Research Article Impact Factor: 0.621 ISSN: 2319-507X
Apurva Taley, IJPRET, 2014; Volume 2 (9): 434-445 IJPRET

INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

EMOTION DETECTION IN SPEECH USING GAUSSIAN MIXTURE MODEL

APURVA. P. TALEY

Department Of Electronics And Telecommunication, K.J Somaiya College Of Engineering, University Of Mumbai.

Accepted Date: 27/02/2014 ; Published Date: 01/05/2014

Abstract: In recent years, research on automatic emotion recognition is growing dramatically due to the development of techniques in computer vision, speech analysis and machine learning. However, humans have a natural ability to recognize the emotion through speech information but the same task of emotion recognition for machines using speech signal is difficult since machines do not have sufficient integellence to analyze emotion from speech. 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. Here in this report study is carried out using Gaussian mixture model classifier used for identification of emotional of speaker’s as we are detecting the emotions in speech i.e whether the speech is a anger, happiness, sad, surprise ,or neutral etc.

Keywords: Emotion recognition, feature extraction, Gaussian mixture model (GMM), MFCC

Corresponding Author: MS. APURVA. P. TALEY

Access Online On:

www.ijpret.com

How to Cite This Article:

Apurva Taley, IJPRET, 2014; Volume 2 (9): 434-445

PAPER-QR CODE

-

INTRODUCTION

Emotion recognition through speech is an area which increasingly attracting attention within the engineers in the field of pattern recognition and speech signal processing in recent years. Automatic emotion recognition paid close attention in identifying emotional state of speaker from voice signal. Emotions play an extremely important role in human life. It is important medium of expressing humans perspective and his or hers mental state to others. Recognition of emotions in speech is a complex task that is further more complicated because there is no unambiguous answer to what the “correct” emotion is for a given speech sample.

Machine can detect what is said by using speaker identification and speech recognition techniques but if we implied emotion recognition system through speech then machine can also detect how it said as emotions plays an important role in rational actions of human being there is a desirable requirement for intelligent machine human interfaces for better human machine communication and decision making. Emotion recognition through speech means detection of the emotional state of human through feature extracted from his or her voice signal. Emotion recognition through Speech is particularly useful for applications in the field of human machine interaction to make better human machine interface. Applications of the emotion recognition system are lie detection in the psychiatric diagnosis, intelligent toys, in aircraft cockpits, in call center’s and in the car board system. In the field of emotion recognition through speech several system are proposed for recognizing emotional state of human being from speakers voice or speech signal. On the basis of some universal emotions which includes anger, happiness, sadness, surprise, neutral, disgust, fearful, stressed etc. for this different intelligent systems have been developed by researchers in last two decades. This different system also differs by different features extracted and the classifier’s used for classification. For the purpose of feature extraction, spectral analysis algorithm such as Mel-frequency Cepstral Coefficients, MFCCs will be used. The extraction and selection of the best parametric representation of acoustic signals is an important task in the design of any speech recognition system as 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. Therefore, here we are using MFCC for spectral feature extraction.

-

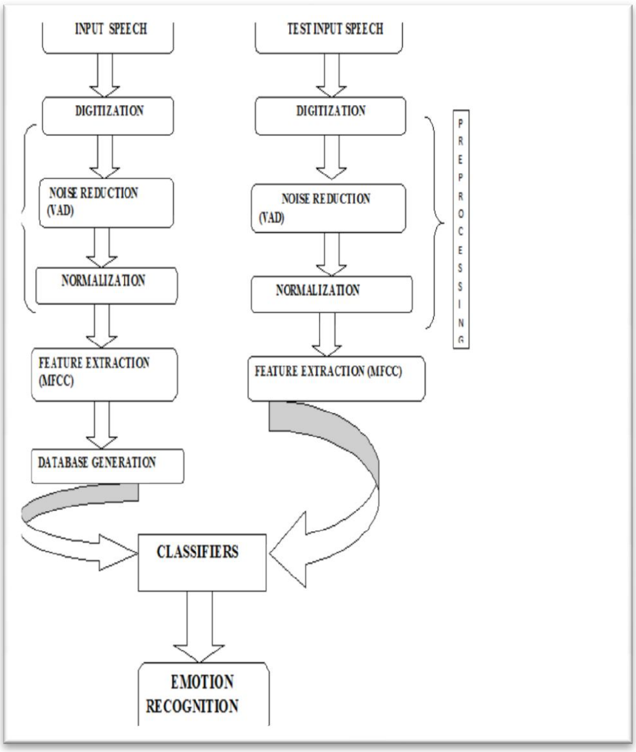


Fig 1.Emotion recognition system

I. Voice Recognition Algorithim

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 2

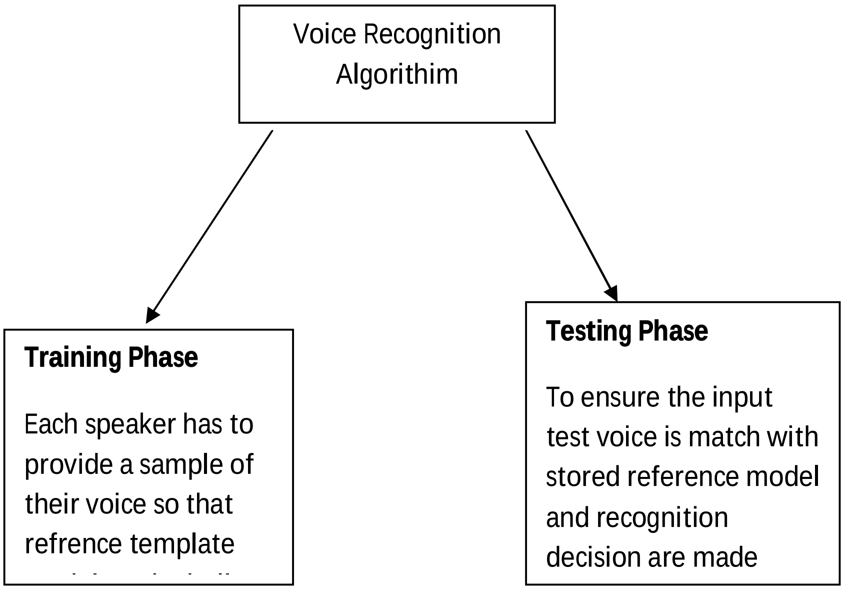

Voice Recognition Algorithim
Training Phase
Each speaker has to provide a sample of their voice so that refrence template
Testing Phase
To ensure the input test voice is match with stored reference model and recognition decision are made

Fig 2. Phases of recognition system

II. 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 Fig2. 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.

III. 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.

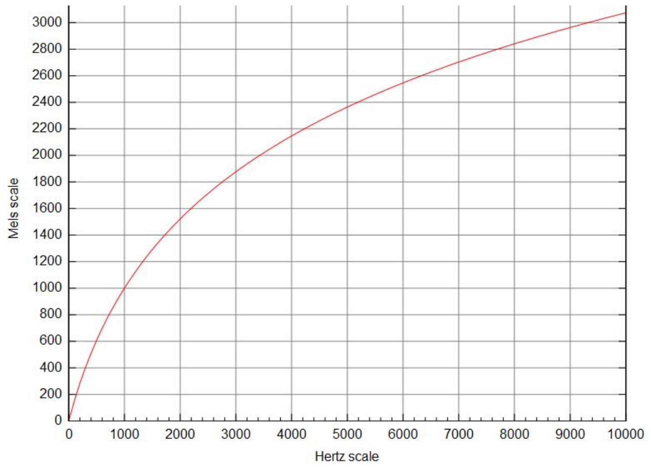
-

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.

F hertz converted to mels as:

$$m = 2595\log_{10}(\frac{f}{700} + 1) = 1127\log_e(\frac{f}{700} + 1)$$



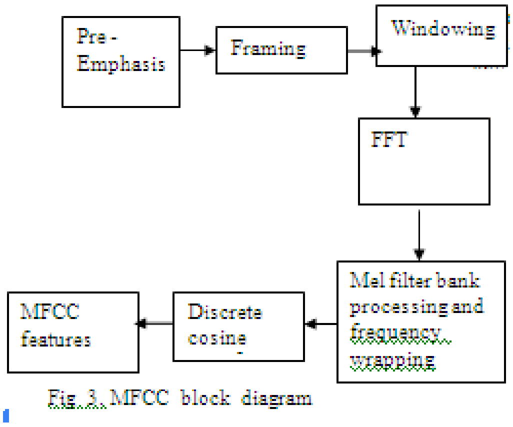

Pre - Emphasis
Framing
Windowing
FFT
Mel filter bank processing and frequency wrapping
Discrete cosine
MFCC features
Fig. 3. MFCC block diagram

Fig 3. MFCC block diagram

-

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. Speech signal sent to a high pass filter

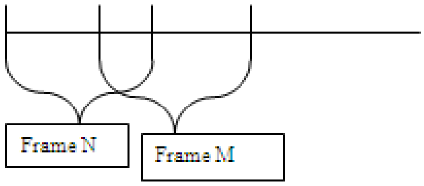
x′(n)= x(n)-a*x(n-1)……………..1

x′(n) is the output of the input signal

a is the pre-emphasized parameter value is in between 0.9 and 1.

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

Here the speech signal s(n) is multiplied by a window w(n) which yeilds a set of speech samples x(n). By shifting w(n) we can examine any part of s(n) through movable window.

So 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 ≤ n ≤ 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

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

-

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

....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:

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

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

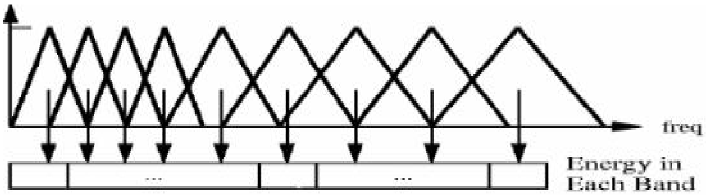


Fig 4 Mel scale filter bank

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

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

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

IV. Classification

The most important aspect of emotion recognition through speech is classification of an emotion. The performance of the system is influenced by the accuracy of the classification, on the basis of different features extracted from the emotional speech samples.

Each emotion is expressed by a codebook, and each codeword is represented as a vector in the feature vector matrix. When we have an input feature vector, we calculate the likelihood between the input and all the code words. Finally, the emotional label of the nearest codeword becomes the classification result.

Gaussian Mixture Model (GMM)

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system. To estimate GMM parameters there are two estimation techniques as Maximum likelihood parameter estimation technique which is computed by using Expectation-Maximization (EM) algorithm and other technique is Maximum A Posteriori (MAP) from a well trained prior model.

In this work Gaussian mixture model(GMM) is adopted to represent the distribution of features. Under the assumption that feature vector sequence X={x1,x2,.....,xn} 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

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

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

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

with mean vector µi and covariance matrix ∑ *i*

The mixture weights satisfy the constraint that:

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

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: λ={ci, µi, ∑i}

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:

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

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 vector.

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 *xt* is the feature vector from the *t* -th frame The probability of *t x* given the speaker model λ is:

$$\cdot\, p(x_t \mid \lambda) = \sum_{i=1}^{M} w_i^{\lambda} p_i^{\lambda}(x)$$

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

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

denote the weight, the covariance matrix and the mean vector of the *i* –th Gaussian of the speaker model λ respectively.

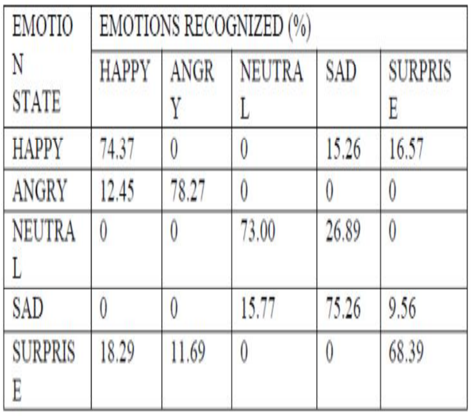
Now for a sequence of T test vector X=x1,x2,.....xn the standard approach to calculate GMM likelihood in log domain is:

$$Log(X/\lambda) = \sum_{i=1}^{r} Log(xt/\lambda t)$$

V. Result

While performing emotion recognition using Gaussian mixture model, first the database is sort out according to the mode of classification. In this study for five modes for five different emotional states features were extracted from the input waveform

Experimental results obtained using GMM



EMOTION STATE	EMOTIONS RECOGNIZED (%)				
	HAPPY	ANGRY	NEUTRAL	SAD	SURPRISE
HAPPY	74.37	0	0	15.26	16.57
ANGRY	12.45	78.27	0	0	0
NEUTRAL	0	0	73.00	26.89	0
SAD	0	0	15.77	75.26	9.56
SURPRISE	18.29	11.69	0	0	68.39

-

VI. 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). Survey indicates that using GMM 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.

REFERENCE

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. Reynolds, D.A., Rose, R.C.: Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models. IEEE Transactions on Acoustics, Speech, and Signal Processing **3**(1) (1995) 72–83

5. McLachlan, G., ed.: Mixture Models. Marcel Dekker, New York, NY (1988)

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. Dempster, A., Laird, N., Rubin, D.: Maximum Likelihood from Incomplete Data via the EMAlgorithm. Journal of the Royal Statistical Society **39**(1) (1977) 1–38

-

9. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing **10**(1) (2000) 19–41

10. 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.

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

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