INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

EXTRACTING MFCC AND GTCC FEATURES FOR EMOTION RECOGNITION FROM AUDIO SPEECH SIGNALS

Minu Babu[1], Dr. Arun Kumar M.N[2], Mrs. Susanna M. Santhosh[3]

[1]MTtech Scholar, Department of Computer Science and Engineering, Federal Institute of Science and Technology (FISAT), Mahatma Gandhi University, Kottayam, Kerala
minubabu4@gmail.com
[2]Associate Professor, Department of Computer Science and Engineering, Federal Institute of Science and Technology (FISAT), Mahatma Gandhi University, Kottayam, Kerala
akmar_mn11@rediffmail.com
[3]Assistant Professor, Department of Computer Science and Engineering, Mar Baselios Institute of Technology and Science (MBITS), Mahatma Gandhi University, Kottayam, Kerala
susannasanthosh@gmail.com

Abstract

Emotion recognition from speech has an increasing interest in recent years given the broad field of applications. The recognition system developed here uses Mel Frequency Cepstrum Coefficient (MFCC) and Gammatone Cepstrum Coefficient (GTCC) as the feature vectors for recognizing emotions in a speech signal. MFCC is the most commonly used feature vector for classification. But, MFCC systems usually do not perform well under noisy conditions because extracted features are distorted by noise, causing mismatched likelihood calculation. By introducing a novel speaker feature, gammatone cepstral coefficient (GTCC), based on an auditory periphery model, and show that this feature captures speaker characteristics and performs substantially better than conventional speaker features under noisy conditions. An important finding in the study is that GTCC features outperform conventional MFCC features under noisy conditions. These features are then used to train the classifier. The classifier used for the system is a cascade feed forward back propagation neural network. The database consists of 240 speech samples. Among them, 180 samples are used for training the system and the remaining 60 samples are used for testing the system. This study compares the differences between MFCC and GTCC for recognizing emotion from speech. The error rate of the system corresponds to MFCC and GTCC is 0.009703 and 0.0090822 respectively.

Keywords*:* Automatic speech recognition, Pre-processing, Feature extraction, Classification, Mel-frequency cepstral coefficient, Gammatone cepstral coefficient.

1. Introduction

Speech is a complex signal which contains information about the message, speaker, language and emotions. Speech is one of the most natural communication forms between human beings. Emotion on other side is an individual mental state that arises spontaneously rather than through conscious effort. Humans also express

their emotion via written and spoken language. Enabling systems to interpret user utterances for a more intuitive human machine interaction therefore suggests also understanding transmitted emotional aspects. The actual user emotion may help system track the user's behaviour by adapting to his inner mental state. Generally recognition of emotions is in the scope of research in the human-machine-interaction. Among other modalities like mimic speech is one of the most promising and established modalities for the recognition. There are several emotional hints carried within the speech signal. The database for the speech emotion recognition system is the emotional speech samples. The classifiers are used to differentiate emotions such as anger, happiness, sadness, surprise, fear, neutral state, etc. The classification performance is based on extracted features. The features extracted from these speech samples are: the energy, pitch, linear prediction cepstrum coefficient, mel frequency cepstrum coefficient. General layout of a Speech Emotion Recognition System is shown in Figure 1.1:

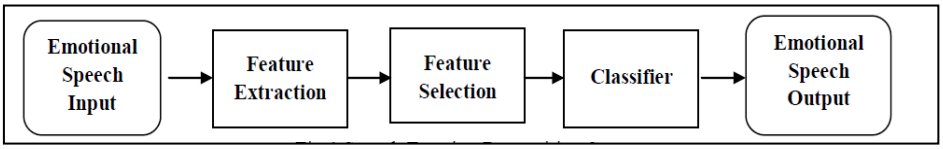

Figure 1.1: General Layout of a Speech Emotion Recognition System

Like typical pattern recognition systems, speech emotion recognition system contains four main modules: speech input, feature extraction, feature selection, classification, and emotion output. Since a human cannot classify easily natural emotions, it is difficult to expect that machines can offer a higher correct classification. Affective computing is a field of research that deals with recognizing, interpreting and processing emotions or other affective phenomena. It plays an increasingly important role in assistive technologies. With the help of affective computing, computers are no longer indifferent logical machines. They may be capable of understanding a user's feelings, needs, and wants and giving feedback in a manner that is much easier for users to accept. Emotion recognition is an essential component in affective computing. In daily communication, identifying emotion in speech is a key to deciphering the underlying intention of the speaker. Computers with the ability to recognize different emotional states could help people who have difficulties in understanding and identifying emotions. Many studies have been conducted in an attempt to automatically determine emotional states in speech. Some of them used acoustic features such as Mel frequency cepstral coefficients (MFCCs) and fundamental frequency to detect emotional cues, while other studies employed prosodic features in speech to achieve higher accuracy of the classification. Various classifiers were applied to recognizing emotions, Hidden Markov Models (HMM), Naïve Bayes classifier, and decision tree classifier.

Deep neural network techniques have recently yielded impressive performance gains across a wide variety of competitive tasks and challenges. For example, a number of the world's leading industrial speech recognition groups have reported significant recognition performance gains through deep network techniques. The large scale image recognition challenge has also recently been won (by a wide margin) through the use of deep neural networks. The Emotion Recognition in the wild challenge is based on an extended form of the acted facial expressions in the wild database in which short video clips extracted from feature length movies have been annotated for different emotions. A core aspect of this approach is the use of a deep convolution neural network for frame-based facial expression classification. To train this model, an additional data composed of images of faces with expressions labelled as one of seven basic emotions (angry, disgust, fear, happy, sad, surprise and neutral). The use of this additional data seems to have made a big difference in the performance by allowing it to train high capacity models without over fitting to the relatively small challenge training data. Importantly, a direct measure of per frame errors on the challenge data does not yield performance that is superior to the challenge baseline; however, the strategy of using the challenge training data to learn how to aggregate the per frame predictions was able to boost performance substantially. These efforts lead to a number of contributions and a number of insights which believe may be more broadly applicable. First, believe that the approach of using the large scale mining of imagery from data image search to train deep neural networks has helped to avoid over fitting in facial expression model. Perhaps counter intuitively, found that the convolution network models learned using only our additional static frame training data sets were able to yield higher validation set performance if the labelled video data from the challenge was only used to learn the aggregation model and the static frames of the challenge training set were not used to train the underlying convolution network. It believed that this effect is also explained in part by the fact that many of the video frames in isolation are not representative of the emotional tag and their inclusion in the training set for the static frame deep neural network classifier further exacerbates the problem of over fitting, adding noise to the training set. The problem of over fitting had both direct consequences on per-model performance on the validation set as well as indirect

consequences on the ability to combine model predictions. The analysis of simple model averaging showed that no combination of models could yield superior performance to an SVM applied to the outputs of our audio-video models. The efforts to create both SVM and Multi-Layer Perceptron (MLP) aggregator models lead to similar observations in those models quickly over fit the training data and no settings of hyper parameters could be found which would yield increased validation set performance. This is due to the fact that the activity recognition and bag of mouth models severely over fit the challenge training set and the SVM and MLP aggregation techniques over fit the data and in such a way that no traditional hyper parameter tuning could yield validation set performance gains. These observations led to develop the novel technique of aggregating the model and per class predictions via random search over simple weighted averages. The resulting aggregation technique is therefore of extremely low complexity and the underlying prediction was therefore highly constrained - using simple weighted combinations of complex deep network models, each of which did reasonably well at this task. As this yielded a marked increase in performance on both the challenge validation and test sets it leads us to the interpretation that given the presence of models that over fit the training data, it may be better practice to search a moderate space of simple combination models compared to more traditional approaches such as searching over the smaller space of SVM hype parameters or even a moderately sized space of traditional MLP hyper parameters including the number of hidden layers and the number of units per layer. In recent seventy years, much research has been done on speech recognition, the human speech processing and converting it into a sequence of words referring to. Although a lot of processing on speech recognition performance, but we are still far from having a natural interaction between human and machine, the machine does not understand human emotion states. This new research field has introduced a sense of speech recognition. Researchers believe that this sense of the speech recognition can be useful to extract meaning from speech and can improve the performance of speech recognition systems. The detected emotions recognized are used in man-machine interfaces to recognize errors in the man machine- interaction by a negative user emotion. If a user seems annoyed after a system reaction, error-recovery strategies are started. On the other hand a joyful user encourages a system to train user models without supervision. First or higher order user preferences can be trained to constrain the potential intention sphere for erroneously recognition instances like speech or gesture input. To do so a system online needs a reference value like a positive user reaction. Furthermore the system initiatively provides help for a seemingly irritated user. Control or induction of user emotions is another field of application that requires the knowledge of the actual emotion. For example in high risk-tasks it seems useful to calm down a nervous person, do not distract her by shortening dialogues, or keep a tired user awake. Other general applications of an emotion recognition system are:

- **Dialog system** for detecting angry users
- **Tutoring system** for detecting student's interest/certainty
- **Lie detection**
- **Social interaction system** for detecting frustration, disappointment, surprise etc

2. Literature Survey

Several studies were conducted in the field of emotion recognition systems. The literature survey for the project includes the earlier works done in the field of emotion recognition from speech. It also describes the various feature vectors and classifiers for recognizing emotion from speech.

2.1 Feature Extraction and Selection

A speech signal composed of large number of parameters which indicates emotion contents of it. Changes in these parameters indicate changes in the emotions. Proper choice of feature vectors is one of the most important tasks in speech recognition. A speech signal composed of large number of parameters which indicates emotion contents of it. Changes in these parameters indicate changes in the emotions. Proper choice of feature vectors is one of the most important tasks in speech recognition. The feature vectors can be classified as long-time and short-time feature vectors. The long-time ones are estimated over the entire length of the utterance. The short-time ones are determined over window of usually less than 100 ms. The long-time approach identifies emotions more efficiently. Commonly used features [1] are energy and related features (the energy is the basic and most important feature in speech signal. The statistics of energy in the whole speech sample can be obtained by calculating the energy, such as mean value, max value, variance, variation range etc.), pitch and related features (the value of pitch frequency can be calculated in each speech frame) and qualitative features (emotional contents of a utterance is strongly related with its voice quality). The acoustic parameters related to speech quality are voice level, signal amplitude, energy and duration, voice pitch, phrase, word, feature boundaries, temporal structures, Linear Prediction Cepstrum Coefficients (LPCC), Mel-Frequency Cepstrum Coefficients

(MFCC) and Perceptual Linear Predictive (PLP) coefficients etc. LPCC embodies the characteristics of particular channel of speech. The Linear Predictive analysis is based on the assumption that the shape of the vocal tract governs the nature of the sound being produced. So these feature coefficients can be used to identify the emotions contained in speech. MFCC is based on the characteristics of the human ear's hearing. It uses a nonlinear frequency unit to simulate the human auditory system. Mel frequency scale is the most widely used feature of the speech, with a simple calculation, good ability of the distinction, anti-noise and other advantages.

In 2013, Dipti D. Joshi and Prof. M. B. Zalte [1] of Mumbai University published a report regarding the various feature vectors and classifiers for emotion recognition from speech. According to their report, feature vectors can be classified as long-time and short-time feature vectors. The long-time ones are estimated over the entire length of the utterance. The short-time ones are determined over window of usually less than 100 ms. The long-time approach identifies emotions more efficiently.

COMMONLY USED FEATRURES

1) Energy and related features

The Energy is the basic and most important feature in speech signal. To obtain the statistics of energy in the whole speech sample by calculating the energy, such as: mean value, max value, variance, variation range, contour of energy.

2) Pitch and related features

The value of pitch frequency can be calculated in each speech frame.

3) Qualitative Features

Emotional contents of a utterance is strongly related with its voice quality. The acoustic parameters related to speech quality are voice level (signal amplitude, energy and duration) ,voice pitch, phrase, word and feature boundaries, temporal structures.

4) Linear Prediction Cepstrum Coefficients (LPCC)

LPCC embodies the characteristics of particular channel of speech. The Linear Predictive analysis is based on the assumption that the shape of the vocal tract governs the nature of the sound being produced. So these feature coefficients can be used to identify the emotions contained in speech.

5) Mel-Frequency Cepstrum Coefficients (MFCC)

MFCC is based on the characteristics of the human ear's hearing. It uses a nonlinear frequency unit to simulate the human auditory system. Mel frequency scale is the most widely used feature of the speech, with a simple calculation, good ability of the distinction, anti-noise and other advantages.

6) Wavelet Based features

Speech signal is a non-stationary signal, with sharp transitions, drifts and trends which is hard to analyze. A time-frequency representation of such signals can be performed using wavelets. For speaker emotional state identification applications the Discrete Wavelet Transform offers the best solution.

2.2 Classifier Selection

Selection of classifier depends on the geometry of the input feature vector. Some classifiers are more efficient with certain type of class distributions. Various Classifiers used are: Hidden Markov Model (HMM), Gaussian Mixtures Model (GMM), Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), and Decision Trees.

1) Hidden Markov Model (HMM)

The Hidden Markov Model (HMM) [2] is a popular statistical tool for modelling a wide range of time series data. In the context of natural language processing (NLP), HMMs have been applied with great success to problems such as part-of-speech tagging and noun-phrase chunking. HMM has been used widely for speech emotion recognition due to its advantage on dynamic time warping capability. That is, its ability to estimate the similarity between two temporal sequences which may vary in time or speed. The Hidden Markov Model (HMM) is a powerful statistical tool for modeling generative sequences that can be characterised by an underlying process generating an observable sequence. HMMs have found application in many areas interested in signal processing, and in particular speech processing, but have also been applied with success to low level NLP tasks such as part-of-speech tagging, phrase chunking, and extracting target information from documents. Andrei Markov gave his name to the mathematical theory of Markov processes in the early twentieth century, but it was Baum and his colleagues that developed the theory of HMMs in the 1960s. However, the classify property of HMM is not satisfactory.

2) Gaussian Mixtures Model (GMM)

GMMs are suitable for developing emotion recognition model when large number of feature vector is available. Gaussian Mixture Models (GMMs) [3] are among the most statistically matured methods for clustering and for density estimation. The GMM and the HMM, are the most used ones for speech emotion recognition. 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. GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm. GMMs are often used in biometric systems, most notably in speaker recognition systems, due to their capability of representing a large class of sample distributions. One of the powerful attributes of the GMM is its ability to form smooth approximations to arbitrarily shaped densities. The classical uni-modal Gaussian model represents feature distributions by a position (mean vector) and a elliptic shape (covariance matrix) and a vector quantizer (VQ) or nearest neighbour model represents a distribution by a discrete set of characteristic templates. A GMM acts as a hybrid between these two models by using a discrete set of Gaussian functions, each with their own mean and covariance matrix, to allow a better modeling capability. The GMM not only provides a smooth overall distribution fit, its components also clearly detail the multi-modal nature of the density. The use of a GMM for representing feature distributions in a biometric system may also be motivated by the intuitive notion that the individual component densities may model some underlying set of hidden classes. For example, in speaker recognition, it is reasonable to assume the acoustic space of spectral related features corresponding to a speaker's broad phonetic events, such as vowels, nasals or fricatives. These acoustic classes reflect some general speaker dependent vocal tract configurations that are useful for characterizing speaker identity.

3) Artificial Neural Network (ANN)

Another common classifier, used for many pattern recognition applications is the artificial neural network (ANN) [4]. There are a large number of different types of networks, but they all are characterized by the following components: a set of nodes, and connections between nodes. The nodes can be seen as computational units. They receive inputs, and process them to obtain an output. This processing might be very simple (such as summing the inputs), or quite complex (a node might contain another network). The connections determine the information flow between nodes. They can be unidirectional, when the information flows only in one sense, and bidirectional, when the information flows in either sense. The interactions of nodes though the connections lead to a global behaviour of the network, which cannot be observed in the elements of the network. This global behaviour is said to be emergent. This means that the abilities of the network super cede the ones of its elements, making networks a very powerful tool.

Networks are used to model a wide range of phenomena in physics, computer science, biochemistry, mathematics, sociology, economics, telecommunications, and many other areas. This is because many

systems can be seen as a network: proteins, computers, communities, etc. An artificial neuron is a computational model inspired in the natural neurons. Natural neurons receive signals through synapses located on the dendrites or membrane of the neuron. When the signals received are strong enough (surpass a certain threshold), the neuron is activated and emits a signal though the axon. This signal might be sent to another synapse, and might activate other neurons. The complexity of real neurons is highly abstracted when modelling artificial neurons. These basically consist of inputs (like synapses), which are multiplied by weights (strength of the respective signals), and then computed by a mathematical function which determines the activation of the neuron. Another function (which may be the identity) computes the output of the artificial neuron (sometimes in dependence of a certain threshold).

ANNs combine artificial neurons in order to process information. The higher a weight of an artificial neuron is, the stronger the input which is multiplied by it will be. Weights can also be negative, so we can say that the signal is inhibited by the negative weight. Depending on the weights, the computation of the neuron will be different. By adjusting the weights of an artificial neuron we can obtain the output we want for specific inputs. But when we have an ANN of hundreds or thousands of neurons, it would be quite complicated to find by hand all the necessary weights. But we can find algorithms which can adjust the weights of the ANN in order to obtain the desired output from the network. This process of adjusting the weights is called learning or training.

There are several types of ANNs and their uses are very high. The differences in them might be the functions, the accepted values, the topology, the learning algorithms, etc. There are a wide variety of ANNs that are used to model real neural networks, and study behaviour and control in animals and machines, but also there are ANNs that are used for engineering purposes, such as pattern recognition, forecasting, and data compression. The back propagation algorithm is used in layered feed-forward ANNs. This means that the artificial neurons are organized in layers, and send their signals “forward”, and then the errors are propagated backwards. The network receives inputs by neurons in the input layer, and the output of the network is given by the neurons on an output layer. There may be one or more intermediate hidden layers. The back propagation algorithm uses supervised learning, which means that we provide the algorithm with examples of the inputs and outputs we want the network to compute, and then the error (difference between actual and expected results) is calculated. Since the error is the difference between the actual and the desired output, the error depends on the weights, and we need to adjust the weights in order to minimize the error. The error of the network will simply be the sum of the errors of all the neurons in the output layer. The classification performance of ANN is usually better than HMM and GMM when the number of training examples is relatively low.

4) Support Vector Machine (SVM)

For the pattern recognition case, SVMs [5] have been used for isolated handwritten digit recognition, object recognition, speaker identification, face detection in images text categorization etc. For the regression estimation case, SVMs have been compared on benchmark time series prediction tests. In most of these cases, SVM generalization performance (i.e. error rates on test sets) either matches or is significantly better than that of competing methods. The use of SVMs for density estimation and decomposition has also been studied. Regarding extensions, the basic SVMs contain no prior knowledge of the problem (for example, a large class of SVMs for the image recognition problem would give the same results if the pixels were first permuted randomly (with each image suffering the same permutation) and much work has been done on incorporating prior knowledge. Although SVMs have good generalization performance, they can be abysmally slow in test phase, a problem addressed in. Recent work has generalized the basic shown connections to regularization theory and shown how SVM ideas can be incorporated in a wide range of other algorithms. The problem which drove the initial development of SVMs occurs in several guises – the bias variance trade off capacity control, over fitting - but the basic idea is the same. Roughly speaking, for a given learning task, with a given finite amount of training data, the best generalization performance will be achieved if the right balance is struck between the accuracy attained on that particular training set, and the “capacity” of the machine, that is, the ability of the machine to learn any training set without error. One of the important classifiers is the support vector machine. SVM classifiers are shown to outperform other well-known classifiers.

2.3 Emotion Recognition Systems

Various studies on recognizing emotion from speech features have attempted. In 2006, B. Schuller and G. Rigoll, [6] studied segment-based speech emotion recognition. Additional sub-phrase level information is believed to improve accuracy in speech emotion recognition systems. Yet, automatic segmentation is a challenge on its own considering word boundaries. Furthermore clarification is needed which timing level leads to optimal results. A variety of potential segmentation schemes exists in general. However, automatic segmentation without the necessity of word- or syllable-boundary detection, which is prone to errors and demands for considerable extra-effort as word alignment with an Automatic-Speech-Recognition (ASR) engine. On examine the impact on accuracy of fast and simple "blind" strategies that neglect spoken content but can easily be realized in real-time and partly allow for direct stream processing. The idea is to test whether speech emotion recognition can be realized at a fixed frame rate compared to speaker recognition. Interestingly, a dynamic number of macro-frames were obtained with respect to the overall length if a single phrase is analyzed. This demands for a different classification strategy as HMM or multi-instance learning with a suited fusion scheme as weighted majority vote.

In 2008, L. V. Berens, conducted a study on Interaction Styles [7]. The study found that the interaction styles are of four types. They are:

- Get-Things-Going: (Saguine temperament): happiness: extraversion
- In-Charge: (Choleric temperament): anger: extraversion
- Behind-The-Scenes: (Phlegmatic temperament): neutrality: introversion
- Chart-The-Course: (Melancholic temperament): sadness**:** introversion

For Chart-The-Course, people of this style focus on knowing what to do and keeping themselves, the group, or the project on track. For Behind-The-Scenes, people of this style focus on understanding and working with the process to create a positive outcome. For In-Charge, people of this style are focused on results, often taking action quickly. They often have a driving energy with an intention to lead a group to the goal. In Get-Things-Going, They thrive in facilitator or catalyst roles and aim to inspire others to move to action, facilitating the process. Their focus is on interaction, often with an expressive style.

In 2011, Wei-Bin Liang and Chung-Hsien Wu [8] defined a flexible emotional segment to extract the acoustic-prosodic features. The results demonstrated that considering the emotional salient segment helped improve recognition accuracy.

In 2012, S. Ntalampiras and N. Fakotakis developed a model [9] for the temporal evolution of acoustic parameters for speech emotion recognition. They used three methods such as short-term statistics, spectral moments, and autoregressive models to integrate subsequent feature values.

In 2013, C.-H. Wu, J.-C. Lin and W.-L. Wei, proposed a Two-level semi-coupled HMM [10] to solve the problem of complex temporal course of emotional expression in a conversation. It enables high-performance of emotion recognition during conversation. The results indicated that, compared with other approaches, it helps to recognize emotion with greater accuracy of 87.5%.

In 2013, Igor Bisio, Alessandro Delfino, Fabio Lavagetto, Mario Marchese and Andrea Sciarrone [11] propose a system that allows recognizing a person's emotional state starting from audio signal registrations. The provided solution is aimed at improving the interaction among humans and computers, thus allowing effective human-computer intelligent interaction. The system is able to recognize six emotions (anger, boredom, disgust, fear, happiness, and sadness) and the neutral state. This set of emotional states is widely used for emotion recognition purposes. It also distinguishes a single emotion versus all the other possible ones, as proven in the proposed numerical results. The system is composed of two subsystems: 1) gender recognition (GR) and 2) emotion recognition (ER). The experimental analysis shows the performance in terms of accuracy of the proposed ER system. The results highlight that the a priori knowledge of the speaker's gender allows a performance increase. The obtained results show also that the features selection adoption assures a satisfying recognition rate and allows reducing the employed features. The performance analysis shows the accuracy obtained with the adopted emotion recognition system in terms of recognition

rate and the percentage of correctly recognized emotional contents. The experimental results highlight that the Gender Recognition (GR) subsystem allows increasing the overall emotion recognition accuracy.

In 2014, Konstantin Mark and Tomoko Matsui [12], proposes an emotion recognition system using Gaussian Processes. Gaussian Processes (GPs) are Bayesian nonparametric models that are becoming more and more popular for their superior capabilities to capture highly nonlinear data relationships in various tasks, such as dimensionality reduction, time series analysis, novelty detection, as well as classical regression. These are two main tasks in the music information retrieval (MIR) field. So far, the support vector machine (SVM) has been the dominant model used in MIR systems. Like SVM, GP models are based on kernel functions and Gram matrices; but, in contrast, they produce truly probabilistic outputs with an explicit degree of prediction uncertainty. In addition, there exist algorithms for GP hyper parameter learning. The system has two sub systems, one for music genre classification and another for music emotion estimation using both SVM and GP models, and compared their performances on two databases of similar size. In all cases, the music audio signal was processed in the same way, and the effects of different feature extraction methods and their various combinations were also investigated. The evaluation experiments clearly showed that in both music genre classification and music emotion estimation tasks the GP performed consistently better than the SVM.

3. Proposed Methodology

Affective computing is a field of research that deals with recognizing, interpreting and processing emotions or other affective phenomena. It plays an increasingly important role in assistive technologies. With the help of affective computing, computers are no longer indifferent logical machines. They may be capable of understanding a user's feelings, needs, and wants and giving feedback in a manner that is much easier for users to accept. Emotion recognition is an essential component in affective computing. In daily communication, identifying emotion in speech is a key to deciphering the underlying intention of the speaker. Computers with the ability to recognize different emotional states could help people who have difficulties in understanding and identifying emotions.

The proposed methodology plans to develop a system for recognizing emotions. An emotion recognition system consists of two phases. They are the training and the testing phase. The basic steps of a recognition system in the training phase are data acquisition, pre-processing, feature extraction, training and classification. The testing phase consists of data acquisition, pre-processing, feature extraction, comparison and classification, recognition. The system considers six emotions, namely- anger, boredom, fear, joy, neutral and sadness. The database used by the system is the Polish Emotional Speech Corpus. It consists of speech samples of 8 speakers for each emotion. All speakers were graduate students of Polish national Film, Television and Theatres School, in Lodz, Poland. Each emotion has 40 samples each. So, total 240 speech samples are used by the system. Among them 30 samples from each emotion is used to train the system and the remaining samples are used for testing.

3.1 Emotion Recognition Systems

An emotion recognition system consists of various stages as shown in the Figure 3.1. The first step is to collect the necessary emotional speech samples to produce the speech database. It is the input to the emotional recognition system. This speech signal is to be processed inorder to identify the corresponding emotion in it. For this, the signal is to be pre-processed first. It will improve the quality of the input signal. From this signal the features are to be collected. The system has to use any classifier inorder to classify the speech signals according to the emotions. By using these features extracted from the signals, the classifier is to be trained. After completing the training, the system becomes familiar with the various emotions and their features. So now, when another speech signal is given, it will properly classify it.

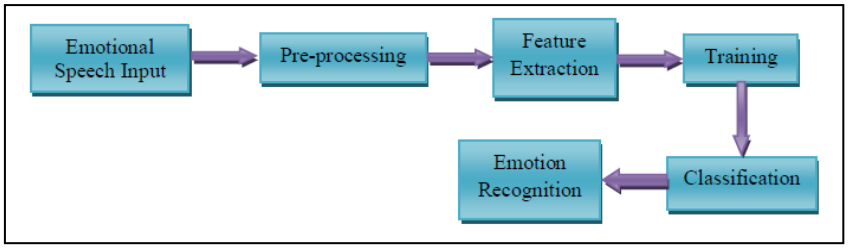

Figure 3.1: Different Stages of Emotion Recognition System

3.1.1 Pre-Processing

Pre-processing of speech signals is considered a crucial step in the development of a robust and efficient speech or speaker recognition system. Pre-processing means segregating the voiced region from the silence/unvoiced portion of the captured signal is usually advocated as a crucial step in the development of a reliable speech or speaker recognition system. This is because most of the speech or speaker specific attributes are present in the voiced part of the speech signals moreover, extraction of the voiced part of the speech signal by marking and/or removing the silence and unvoiced region leads to substantial reduction in computational complexity. Other applications of classifying speech signals into silence /unvoiced region and voiced region are: Fundamental Frequency Estimation, Formant Extraction or Syllable Marking, End Point Detection etc.

3.1.2 Feature Extraction

A speech signal composed of large number of parameters which indicates emotion contents of it. Changes in these parameters indicate changes in the emotions. Proper choice of feature vectors is one of the most important tasks in speech recognition. The feature vectors can be classified as long-time and short-time feature vectors. The long-time ones are estimated over the entire length of the utterance. The short-time ones are determined over window of usually less than 100 ms. The long-time approach identifies emotions more efficiently. Commonly used features [13] are mel-frequency cepstrum coefficients (MFCC).

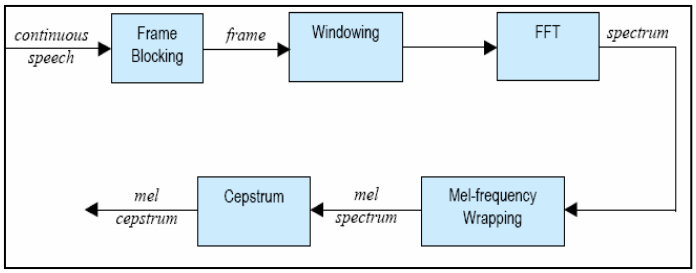

Figure 3.2: MFCC Calculation

MFCC is based on the characteristics of the human ear's hearing. It uses a nonlinear frequency unit to simulate the human auditory system. Mel frequency scale is the most widely used feature of the speech, with a simple calculation, good ability of the distinction, anti-noise and other advantages.

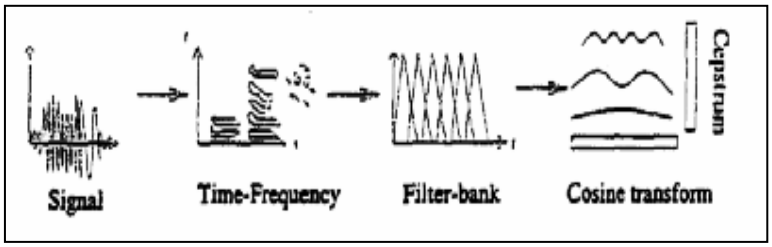

Figure 3.3: MFCC of a Signal

3.1.2.1 MFCC Feature Extraction

1. Pre-emphasize input signal.
   - It includes 2 processes: - framing and windowing.
   - Framing
     - It divides the input signal into frames each of size N with M overlapping values.
     - If length (last frame < N), do padding.
   - Windowing
     - It is used to minimize the signal discontinuities at the beginning and end of each frame.
     - It applies the Hamming Window Function for each frame.

2. Perform discrete Fourier analysis to get power spectrum.
3. Wrap the power spectrum into Mel-spectrum using triangular Mel-Filter. The equation for converting the frequencies in Hz into Mel scale is given below.

$$fMel = 2595 \log_{10}(1 + \frac{fHz}{700}) \dots\dots (3.1)$$

where:

$fMel$ is the frequency in Mel scale
fHz is the frequency in Hz

4. Take the log operation on the Mel-power spectrum.
5. Apply the discrete cosine transform (DCT) on the log-power- spectrum to derive mel-frequency cepstral features.

Similar to Mel Frequency Cepstrum Coefficient (MFCC), there is another feature vector called Gammatone Cepstrum Coefficient (GTCC) [14] or Gammatone Frequency Cepstrum Coefficient (GFCC). MFCC systems usually do not perform well under noisy conditions because extracted features are distorted by noise, causing mismatched likelihood calculation. By introducing a novel speaker feature, gammatone cepstral coefficient (GTCC), based on an auditory periphery model, and show that this feature captures speaker characteristics and performs substantially better than conventional speaker features under noisy conditions. An important finding in the study is that GTCC features outperform conventional MFCC features under noisy conditions.

Broadly speaking, there are two major differences between MFCC and GTCC. The obvious one is the frequency scale. GTCC, based on equivalent rectangular bandwidth (ERB) scale, has finer resolution at low frequencies than MFCC (mel scale). The other one is the nonlinear rectification step prior to the DCT. MFCC uses a log while GTCC uses a cubic root. Both have been used in the literature. In addition, the log operation transforms convolution between excitation source and vocal tract (filter) into addition in the spectral domain. By carefully examining all the differences between MFCC and GTCC, it concludes that the nonlinear rectification mainly accounts for the noise robustness differences. In particular, the cubic root rectification provides more robustness to the features than the log. The cubic root operation better might be the case that some speaker information is embodied through different energy levels. In a noisy mixture, there are target dominant T-F units or segments indicative of this energy information. The cubic root operation makes features scale variant (i.e. energy level dependent) and helps to preserve this information. The log operation, on the other hand, does not encode this information. Thus, GTCC provides more accurate result than MFCC. This study compares the differences between MFCC and GTCC for recognizing emotion from speech.

3.1.2.2 GTCC Feature Extraction

1) Gammatone Filter Properties

The Gammatone function can be used for the modeling of the human auditory filter responses. The correlation between the impulse response of the Gammatone filter and the one obtained from the humans observed that the properties of the frequency selectivity of the cochlea and those psychophysically measured in human beings seem to converge, since: 1) the magnitude response of a fourth-order GT filter is very similar to *reox* function (commonly used to represent the human auditory filter response) and 2) the filter bandwidth corresponds to a fixed distance on the basilar membrane.

The Gammatone filter takes its name from the impulse response, which is the product of a Gamma distribution function and a sinusoidal tone centered at the frequency, being computed as:

$$g(t) = Kt^{(n-1)}\, e^{-2\pi Bt} \cos(2\pi f_c t + \emptyset) \qquad t > 0 \dots\dots (3.2)$$

where:

$g(t)$ is the Impulse Response of Gammatone Filter
K is the amplitude factor
n is the filter order
f_c is the central frequency in Hertz

$\emptyset$ is the phase shift
B is the duration of the impulse response

2) *Equivalent Rectangular Bandwidth*

The equivalent rectangular bandwidth is a psychoacoustic measure of the auditory filter width at each point along the cochlea. An ERB filter models the spectral integration derived from the channelling effectuated by the inner hair cells, which send signals of a certain bandwidth to the brain.

$$EBR = [\left(\frac{f_c}{EarQ}\right)^p + minBW^p]^{1/p} \dots\dots (3.3)$$

where:

EarQ is the asymptotic filter quality at high frequencies
minBW is the minimum bandwidth at low frequencies
p is commonly 1 or 2

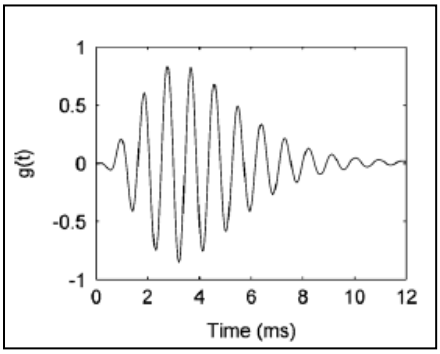

Figure 3.4: Impulse Response of a Gammatone Filter

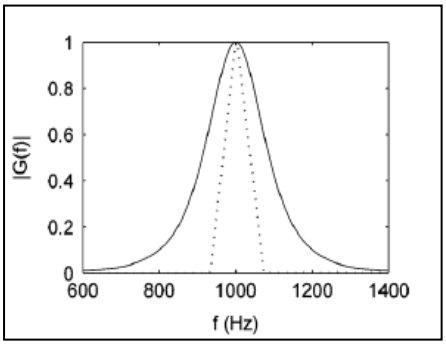

Figure 3.5: Frequency response of a Gammatone Filter (dark line) and a Mel Filter (dotted line)

3) *GT-ERB Filter Bank*

The distribution of the filters along the bandwidth of interest is particularly relevant when it comes to implementing an audio filter bank. In the human auditory system there are about 3000 inner hair cells along the cochlea, each one resonating at its characteristic frequency within a certain bandwidth. Typically, this filter density is computationally approximated by a lower number of band-pass filters with certain spectral overlap. Considering the biologically motivated ERB bands, the central frequency of each GT filter is given by:

$$f_{ci} = \left(f_{hight} + EarQminBW\right)e^{-i\,step/EarQ} - EarQminBW \ldots\ldots (3.4)$$

where:

f_{ci} is central frequency of GT filter
f_{hight} is the highest frequency considered by the filter bank
$EarQ \ and \ minBW$ are the ERB parameters
i is the GT filter index
step is the gap between consecutive filters, which can be calculated as :

$$step = \frac{EarQ}{N}\log\left(\frac{f_{hight} + EarQminBW}{f_{low} + EarQminBW}\right) \ldots\ldots (3.5)$$

where:

f_{low} is the lowest frequency considered
N is the number of GT filters

4) Gammatone Cepstral Coefficients

The computation process of the proposed Gammatone cepstral coefficients is analogous to the MFCC extraction scheme.

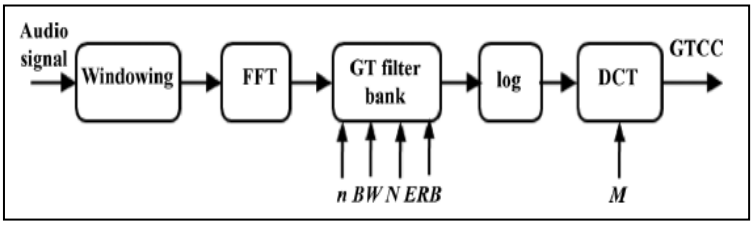

Figure 3.6: GTCC Feature Extraction

The audio signal is first windowed into short frames; usually of 10–50 ms. This process has a twofold purpose.
1) the non-stationary audio signal can be assumed to be stationary for such a short interval a
2) the efficiency of the feature extraction process is increased.

Subsequently, the GT filter bank (composed of the frequency responses of the several GT filters) is applied to the signal's fast Fourier transform (FFT), emphasizing the perceptually meaningful sound signal frequencies. Finally, the log function and the discrete cosine transform (DCT) are applied to model the human loudness. The overall computation cost is almost equal to the MFCC computation as:

$$GTCC_m = \sqrt{\frac{2}{N}}\sum_{n=1}^{N}\log(X_n)\cos\left[\frac{\pi n}{N}\left(m - \frac{1}{2}\right)\right] \quad 1 \le m \le M \ldots\ldots (3.6)$$

where:

X_n is the energy of the signal in the n th spectral band
N is the number of Gammatone filters
M is the number of GTCC

3.1.3 Training and Classification

Selection of classifier depends on the geometry of the input feature vector. Some classifiers are more efficient with certain type of class distributions. Various classifiers used are Hidden Markov Model (HMM), Gaussian Mixtures Model (GMM), Support Vector Machine (SVM), Artificial Neural Network (ANN) etc. The GMM and the HMM, are the most used ones for speech emotion recognition. Another common classifier, used for many pattern recognition applications is the artificial neural network (ANN). Their classification performance is usually better than HMM and GMM when the number of training examples is relatively low.

Hence, in the proposed system the classifier used is an artificial neural network. An artificial neuron is a computational model inspired in the natural neurons. Natural neurons receive signals through synapses located on the dendrites or membrane of the neuron. When the signals received are strong enough (surpass a certain threshold), the neuron is activated and emits a signal though the axon. This signal might be sent to another synapse, and might activate other neurons. The complexity of real neurons is highly abstracted when modelling artificial neurons. These basically consist of inputs (like synapses), which are multiplied by weights (strength of the respective signals), and then computed by a mathematical function which determines the activation of the neuron. Another function (which may be the identity) computes the output of the artificial neuron (sometimes in dependence of a certain threshold).

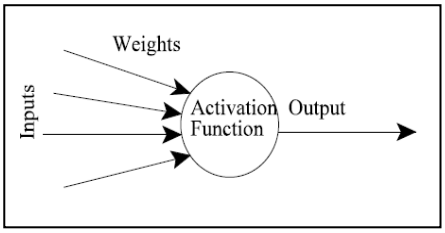

Figure 3.7: Artificial Neuron

ANNs combine artificial neurons in order to process information. The higher a weight of an artificial neuron is, the stronger the input which is multiplied by it will be. Weights can also be negative, so we can say that the signal is inhibited by the negative weight. Depending on the weights, the computation of the neuron will be different. By adjusting the weights of an artificial neuron we can obtain the output we want for specific inputs. But when we have an ANN of hundreds or thousands of neurons, it would be quite complicated to find by hand all the necessary weights. But we can find algorithms which can adjust the weights of the ANN in order to obtain the desired output from the network. This process of adjusting the weights is called learning or training. The proposed system employs a cascade feed forward neural network [15] for its training and classification. Feed forward back propagation artificial neural network model consists of input, hidden and output layers. Back propagation learning algorithm is used for learning the network. During training this network, calculations were carried out from input layer of network toward output layer, and error values were then propagated to prior layers. Feed forward networks often have one or more hidden layers of sigmoid neurons followed by an output layer of linear neurons. Cascade forward back propagation mode is similar to feed-forward networks, but include a weight connection from the input to each layer and from each layer to the successive layers. While two-layer feed forward networks can potentially learn virtually any input output relationship, feed-forward networks with more layers might learn complex relationships more quickly. Cascade forward back propagation ANN model is similar to feed forward back propagation neural network in using the back propagation algorithm for weights updating, but the main symptom of this network is that each layer of neurons related to all previous layer of neurons.

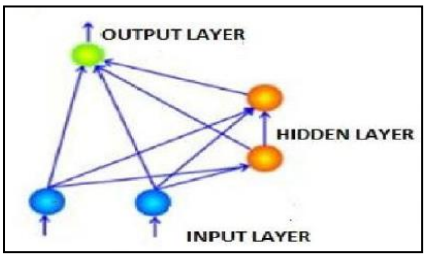

Figure 3.8: Cascade Feed Forward Neural Network

Tan-sigmoid transfer function was used to reach the optimized status and is given by:

$$tansig(x) = \frac{2}{1 + \exp(-2x)} - 1 \ldots\ldots (3.7)$$

where :

x is the input to tansig function

The plot for tan-sigmoid transfer function is shown below.

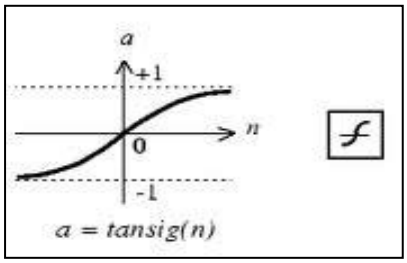

Figure 3.9: Tan-sigmoid Transfer Function Plot

The performance of cascade forward back propagation is evaluated using Mean Square Error (MSE).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(observed - predicted)^2 \quad \ldots\ldots (3.8)$$

where:

observed is the current output
predicted is the actual output

4. Results

The emotion recognition system developed here considers six emotions – anger, boredom, fear, joy, neutral and sadness. The system uses both MFCC and GTCC features for recognizing various emotions. First, the system is trained with MFCC features and result is obtained. Then, the system is trained with GTCC features and result is obtained. For training and classification a cascade forward back propagation neural network is used. Finally, the performance and recognition accuracy of the system while using each feature are compared. The results obtained from the system are given in the following figures and tables.

It is necessary to enter the number of emotions and also the number of speech samples from each emotion to be considered for training. It considers six emotions and each emotion have forty speech samples (total 240 samples). The system uses 30 samples from each emotion for training (total 180 samples) and the remaining samples (10*6=60) are used for testing the system.

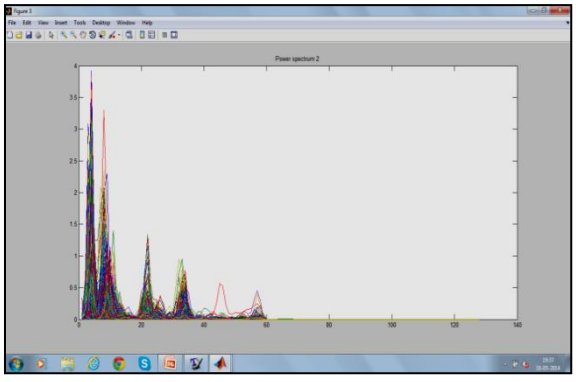

Figure 4.1: Mel-Power Spectrum

The figure 4.1 shows the Mel-Power Spectrum corresponds to a speech signal. It is obtained by wrapping the power spectrum of the speech signal with the Mel-Filter.

The figure 4.2 shows the Mel Frequency Cepstrum Coefficient corresponds to the speech signal. It is obtained by taking the discrete cosine transformation the Mel-Power Spectrum. Then, training is done by using the MFCC features.

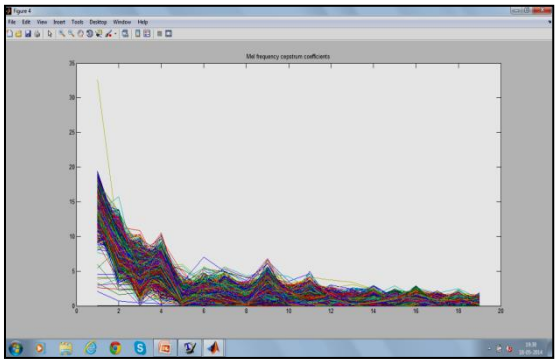

Figure 4.2: Mel Frequency Cepstrum Coefficient Plot

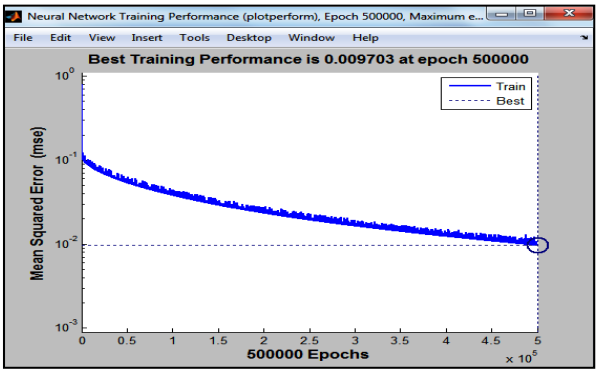

Figure 4.3: Performance Plot of the NN after Training

The figure 4.3 shows the performance plot of the neural network. Here, the MSE of the system decreases as the training proceeds. The MSE corresponds to the total number of iterations (5,00,000) is represented using the dotted line.

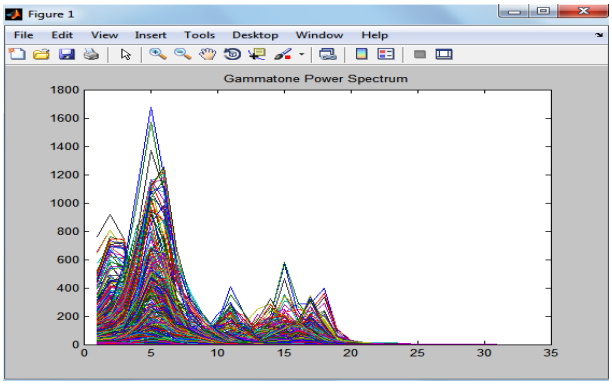

Figure 4.4: Gammatone Power Spectrum

The figure 4.4 shows the Gammatone-Power Spectrum corresponds to a speech signal. It is obtained by wrapping the power spectrum of the speech signal with the Gammatone-Filter.

The figure 4.5 shows the Gammatone Cepstrum Coefficient corresponds to the speech signal. It is obtained by taking the discrete cosine transformation the Gammatone-Power Spectrum. Then, training is done by using the GTCC features.

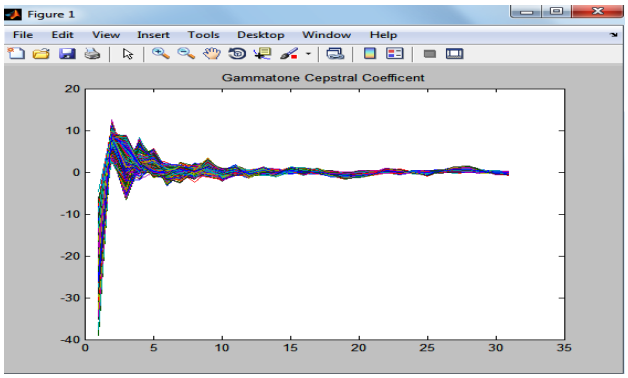

Figure 4.5: Gammatone Cepstrum Coefficient Plot

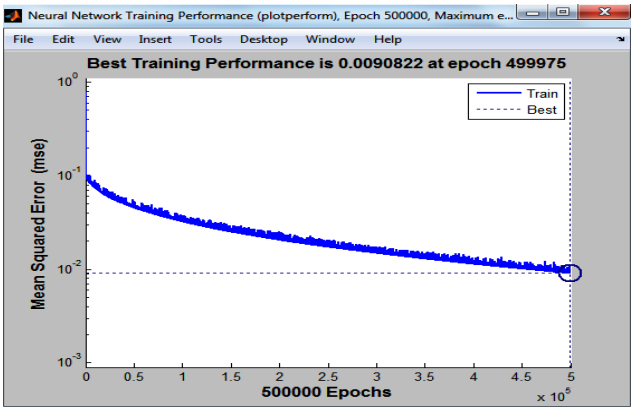

Figure 4.6: Performance Plot of the NN after Training

The figure 4.6 shows the performance plot of the neural network. Here, the MSE of the system decreases as the training proceeds. The MSE corresponds to the total number of iterations (5,00,000) is represented using the dotted line.

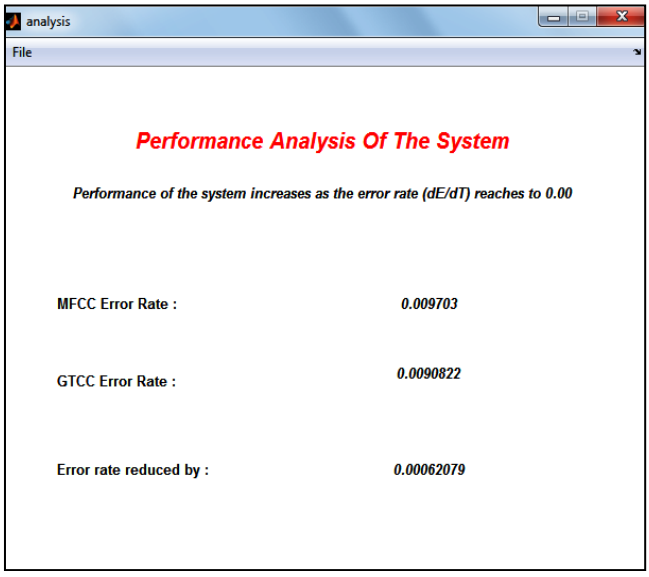

Figure 4.7: Performance Analysis of the System using MFCC and GTCC Features

The figure 4.7 shows the window that gives the performance analysis of the system. The performance of the system increases as the error rate reaches to 0.00. Here, the error rate of the system reduced by 0.0062079 when it uses GTCC features.

The accuracy of the system can be further improved by continuing the training by varying the number of iterations, number of layers, number of nodes in each layer etc. The error rate obtained for the system during the trainings done prior to reach the current state is given by the table 4.1. Among the results the best one (ie. minimum error) is selected for the system.

Table 4.1: Error Rate of the System during Various Trainings

Number of iterations	Number of layers	Number of nodes corresponds to each layer	Error rate	
			While using MFCC	While using GTCC
30,000	4	10, 100, 80,6	0.048006	0.040334
50,000	4	10, 100, 160, 6	0.039847	0.037756
1,00,000	5	10, 100, 180, 160, 6	0.026438	0.024311
1,50,000	6	10, 100, 140, 180, 160, 6	0.024120	0.021243
2,00,000	6	10, 100, 140, 180, 160, 6	0.022336	0.021691
3,00,000	6	10, 100, 140, 180, 160, 6	0.019328	0.01661
5,00,000	6	10, 150, 180, 200, 160, 6	0.009703	0.0090822

5. Conclusion

Emotion recognition from speech signal has become a major research topic in the field of human computer interaction in the recent times due to its many potential applications. It is being applied to growing number of areas such as humanoid robots, car industry, call centres, mobile communication, computer tutorial applications etc. The most commonly used acoustic feature is the MFCC. Such systems usually do not perform well under noisy conditions because extracted features are distorted by noise, causing mismatched likelihood calculation. By introducing a novel speaker feature, gammatone cepstral coefficient (GTCC), based on an auditory periphery model, the system performs substantially better than the conventional speaker feature. The emotion recognition system implemented here compares the system performance while using the MFCC and GTCC features. The error rate of the system corresponds to MFCC and GTCC is 0.009703 and 0.0090822 respectively.

REFERENCES

[1] Dipti D. Joshi, Prof. M. B. Zalte, "Speech Emotion Recognition: A Review", *IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)(Jan. - Feb. 2013).*

[2] Phil Blunsom, pcbl@cs.mu.oz.au, August 19, 2004, "Hidden Markov Models".

[3] Douglas Reynolds, "Gaussian Mixture Models", MIT *Lincoln Laboratory, 244 Wood St., Lexington, MA 02140, USA.*

[4] http://www.cse.unr.edu/~bebis/MathMethods/NNs/lecture.pdf

[5] CHRISTOPHER J.C. BURGES burges@lucent.com *Bell Laboratories, Lucent Technologies,* Kluwer Academic "A Tutorial on Support Vector Machines for Pattern Recognition", *Boston. Manufactured in The Netherlands.*

[6] B. Schuller and G. Rigoll, "Timing levels in segment-based speech emotion recognition," in *Proc. INTERSPEECH,* 2006, pp. 1818–1821.

[7] L. V. Berens, *Understanding Yourself and Others: An Introduction to Interaction Styles*. Brighton, U.K.: Telos, 2008.

[8] W.-B. Liang, C.-H. Wu, C.-H. Wang, and J.-F. Wang, "Interactional style detection for versatile dialogue response using prosodic and semantic features," in *Proc. INTERSPEECH*, 2011.

[9] S. Ntalampiras and N. Fakotakis, "Modeling the temporal evolution of acoustic parameters for speech emotion recognition," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 116–125, Jan.-Mar. 2012.

[10] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Two-level hierarchical alignment for semi-coupled HMM-based audiovisual emotion recognition with temporal course," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1880–1895, Dec. 2013.

[11] IGOR BISIO, ALESSANDRO DELFINO, FABIO LAVAGETTO, MARIO MARCHESE, and ANDREA SCIARRONE, "Gender-Driven Emotion Recognition Through Speech Signals for Ambient Intelligence Applications", IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING, 2013.

[12] KONSTANTIN MARKOV and TOMOKO MATSUI, "Music Genre and Emotion Recognition Using Gaussian Processes ", IEEE ACCESS, .2014.

[13] Mel Frequency Cepstral Coefficients for Speaker Recognition Using Gaussian Mixture Model-Artificial Neural Network Model, Cheang Soo Yee and Abdul Manan Ahmad, *Faculty of Computer Science and Information System, University of Technology Malaysia.*

[14] Gammatone Cepstral Coefficients: Biologically Inspired Features for Non-Speech Audio Classification Xavier Valero, *Student Member, IEEE,* and Francesc Alías, *Member, IEEE, 2014.*

[15] Cascade and Feed Forward Back propagation Artificial Neural Network Models for Prediction of Compressive Strength of Ready Mix Concrete, *IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE).*

Minu Babu – Currently doing her final year post graduation in Computer Science from Mahatma Gandhi University, Kottayam, Kerala. She took graduation from the same university in 2012 with distinction. This work is done as a part of her PG mini project. She is very much interested in Pattern Recognition, Image Processing, Machine Learning and Artificial Intelligence areas. Recently, she is interested in Emotion Recognition from speech and doing her main project in this area.