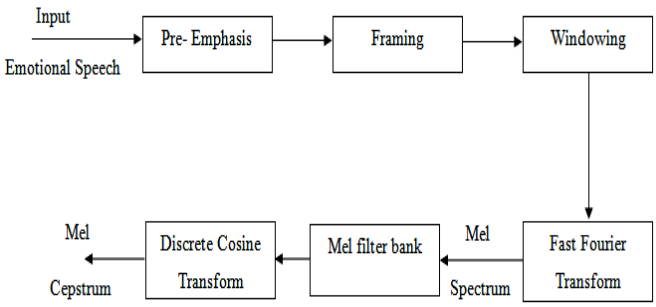

Fig 2. Steps involves in MFCC [6]

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.

$$Y[n] = X[n] - aX[n-1] \qquad 0.9 \le a \le 1.0 \qquad (3)$$

The Pre-emphasizer is implemented as a fixed coefficient filter or as an adaptive one, where the coefficient is adjusted with time according to the auto-correlation values of the speech. The aim of this stage is to boost the amount of energy in the high frequencies. The drop in energy across frequencies is caused by the nature of the glottal pulse. Boosting the high frequency energy makes information from these higher formants available to the acoustic model.

Step 2: Framing

The process of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length with-in 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).

Step 3: Windowing

In this step processing is to hamming window each individual frame minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by the window to taper the signal to zero at the beginning and end of each frame.

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. If the window is define as,

$$w(n), \quad 0 \leq n \leq N-1,$$

(4)

Where N = number of samples in each frame

$$y(n) = x(n)w(n), \qquad 0 \leq n \leq N-1$$

(5)

Where $y(n)$ = Output signal

$x(n)$ = Input signal

$w(n)$ =Hamming window, then the result of windowing signal is:

$$w(n) = 0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1 \quad (6)$$

Step 4: Fast Fourier Transform (FFT)

To convert the signal from time domain to frequency domain preparing to the next stage (mel frequency wrapping). When we perform FFT on a frame, we assume that the signal within a frame is periodic, and continuous when wrapping around, the equation below:

$$Y(w) = FFT[h(t) \cdot x(t)] = H(w) * X(w) \qquad (7)$$

Step 5: Mel Filter Bank

Psychonomics studies have shown that 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, f, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. The mel frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Therefore they have used the following approximate formula to compute the mels for a given frequency fin Hz:

$$F(mel) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \qquad (8)$$

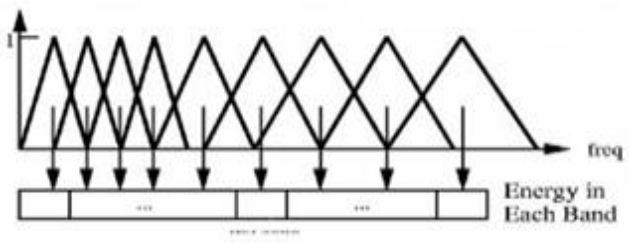

Fig.3 Mel filter bank [12]

The subjective spectrum is to use a filter bank, spaced uniformly on the mel scale. The filter bank has a triangular band pass frequency response. The bandwidth of each filter is determined by the center frequencies of the two adjacent filters and is dependent on the frequency range of the filter bank and number of filter chosen for design. The bank of filters corresponding to Mel scale as shown in Fig.3 [12].

This above Fig.3 depicts a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter's magnitude frequency is triangular in shape and equal to unity at the center frequency and decline linearly to zero at center frequency of two adjacent filters. Then, each filter output is the sum of its filtered spectral components.

Step 6: Discrete Cosine Transform (DCT)

In this final step, they convert the log Mel spectrum to time domain. The result is called the MFCC (Mel Frequency Cepstrum Coefficients). This representation of the speech spectrum provides a good approximation of the spectral properties of the signal for the given frame analysis. The Mel spectrum coefficients being real numbers are then converted to time domain using Discrete Cosine Transform (DCT).

Linear Predictor coefficient

The LPC method considers a speech sample at time n, s(n) and approximates it by a linear combination of the past p speech samples in the following way[17]:

$$s(n) = a_1 s(n-1) + a_2 s(n-2) + \ldots\ldots a_p s(n-p)$$

(9)

Where $a_1 \ldots a_n$ constant coefficients. The eq^n(9) including by an excitation term G(u(n)).

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + G(u(n)) \qquad (10)$$

Where G is the gain and u(n) is the normalized excitation. Transforming equation (10) to z-domain,

$$s(z) = \sum_{i=1}^{p} a_i z^{-i} s(z) + GU(z)$$

(11)

And the transfer function H(z) becomes,

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1-\sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)} \qquad (12)$$

A linear predictor with coefficients α_k is defined as follows:

$$P(z) = \sum_{k=1}^{p} \alpha_k z^{-k} \qquad (13)$$

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \quad (14)$$

The prediction error e(n) is defined as,

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \quad (15)$$

It is required to minimize this error for good prediction. Initially, speech is segmented into intervals of 10ms, over which speech is assumed to be stationary. For each segment, one has to calculate the LP coefficients. The error minimization, during the calculation of the predictor coefficients, is done over each segment [2]. Segmentation, however, brings with it a problem of its own, that of spectral leakage. To reduce this effect, each frame is multiplied by a hamming window is defined by,

$$w(n) = 0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right) \quad (16)$$

Now, error is minimized in the least square sense, on each frame.

$$E_n = \sum_m {e_n}^2(m) = [s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k)]^2 \quad (17)$$

Differentiate the above equation with respect to a_k to obtain its minimum. Solving this equation gives the prediction coefficients for each frame.

4. Classification Technique

4.1 DTW (Dynamic Time Wrapping)

Dynamic Time Wrapping is based on dynamic programming techniques. This algorithm is to measure the similarity between two time series that can vary in time or speed. This technique is also used to find optimal alignment between sets of twice if a time series can be "deformed" non-linearly by stretching or contraction along its time axis. This deformation between two time series can then be used to find corresponding regions between the two time series or to determine the similarity between the two time series. In Figure 4 shows, the example of how one set of times "deforms" another and each vertical line connects one point of a time series with its similarly similar point in the other time series. The lines have similar values on the y-axis, but have been separated so that the vertical lines between them can be seen more easily.

If both time series in Fig.4 were identical, all of the lines would be straight vertical lines because no warping would be necessary to “line up” the two time series. The warp path distance is a measure of the difference between the two time series after they have been warped together, which is measured by the sum of the distances between each pair of points connected by the vertical lines in Fig.4.

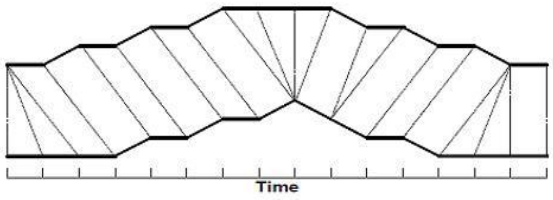

Fig.4: A Warping between two time series [11]

Thus, two time series that are identical except for localized stretching of the time axis will have DTW distances of zero. The principle of DTW is to compare two dynamic patterns and measure its similarity by calculating a minimum distance between them. The classic DTW is computed as below [11].

$$Q = q_1, q_2, \ldots\ldots, q_i, \ldots\ldots q_n,$$

$$C = c_1, c_2, \ldots\ldots c_j, \ldots . c_m$$

To align two sequences using DTW, an n-by-m matrix where the (i^{th}, j^{th}) element of the matrix contains the distance d (q_i,c_j) Between the two point’s q_i and c_j is constructed. Then, the absolute distance between the values of two sequences is calculated using the Euclidean distance calculation.

$$d(q_i, c_j) = (q_i - c_j)^2 \quad (18)$$

Each matrix element (i,j) corresponds to the alignment between the points q_i and c_j. Then, accumulated distance is measured by,

$$D(i, j) = \min[D(i-1, j-1), D(i-1, j), D(i, j-1)] + d(i, j) \quad (19)$$

4.2 Artificial Neural Network (ANN)

Artificial neural networks are the result of academic research using mathematical formulations to model the operations of the nervous system. The resulting techniques are being successfully applied in a variety of everyday business applications. The neural network is used to learn patterns and relationship in the data. The data can be the results of any research. Neural networks do not require explicit coding of problems. The neural network sorts through this information and produces and understanding of the factors affect sales. A very important tool for study the structure-function relationship of the human brain, Due to the complexity and incomplete understanding of biological neurons, several architectures have been reported in the literature. The ANN structures used for many applications often consider the behaviour of a single neuron as the basic computational unit for describing neural information processing operations. Each computing unit, i.e. the artificial neuron in the neural network is based on the concept of an ideal neuron. The goal of neural networks is to mimic the human ability to adapt to changing circumstances and the current environment. This depends to a great extent on being able to learn from the events that have occurred in the past and to be able to apply this to future situations [1].

4.3 Support Vector Machine (SVM)

Support Vector Machine classifiers are mainly based on the use of kernel functions to nonlinear mapping of input patterns into more dimensional space.

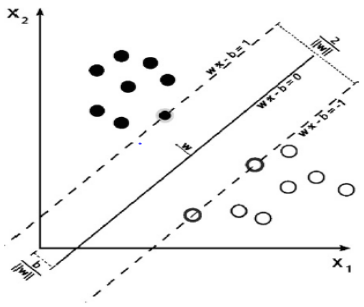

Fig.5 Maximum – margin hyperplane [14]

Support Vector Machine (SVM) classifiers are also used for the problem of emotion recognition using speech and speaker recognition. Basically the idea of a support vector machine depends on two mathematical operations [14]:

(i) Nonlinear mapping of input patterns into high-dimensional feature space.

(ii) Construction of optimal hyper plane for linearly separating the features discovered in step (i).

The region delimited by these two hyper planes is called the "margin", and the maximum-margin hyper plane is the hyperplane that is halfway between them. SVM starts with the easiest classification problem: binary classification of linearly separable data. The training dataset of n points of the form,

$$(\overrightarrow{x_1}, y_1), \dots\dots\dots, (\overrightarrow{x_n}, y_n) \qquad (20)$$

Where $x_i = [x_{i1}\, x_{i2} \dots x_{in}]^T$ is an n- dimensional input vector for the ith example a real- valued space $X \subseteq R^n$; y_i is its class label, and y_i is either +1 or -1 , +1 denoted is class 1 and -1 is denoted class 2. SVM finds a linear function of the form,

$$f(x) = w^T x + b \qquad (21)$$

Hence, $f(x)$ is a real valued function $f: X \subseteq R^n \rightarrow R$. $w = [w_1\ w_2 \dots w_n]^T \epsilon R^n$ Is called the weight vector and $b \in R^n$ is called the bias. In principal, Support Vector Machine finds a hyper plane,

$$w^T x + b = 0 \qquad (22)$$

The plane is called the separation hyper plane. The problem of finding the optimal separation hyper plane becomes an optimal problem as follows:

$$maximize\ L(\lambda) = \sum_{i=1}^{p} \lambda_i - \frac{1}{2}\sum_{i,j=1}^{p} \lambda_i \lambda_j y_i y_j x_i^T x_j \qquad (23)$$

$$subject\ to\ \sum_{i=1}^{p} \lambda_i = 0$$

$$0 \leq \lambda_i \leq C; i = 1, \dots, p$$

In SVM use of kernel function, denoted by k:

$$K(x_i, x_j) = \Phi^T (x_i)\Phi(x_j) \qquad (24)$$

Commonly used kernels include:

$$Polynomial\ of\ degree\ d:\ K(x_i, x_j) = (x_i^T x_j + 1)^d \qquad (25)$$

$$Gaussian\ RBF:\ K(x_i, x_j) = exp\left(-\frac{1}{2\sigma}\|x_i - x_j\|^2\right)$$

4.4 K- Nearest Neighbor (K-NN)

The nearest neighbor technique bases the classification of an unknown sample on the "votes" of K of its nearest neighbor rather than on only it's on single closest neighbor. If the error costs are the same for each class, the estimated class of an unknown sample is chosen to be the class most commonly represented in the collection of its K nearest neighbor.

Let, the K nearest neighbors to y be $N_K(Y)$ and $c(z)$ be the class label of z. The cardinality of $N_K(Y)$ is equal to k and the number of classes is l. Then the subset of nearest neighbor within class $j \in \{1, \dots, l\}$is [2]:

$$N_k^j = \{Z \epsilon\ N_K(Y) : c(z) = j\} \qquad (26)$$

The classification result $\{1, \dots, \} * j \in l$ is defined as the majority vote:

$$j^* = argmax_j \left| N_k^j(Y) \right| \qquad (27)$$

4.5 Gaussian Mixture Model (GMM)

Gaussian Mixture Model approach is proposed where the emotions of speech are modelled as a mixture of Gaussian densities. The use of this model is motivated by the interpretation that Gaussian components represent some spectral form dependent general emotion and the ability of Gaussian blends to model arbitrary densities. The advantage of loudspeaker models is mathematical traceability where the density of complete Gaussian mixtures is represented by the mean vectors, covariance matrices and mixing weights of all component densities [2].

Table 1: Summary of literature based on Berlin Emotion Database

Publish (Year)	Types of database	Emotion	Features	Classifier	Conclusion
Elsevier (2011)[1]	Berlin Emotional database	Happy, Anger, Neutral, Sadness Surprised, Fearful	Continuous, Qualitative, Spectral, TEO-Based	GMM,SVM ANN,HMM	MFCC+GMM Accuracy: 74.83-81.94% Average training time is smallest
Elsevier (2015) [2]	Berlin Emotional database	Happy, Anger, Neutral, Sadness Surprised, Fearful	MFCC, Wavelet feature, Pitch features of speech	GMM, K-NN	-GMM technique detect the 'Angry' emotion recognize with rate of 92% rate. -K-NN technique detect the 'Happy' emotion recognize with 90% rate. -GMM is best result.
IEEE (2013) [7]	Berlin, Japanese, Thai	Happy, Anger, Neutral, Sadness, Fearful, Disgust	MFCC, LPC, Pitch, Energy, ZCR	SVM	- Accuracy for Berlin database 89.80%, Japanese database 93.57% and Thai database 98.00% - Best accuracy in Thai database

Table 2: Summary of literature based on Neutral Emotion Database

Publish (Year)	Types of database	Emotion	Features	Classifier	Conclusion
IEEE (2007) [16]	Neutral Speech database	Anger, Happiness, Surprise, Sadness, Fear	MFCC, LPCC	GMMKL, SVM super vector, 3rd order polynomial	- GMMkl divergence Kernel has the best performance in the gender-dependent and gender-independent
IJCSI (2011) [6]	Neutral Speech database	Happy, Sad, Angry, Surprise, Neutral	MFCC, ΔMFCC, ΔΔMFCC	SVM, DTW	- ΔΔMFCC(39) using DTW gives better accuracy

5.CONCLUSION

In this paper, a review of emotion recognition through speech signal system has been given. Speech emotion recognition systems based on the several speech feature and classifiers is illustrated. As per literature the speech signal analyses by using MFCC provide a spectrum factor which represents the exact vocal system for stored speech emotions. MFCC provide a high level of perception of the human voice and achieving high recognition accuracy. The DTW technique was able to authenticate the particular speaker based on the individual information that was included in the speech signal. The recognition accuracy obtain in [13], using Euclidian Distance was 57.5%. The classification accuracy of ANN is fairly low compared to other classifiers. The speed of computation is fast for K-NN classifier makes it as one of the optional techniques that can be used widely if time constraint is critical. The time of computation increased for GMM classifier when the number of speech features increased in training phase.

REFERENCES

[1] Moataz ElAyadi, MohamedS.Kamel, FakhriKarray, "Survey on speech emotion recognition: Features, classification schemes, and databases", Pattern Recognition 44 (2011) 572–587, Elsevier.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[2] Rahul B. Lanjewar, Swarup Mathurkar,Nilesh Patel, "Implimentation and Comparision of speech E.R. system using Gussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques", Procedia Computer Science 49 (2015) 50 – 57,Elsevier.

[3] Lijiang Chen, Xia Mao, Yuli Xue, Lee Lung Cheng, "Speech emotion recognition: Features and classification models", Digital Signal Processing 22 (2012) 1154–1160, Elsevier.

[4] Tin Lay Nwe , Say Wei Foo , Liyanage C. De Silva, "Speech emotion recognition using hidden Markov models", Speech Communication 41 (2003) 603–623, Elsevier .

[5] Anurag jain, Nuour Prakash, S.S.agrawal, "Evaluation of MFCC for Emotion Identification in Hindi Speech", 2011, IEEE Conference.

[6] N. Murali Krishna, P.V.Lakshmi, Y.Srinivas, J.Sirisha Devi, "Emotion Recognition using Dynamic Time Warping Technique for Isolated Words", IJCSI International Journal of Computer Science, ISSN, September 2011.

[7] Thapanee Seehapoch, Sartra Wongthanavasu," Speech Emotion Recognition Using Support Vector Machine",2013,IEEE Conference.

[8] J. Nicholson, K. Takahashi and R. Nakatsu, "Emotion Recognition in Speech Using Neural Networks", Neural Comput & Application (2000)9:290–296, Springer.

[9] Sreenivasa Rao Krothapalli, shashidhar G.Koolagudi,"Emotion Recognition using speech signal",2013,Springer.

[10] Rabiner Lawrence, Juang-Hwang, "Fundamentals of speech recognition", Prentice Hall, New Jersey, 1993,ISBN 0-13-015157-2.

[11] Linddasalwa Muda, Mumtaj Begam and l.Elamvazuthi," Voice Recognition Algorithms Using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Technique", Journal of computing, volume 2, ISSUE 3, March 2010, ISSN 2151-9617.

[12] Purnima Chandrasekar, Santosh Chapaneri, Dr. Deepak Jayaswal," Automatic Speech Emotion Recognition: A Survey", 2014, IEEE Conference.

[13] Nidhi Desai, Prof. Kinnal Dhameliya, Prof. Vijayendra Desai, "Recognizing voice commands for robot using MFCC and DTW", ISSN ,May 2014.

[14] Neural Networks - A Comprehensive Foundation - Simon Haykin. [Online].

[15] Surabhi vaishnav, saurabh Mitra," speech emotion recognition: A Review", International Research Journal of Engineering & Technology (IRJET), volume:03, Issue:04/Apr-2016.

[16] Hao Hu, Ming-XingXu, and Wei Wu, "GMM super vector based SVM with spectral features for Speech Emotion Recognition", 2007 IEEE Conference.

[17] Jiang Hai and Er Meng Joo, "Improved linear predictive coding method for speech recognition," Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint, 2003, pp. *1614-1618* vol.3