Identification Of Individual Melodies Using Artificial Neural Network Classifier

Dissertation submitted in the partial fulfillment of requirements for the award of degree of

Master of Engineering

In

Wireless communications

Submitted by:

Priya

Roll No: 801363022

Under the guidance of:

Dr. Ravi Kumar

Assistant Professor, ECED

Electronics And Communication Engineering Department

THAPAR UNIVERSITY

(Established under the section 3 of UGC Act, 1956)

PATIALA – 147004 (PUNJAB)

July, 2015

DECLARATION

I, hereby declare that the work, which is being presented in this thesis entitled **"Identification Of Individual Melodies Using Artificial Neural Network Classifier"** in partial fulfillment of requirements for the award of the degree of **Master of Engineering** in Wireless Communications from Electronics and Communication Department, Thapar University, Patiala is an authentic record of my own work carried out under the supervision of Dr. Ravi Kumar (Assistant Professor, ECED and refers other research's work which are duly listed in reference section. The matter embodied in this work has not been submitted anywhere else for the award of any other degree.

Priya

Date: 11 July, 15

Place: Patiala

Priya

Roll No. 801363022

It is certified that the above statement made by the student is correct to the best of my knowledge and belief.

Ravi

Date: 11 July, 15

Place: Patiala

Supervisor:

Dr. Ravi Kumar

Assistant Professor

ECED, Thapar University

Dr. Sanjay Sharma

Professor and Head (ECED)

Thapar University, Patiala

Dr. S. S. Bhatia

Dean, Academic Affairs

Thapar University, Patiala

Acknowledgement

With profound sense of gratitude, I take it as a highly esteemed privilege in expressing my sincere thanks to my supervisor and guide, **Dr. Ravi Kumar (Assistant Professor)**, Electronics and Communication Engineering Department, Thapar University, Patiala, for his technical guidance, sound advice, excellent supervision and valuable suggestion to accomplish this dissertation. The abundance enthusiasm with which he solved my difficulties whenever required will be remembered forever with plentiful greatness. Without his wise counsel and able guidance, this dissertation would not have been possible.

I express my heartfelt sense of gratitude and thanks to to **Dr. Sanjay Sharma, Head of Department (ECED), Dr. Amit Kumar Kohli, P.G. Coordinator (ECED)**, whose timely guidance and constant motivation helped me to complete my dissertation. I am also thankful to the staff of ECE department for providing me helpful cooperation during my work.

This acknowledgement would be incomplete if I do not express my deep sense of gratitude towards my family as no words describe their incessant encouragement and constant motivation throughout my dissertation. I am thankful to my parents whose love, moral support and encouragement have been a great source of inspiration to me. I also thank my friends and well-wishers whose well wishes helped me during the dissertation.

Above all, I bow to the almighty for blessing me with wisdom, ability and strength to carry out this work with sincerity and dedication.

Priya
Roll No. 801363022

ABSTRACT

New trends in music distribution and storage have led to tremendous interest in the processing of music signals. From surfing music collections, to discovery new performers, and preventing the records from privacy computers have played a very vital role, thus processing of music is need of the hour.

Every part of music can be classified on the basis of genre, artist or one of the many other parameters. The music samples have wide variety of information and details which are very important for different kind of applications. There are various properties of audio signals which are defined such as fundamental periodicity, known as pitch, and amount of overlapping of music instruments in a sample, known as polyphony, and types of characteristics, known as timbre. Some of the properties are just defined in lay-man language but do not have exact mathematical definition, timbre is not defined mathematically. In this report feature extracted and used for processing is chroma feature.

Artificial neural networks(ANN) have been applied to many research fields like speech recognition, classification of cancers and gene prediction. In this dissertation ANN is used on music melody samples. ANN is made to learn the data and its formation and then it is tested. Two techniques of ANN, namely Back Propagation Algorithm and Radial Basis Function, are used for this purpose.

The chromagrammed data is used as as learning and testing data. Principal Component Analysis (PCA) and Continuous Wavelet Transform (CWT) are both used typically used for large set of data. These two techniques are used as pre-processing techniques for the data set.

A comprehensive analysis was carried out in between the learning error rate performance and testing performance of all the pre-processed and non-processed data sets.

The comparative results demonstrate the effectiveness of the proposed method.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1- INTRODUCTION

Music is a very important part of billions of people's lives all around the globe. Since ages music has been existing in the society. Creating and performing music perfectly have been one of the most intricate art form since long. Music itself is invisible but has visible effect on emotions of the listener. Few people flow with the music only. Music can be classified into enormous range of styles and forms, just like simple, folk songs of different regions, orchestras, and creation of electronic music is very difficult, it takes hard work of months in studio.

1.1 Music Information Retrieval

This section gives brief overview of the different applications, the problems and the methods related to pattern recognition for the analysis of music. Many of the uses of musical pattern recognition are connected to MIR (Music Information Retrieval). In MIR various disciples of study are involved while processing. This field bridges the domains of pattern recognition, machine learning, digital audio signal processing and software design system together. in other words it can be said that, by using MIR algorithms a computer listens and makes sense of any kind of audio data as per the algorithm. The data normally used is either MP3 or WAVE format. mostly .wav files are used over .mp3. the data set can be any of the personal collections, gigabytes of sound effects, or live streaming audio. The expected result of MIR is to try to reduce gap between higher-lever music information and lower level aural data. The processing of MIR should be such that, it should be such that the computer should be able to detect the characteristics of music and sound such as key, chord, tempo, genre, chord progression etc. The main reason why do we need MIR is that it enables us to recognize and thus extract information from samples, it enables system to perform rigorous searching and sorting and recommending music. This area covers fields like signal processing, information retrieval, pattern recognition, artificial intelligence, computer music processing, databases and music cognition. This chapter of report focuses on methods and problems related to pattern recognition and signal processing. Content-based music retrieval and Automatic music transcription are some of the problems those have received the maximum attention within

this area. Audio queries based on content-based retrieval are somewhat dependent on the transcription, though to find similarity a full transcription may not be necessary.

Other problems like music summarization genre classification and musical instrument recognition are also studied. There are various other problems those are related to music retrieval.

MIR (Music Information Retrieval) has been defined by Stephen Downie as 'a multidisciplinary research endeavor that strives to develop innovative content-based searching schemes and novel interfaces'.[28] Many researchers have personal interest in music and therefore they use their personal audio collection.

The basic objective for the research in this field is to design a classifier which would classify the music into different types genres Indian Raagas like Asa raag, Bhairav Raag, Gauri Raag, Malar Raag etc on the basis of Tempo, Beat, and Rhythm.

1.2 Motivation

Most of the musicians face a problem that they can't work to their best with just reading through the sheets of music, for it they have to attain a specific level of music achievement. For example, only few musicians can guess properly which note is being played by just listening to the node, unless they read the sheet music like as in case of C-sharp(#) and C-natural. They can only identify, as if their ears are tuned to each of the note fairly well. But budding musicians don't have the skill. The point is, it is very difficult to identify the notes being played unless sheet is read. A beginner or not that expert musician can also identify the music notes when given to him in the form of sheets not in video or audio. Thus it is useful to give them the processed form of music which they can identify.

Even after so much research on music such program is not made yet which can convert mp3 in the form of sheets. To create sheet, one has to play score on keyboard, guitar or one can enter at their own in the form of sheets, but can't play on mp3. MIDI files are built in such a way that they already contain such details (such as duration of data or pitch) which are needed by the software to interpret the notes in the file, whereas mp3 are sound waves which contain different set of information to be used in sound editing software. This has remained as a area where musicians can still make more research.

For working on Music Information Retrieval, deep understanding of music theory is not required, but the basic of music structure are necessary to be understood. Surprisingly, knowledge regarding audio perception has a very limited role is most of the music signal processing systems, but since music exists only to be heard, hearing sciences promise to help to advance our capability to understand music perception and should therefore inform the analysis of the complex signals.

1.3 Speech v/s Music Signals

A) Speech v/s music spectra

Speech is generated by a uniform set of cavities and tubes. The human vocal tract is about 17 cm from vocal chords to lips. The human vocal tract is operated by fundamental rules of acoustics which are not dependent on the language spoken.

For example, the frequency of the resonance of the vocal tract, formants, is governed basically by constriction in the mouth and length of the vocal tract tube. Length of the vocal tract cannot change significantly. It is to understand that an adult vocal tract can generate a limited set of outputs. If the sample is put together for over a period of time , it can be referred as “long term speech spectrum”.

In comparison, to the well-defined human vocal tract output, there is no well-defined, long term music spectrum. The output of various musical instruments is very highly variable ranging from a low-frequency to a high-frequency emphasis. In various cases the output spectrum may resemble the speech spectrum, sometimes in others, there is no resemblance.

B) Physical output against perceptual requirements of the listener

In speech, there are very slight differences between various languages in proportion of audible cues which are important for speech perception, has been summarized under the "articulation index (AI)" research. Measures like AI are used for decades in the hearing aid industry. Results for AI importance weighted as a frequency function may vary slightly from language to language, but normally it is seen that for speech, major important sounds for speech clarity belong to bands above 1000 Hz, whereas most of the loudest speech perception belong to those bands, which are below 1000 Hz.

In speech, the most of the energy, or intense region is the lower frequencies in the spectrum, and the clarity is obtained from the higher frequencies. Speech is more prominent if in the lower frequencies, and is phonemically of more important if in the higher frequencies. The auditory perception of speech has an obvious different weighting than the physical output from the speaker's mouth. Despite the differences between the physical output of speech and frequency requirements for optimal understanding of speech, the differences are the consistent and predictable- low frequency loudness clue and high frequency clarity clue.

Unlike speech, the spectrums of music are highly variable. Ignoring, the physical output of musical instruments, the visceral needs of the musicians and listeners vary depending upon the instruments. String based instrument musicians need to be able to know the exact relationship between fundamental energy of lower frequency and harmonic structure of the higher frequencies. If violinist says, "the instrument sounds great", he is saying that the relationship between the harmonics and the fundamental has desired balance - both in exact spectral location and relative intensity. One can also say that the violinist has a broadband phonemic requirement. Along with generating a wide range of frequencies, violinist needs to be capable to hear and identify those frequencies.

In contrary, a woodwind player such as a clarinetist is to be capable to recognize the lower frequency inter-resonant breathiness. If a clarinet player says "sound is good", he is saying that there is lower frequency noise in amidst resonances of their instrument has a specific level. High frequency information is not very crucial to a clarinet player (other than for loudness perception). Therefore, one can say that, a clarinet player has a low frequency phonemic requirement, against the fact that clarinet player can also produce as many higher frequency sounds as the violinist can.

C) Loudness summation, loudness, and intensity

The "source" of sound in human vocal tract is the vibration in the vocal cords. If we relate it in physics, due to the way the vocal cords are held close to the larynx, their function is as one half wavelength resonator. This doesn't mean that there is only the fundamental energy (usually 120-130 Hz for men and 180-220 Hz for women) but

also there are harmonics at integer multiples of the fundamental evenly spaced. For a man's voice with a fundamental frequency of 130 Hz, there are harmonics at 260 Hz, 390 Hz, 520 Hz, and so on. Rarely there is a fundamental frequency below 100 Hz. Thus the minimum spacing between two harmonics in speech is to be of at least 100 Hz. Also in other words, none of the two harmonics would fall in same critical band, which results that there is minimal loudness summation. It can be also said that soft sounding speeches are lesser intense in comparison to loud sounding speeches. In speech, there is good correlation between one's perception of physical vocal intensity and the loudness

Some of the musical instruments are speech-like, they generate mid- frequency fundamental energies with evenly spaced harmonics. saxophones, Oboes and violins fall under this category. However, some bass stringed instruments such as the cello and the string bass are also half wavelength resonating instruments- like speech- but are perceived as quite loud because more than one harmonics fall within range of one critical bandwidth therefore resulting in increase in loudness (because of loudness summation), with no change the intensity. For the bass and cello, there is very poor correlation in between perceived loudness and measured intensity. A so called "music" channel for cello and bass players, need to be set with less of low-frequency and mid-frequency gain in comparison to other treble-oriented instruments.

D) Different intensities of speech and music

Typical outputs for normal intensity of speech range from 53 dB SPL for [th] as in 'think' to nearly 77 dB SPL for [a] in 'father'. Shouted speech may reach 83 dB SPL. The 24 dB range (+/- 12 dB) can be related to the characteristics of human vocal chords and vocal tract. Whereas, music is in the order of 100 dB SPL with peaks and valleys in the spectrum with +/- 18 dB. For the matter of fact, peaks for a 100 dB SPL musical input can result in distortion of conventional hearing aid microphones (as the maximum transduction capability is 115 dB SPL).

1.4 Musical signal features and their properties.

1.4.1 Audio Features

The basis of any algorithm for auditory signal analysis is short-time feature vector extraction, where the audio files are segmented into small segments in time domain and for each of these segments a feature vector is calculated. Features which can describe the audio signal can typically be divided into two categories- physical and perceptual features. The statistical or mathematical properties of the signals tell about the physical features of segment, while the way humans hear sound gives idea about the perceptual features of the segment. The physical features and the perceptual features usually related to each other[12].

1.4.2 Structure

Popular music tunes are structured around a sequence of chords. A normal band would have a pianist playing the chords (piano or keyboard or guitar), a drummer keeping the beat, a bassist outlining the chords with a walking bass line. When jamming, a lead player usually exercises the chords and melody. Musicians use scales or sets of notes those work with particular chords, and thus build their melodic lines to fit when there are changes in the chords. A set of keys, or tonal centers, for a song are defined with the use of chords. For instance a given key is linked with a scale, and sometimes many successive chords within a song will fit into one single key, making a musician to play with one scale for an extended period of time.

1.4.3 Psychoacoustics[12]

It is the scientific study of sound perception. Basically, it is the branch of science studying the physiological and psychological responses towards sound (including both speech and music). It can be also considered as a branch of psychophysics.

Hearing is not only a purely mechanical phenomenon of wave propagation, but is also a perceptual and sensory event, in other words, when a person feels he has heard something, that something has arrived at his ear as a mechanical (sound) wave traveling through the air, compression of air molecules, but within the ear wave is converted into neural action potentials. These nerve pulses then travel to the brain where they are

recognized. In many problems related to acoustics, such as for audio processing, it is useful to take note that, not just the mechanics of the surrounding environment, but also both the ear and the brain of person are involved in his listening experience.

For example, the ear does significant amount of signal processing in converting sound waveform into neural signal, so certain differences between waveforms may be invisible. Data compression techniques, such as MP3, make use of this fact. In addition, the ear has a nonlinear response to the different intensity levels of sound; this response is called loudness. Audio noise reduction systems and telephone networks make use of the fact by compressing data samples before transmission nonlinearly, and for playback expanding them. Another effect of the ear's nonlinear response is that sounds that are close in frequency produce phantom beat notes, or inter modulation distortion products. A brief introduction of the psychoacoustic parameters follows-

a) **Intensity-**Sound is a wave and waves have amplitude. **Amplitude** is the measure of energy. More energy a wave has, the higher is its amplitude. Increase in amplitude shows the increase in intensity and vice versa. **Intensity** is the amount of energy a signal has over an area. The same sound is heard with more intensity if you hear it in a smaller area. We are used to measure the sounds we hear in term of loudness. The sound of a someone yelling is louder, while the sound of one's own breathing is very soft. A specific number can't be assigned to loudness, but to intensity it can be. Unit to measure intensity is decibel.

   Human ears are more sensitive to higher frequencies than lower, therefore just react much more to higher frequencies than lower even if they have same intensity. Decibels and intensity, however, are not the features of the ear. Instruments are used to measure these. A whisper is about 10 decibels while thunder is 100 decibels. Hearing to loud sounds, sounds with intensities above 85 decibels, prove to harmful for ears. If the sound is above the intensity of 120 dB, is considered to be loud and causes pain when heard as 120 dB is the threshold of pain.

b) **Pitch**- Pitch is the basic feature, which help to understand low or high sounds. For instance, a singer sings a same note twice, one normally and other octave above

the previous one. One (non musician) can tell that there is some a difference between these two sounds. This is because both sounds have different pitch.

Pitch of sound wave depends upon its frequency. **Frequency** is referred to as no of wavelength in a unit time one wavelength is equal to one compression and one rarefaction. Even if the singer sings the same note, because the sounds had different pitches (frequencies), we hear them as different. Unit of frequency is hertz. One hertz is equal to one cycle of compression and rarefaction in one second. High sounds have higher frequencies and low sounds have lower frequencies. Thunders can have frequency of only 50 hertz, while a whistle can have frequency upto 1,000 hertz.

Frequencies ranging from 20 to 20,000 hertz is range of audible frequencies for humans. Some animals have capability to hear sounds at even higher frequencies. Humans cannot hear dog whistles, while other dogs can because the frequency of whistle is too high to be heard by human ears. Sounds which are too high for humans to hear are called **ultrasonic**.

Ultrasonic waves are very useful. For example for bats, bats emit ultrasonic waves and listen to the echoes, this help them know where the walls are or where they can to find prey. Captains of ships, submarines and other boats use special machines that send out and receive ultrasonic waves, which help them to guide their boats through the water and warn them when another boat or danger is near.

c) **Harmonics & Timbre (tone) -** When a sound is produced, it is produced due to vibrations. When a source is made to produce a sound, it actually vibrates with many frequencies at the same time. Each of the frequency produces a wave. Each wave would correspond to a specific sound. And if many waves are produced at a same instant using one instrument the quality of sound depends upon the combination of the sound waves. Another difference noticeable between sounds is that some sounds are pleasant while others are unpleasant. A violin beginner would sound as good as violin player in a symphony, even if the same note is being played by both of them. Sound of violin is different from a flute playing the same pitch. It is because they have a different tone, and they both use different modes to produce sound. Violin uses strings where as flute uses air compression.

If a sound is produced using a guitar. Imagine if its string is tightly stretched. If musician strum it, the energy from his finger gets transferred to the string, causing it to vibrate and produce sound. When whole of the string vibrates, we hear the lowest pitch. This lowest pitch is called the **fundamental tone**. Do not forget, the fundamental is only one of many pitches that a string can produce. Parts of string vibrating at frequencies higher than the fundamental tone are called **overtones**, and those vibrating in whole number multiples of the fundamentals are called **harmonics**. The frequency of two times the fundamental will sound one octave higher than fundamental and is called the **second harmonic**. The **fourth harmonic** is the frequency four times of the fundamental and will sound two octaves higher. Fundamental is one times itself, so it is also called the first harmonic. **Timbre** is a parameter which is not well defined mathematically. Timbre is defined as the "attribute of auditory sensation in terms of which a listener can judge two sounds similarly presented and having the same loudness and pitch as dissimilar"[12].

1.4.4 Spectrogram

In audio-related applications, the most commonly used tool for representing the time-varying energy corresponding to different frequency bands is **"Short-Time Fourier Transform** (**STFT)"**,[12] which, when visualized using its magnitude, gives a graph known as spectrogram (in Figs. 1.1 and 1.2). Formally, let x be a discrete-time signal obtained by uniform sampling of a waveform at a sampling rate of Fs Hz. Using an –N point tapered window w (e.g., Hamming for w(n)=0.54-0.46cos(2πn/N)) and an overlap of half a window length, we obtain the STFT[12]

$$X(t,k) = \sum_{n=0}^{N-1} w(n)x(n + tN/2) \exp(-j2\pi nk/N) \qquad (1.1)$$

t ϵ [0:T-1] and k ϵ [0:K]. Here, T determines the number of frames, K=N/2 is the index of the last unique frequency value, and thus X(t, k) corresponds to the window beginning at time t.=N/2Fs in seconds and frequency

$$fcoef(k) = \left(\frac{k}{N}\right) Fs \qquad (1.2)$$

in Hertz (Hz). Typical values of Fs=44100 ,N=4096 and give a window length of 92.8 ms, a time resolution of 46.4 ms, and frequency resolution of 10.8 Hz.[12]

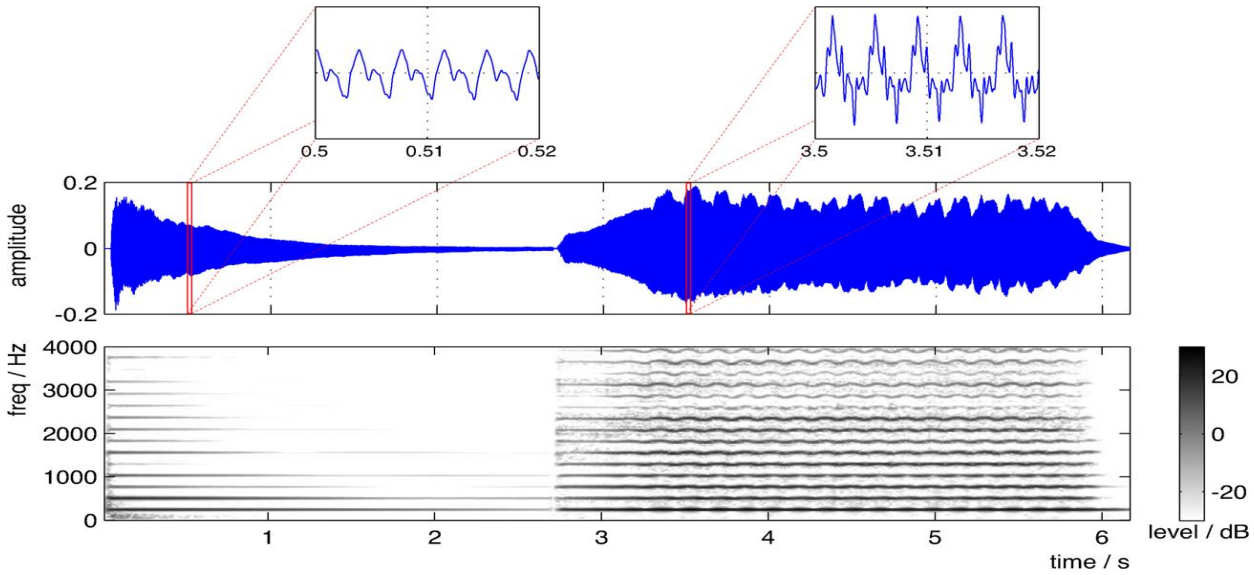

Figure 1.1 Middle C (262 Hz) played on a piano and a violin. The top pane shows the waveform, with the spectrogram below. Zoomed-in regions shown above the waveform reveal the 3.8-ms fundamental period of both notes. [12]

1.4.5 Scales and Chroma

Although different musical conventions have been developed in different cultures, a similar parameter is the musical "scale", which can be described as set of discrete pitches that is repeated after every octave, out of which melodies(tunes) are generated. Consider an example, basic of coexisting western music is "equal tempered" scale, which, by a happy mathematical coincidence, permits the octave scale to be fragmented into 12 similar steps on a logarithmic scale, while sustaining number of intervals corresponding to the popular pleasant note combinations. The equal division makes next frequency greater than its predecessor, this interval is called as a semitone. The coincidence is that there is a possibility to fragment octaves uniformly in smaller steps with close correlation. If there is not an strong correlation, it matches to simple integration ratios that will produce harmonics. Like, The western major scale operated on the octave by

utilizing seven of the twelve steps, having symbolic notation C, D, E, F, G, A, B. by over the octave using seven of the twelve steps —the "white notes" on a piano, denoted by C, D, E, F, G, A, B. Spacing of 2 semitones (like c-sharp(#)), exist in between two simultaneous notes, except for E/F and B/C which have only one semitones distance. Other notes in between white notes are called "black notes" with reference to immediately below or above note, on the basis of musicological conventions. The octave degree given by these symbols is called as the **Chroma** of pitch. Any specific pitch can be identified using integration of chroma and octave number (where spanning of C to B for each octave number). The note which found to be lowest on a piano is A0 (27.5 Hz), the highest note is C8 (4186 Hz), and middle one is C (262 Hz) is C4.[12].

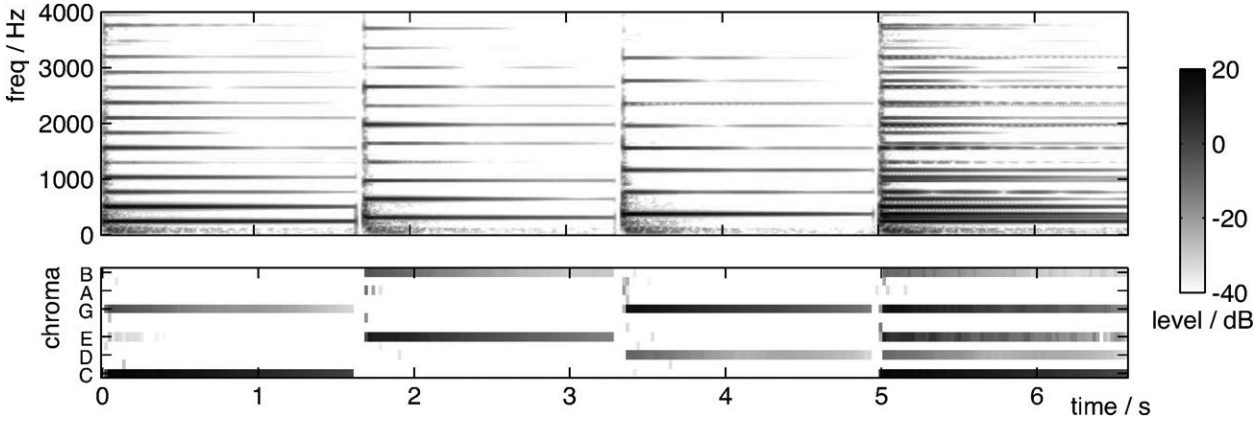

Figure 1.2 Middle C, followed by the E and G above, then all three notes together—a C Major triad—played on a piano. Top pane shows the spectrogram; bottom pane shows the chroma representation [12].

1.4.6 Harmony

Melodies are generated by a sequence of pitches- the "tune" of musical piece While sequences of pitches create melodies—the "tune" of a musical piece, and "voice" only part that can be regenerated by a monophonic instrument—other mandatory parameter of much music is harmony, which is consecutive representation of notes at distinct pitches. Distinct combination of notes gives distinct musical colors or "chords", which can be recognized regardless of the instrument used to play them. [12]

1.4.7 Tempo, Beat, and Rhythm

A primary role is played by musical aspects of tempi, beat and rhythm to understand and interact with music. Beat refered to as a steady pulse that play music forward and provide temporal framework of a piece of music. It can be defined as a sequence of perceived pulses regularly spaced in time and corresponds to pulse that a human taps while listening to music.

The term *tempo* defined as the rate of the pulse. Musical pulses generally oscillate along notes onset or percussive events. It is a fundamental task to locate such events within a given signal, which is also known as *onset detection.* In this section, an overview regarding latest methods for extracting onset, tempo, and beat information from music signals are given and after that indication of information regarding application to higher-level rhythmic patterns is given.[12]

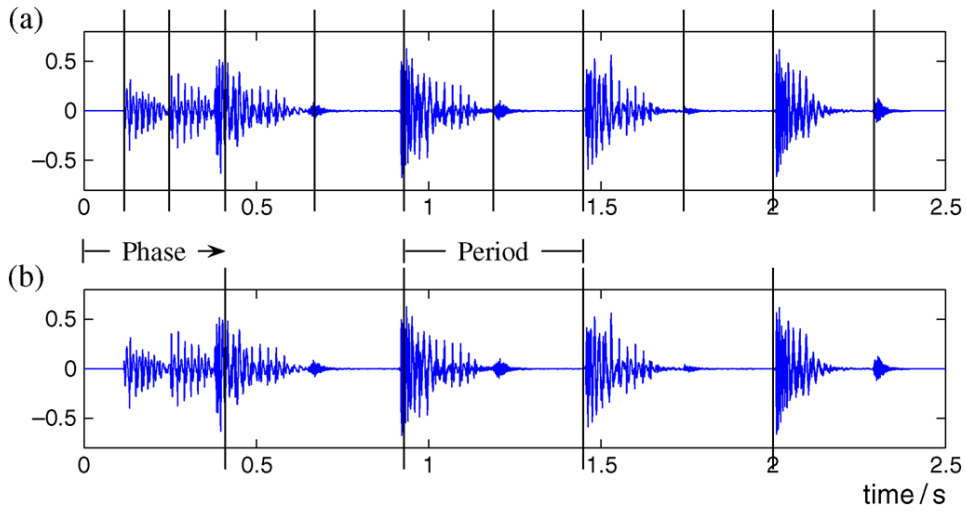

Figure 1.3 Waveform representation of the beginning of *Another one bites the dust*by Queen. (a) Note onsets. (b) Beat positions. [12].

1.4.8 MIR Toolbox in MATLAB[18]

MIR toolbox is a toolbox to extract musical features from audio, given by Olivier Lartillot and Petri Toiviainen. It is an integrated set of functions described in Matlab, dedicated to the musical features extraction from audio files. Functions for statistical analysis, segmentation and clustering are also included in the toolbox. Out of these feature extraction methods each one accepts an audio file or any preliminary result from

intermediary stages of sequence of operations. The distinct musical features which are extracted from the audio files are highly interdependent.

An overview showing primary features implemented in toolbox is presented in figure 1.4 Proceeding to the right a chain of operation is formed and all distinct processes starts from the audio signal(on the left). Each musical feature is resembled traditionally to one of the musical dimensions defined in music theory. Features resembling pitch and tonality are highlighted by boldface characters. Bold italics shows features corresponds to rhythm. Simple italics shows a wide set of features that can be resembled to timbre and dynamics. Among them, operators which are in grey italics can be applied to various distinct representations.

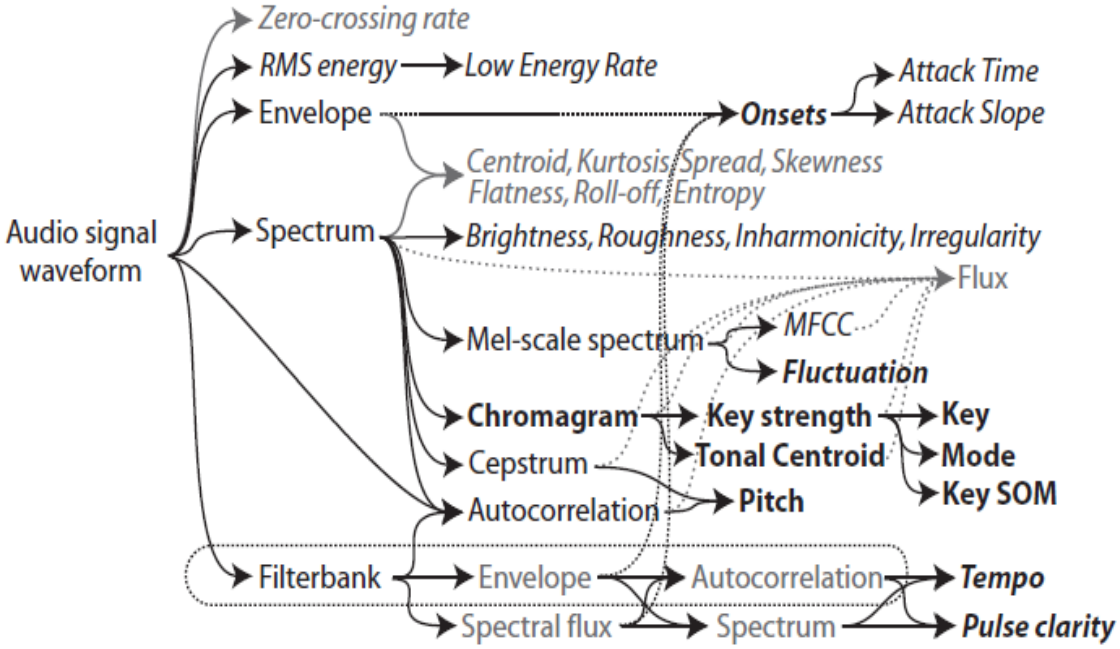

Figure 1.4 Overview of the musical features that can be extracted with MIRtoolbox [18].

1.5 Challenges

a) The main challenges for musical research is to extract relevant data from music signal and data as there is no correlation genre and statistical properties of sample data.

b) In the absence of properties mentioned in (a) , it is difficult to perceive model which could differentiable between genres and melodies.

c) A technique for pre-processing of data is also needed for the music signals to get the perfect result.

1.6 Novel Aspects of this work

a) Genre Classification of Indian raga has been attempted for the first time using ANN. ANN has already been used in other types of processing like signature authenication verification, speech detection, study on iris etc.

b) To the best of the author's knowledge features from wavelet domain have been used in musical genre classification tasks for the first time. In this thesis, new attempt has been made on music signal processing. Wavelet transform used here is continuous wavelet transform where as earlier similar techniques resembling continuous wavelet transform.

1.7 Thesis organization

This report consists of following 5 chapters.

Chapter 2: In this chapter, detailed review of the work in the field of music-analysis, has been presented along with the reported features extraction methods. Also gaps in the present state of the art have been identified.

Chapter 3: In this chapter, Back propagation and radial basis function neural network as the processing technique has been described.

Chapter 4: In this chapter, techniques for data extraction have been described. Pre-processing techniques are also presented.

Chapter 5: In this chapter, results obtained after processing are discussed in detail.

Chapter 6: In this chapter, whole work is concluded on the basis of results and observations.

CHAPTER 2- LITERATURE SURVEY

Any successful application must be able to show the specific properties of musical signals more precisely. Surpassing the colossal diversity of the musical notes, an engineer must be able to reduce the data to a number of key properties to work with. Even though the history of music is as old as mankind itself, use of machines for music production started in the last century around 1920 with the invention of Theremin, an instrument in which a person is able to modify the sound output by varying its capacitance. Further development in the next two decades led to initial steps towards electronic music, when some musicians used ring modulators as electronic instruments. But the golden age of electronic music can be said to have started with the arrival of computers for general purpose in the 1960s to 1970s. Many researchers such as Max Matthews and John Pierce used them to synthesize music. It also led to diversification within the music industry, with 'Blues', 'Rock', 'Metal', 'Jazz' etc. coming up as distinct genres. Notable are the advances made by bands like 'Pink Floyd' whose 'Psychedelic Rock' contained unique musical sounds, previously unheard, which could only be produced digitally. Further down the lane, it lead to emergence of new fields of musical data processing like Music Information Retrieval (MIR). International Symposium on Music Information Retrieval (ISMIR) is held annually since 2000 and showcases more than a hundred papers every year.

Notable papers from IEEE journals and conferences, and ISMIR conferences have been discussed below:

2.1 Literature Survey

- **Wasiq Khan, *et al*. 2014. [1]** This paper aims to understand the different techniques used for the recognition of an isolated work in continuous speech. This task is rendered difficult owing to speech having a dynamic nature. Variation in time may lead to the same word being recorded differently. The length and frequency level of each isolated word may differ, resulting in failing to find the best match using similarity measurement techniques. To overcome this, the literature proposes

filtration of frequencies causing mismatch using the wavelet transform. The performance and accuracy is much higher than those based on STFT and MFCC. This is depicted by experiments. For further improvements in the results, the pitch, frequency and vocal tract normalization should be considered.

- **Vahab Iranmanesh, *et al.* 2014. [2]** the authors, proposed the use of multilayer preceptor which worked on the basis of subset of principal component analysis features (PCA). Reduced error rates are achieved by doing a feature selection using PCA. 4000 signature samples from SIGMA database were taken generating False Acceptance Rates of 7.4% and Rejection rates of 6.4%.

- **Siddharth Sigtia, *et al.* 2014. [3]** examined 3 ways to get better feature learning for audio data using neural networks: 1. using Rectified Linear Units (ReLUs) instead of standard sigmoid units; 2. using a powerful regularization technique called Dropout; 3. using Hessian-Free (HF) optimization to get better training of sigmoid nets. They show that these methods provide considerable improvements in training time and the features learnt are better than the state of the art handcrafted features, with a genre classification accuracy of 83 ± 1.1% on the Tzanetakis (GTZAN) dataset. They found that the rectifier networks learnt better features than the sigmoid networks. They also demonstrated the capacity of the features to capture relevant information from audio data by applying the genre classification on the ISMIR 2004 dataset.

- **Franz A. de Leon, *et al.* 2014. [4]** gave several approaches for computer based genre classification of music using polyphonic timbre models. The authors, specifically compares the performance of the Gaussian mixture model (GMM), the Support Vector Machine (SVM), and the *k*-nearest neighbor model(*k*-NN). Features are extracted to model major attributes of timbre such as spectral envelope, range between tonal and noise like character, and spectra temporal evolution of sound. To address the scalability problem, which is a modified filter and then refine method is integrated with the *k*-NN classifier. With results showing that the 1-NN classifier with

filter-and refine method achieved the highest classification accuracy on the GTZAN and ISMIR2004 datasets.

- **Bilal Hadjadji, *et al.* 2014. [5]** Auto Associative Neural Networks are used to select most representative training samples. And later the same AANN is used to recreate the samples for generating the most valid representative model. Further, more experimentation was conducted for generating many real world benchmarks confirming the effective use of Selected training Samples for AANN as opposed to training on entire set. These systems have been effectively used in various recognition application given its superb track record, its main advantage being describing samples more accurately than other OCCs. On the down side, its sensitivity to noise or other outliers badly affect the representative model.

- **Alex Alexandridis, *et al*. 2014. [6]** This work aims to develop a system of high intellect with the ability of classifying different genres of music with an increase in the accuracy. An MGC which is based on RBF networks is presented, and it is trained using the algorithm PSO-NSFM. The algorithm is modified to suit the needs of classification problems. To replace accuracy as a function of cost, utilization of MCC is done. The results reveal that this method performs better than the standard algorithm using SFM.

Guangzhao Bao, *et al*. 2013. [8] the authors discuss under-determined blind source separation (BSS) based on two staged method of compressed sensing (CS) approach. A K-means method to predict unknown mixing matrix is exploited and sources from combined signals were separated using estimated mixing matrix from the first one. After that a 2 layer sparsity model was utilized. Assumption is made by this 2 layer sparsity model that components at low frequency for speech signal sparsed on K SVD dictionary and at high frequencies components were sparsed on DCT dictionary. This model takes advantage of two dictionaries which were designed at second stage, which can generate significant separation even if there is no sparsment of sources in time-frequency (TF) domain

- **Ching-Hua Chuan, *et al.* 2012. [9]** proposed an audio classification system based on wavelet transform to extract acoustic feature for low level in this paper. Many multiple-level decomposition by utilizing DWT to obtain acoustic features at various scales is performed and audio recording time was computed. The job of translating it into a compact vector representation is carried out. After that an Expectation Maximizing Algorithm is used to build sound classes model. Three kind of audio classification tasks are modeled to examine the system: 1) speech/music classification, 2) male/female speech classification, and 3) music genre (classical, pop, jazz, and electronic) classification. The evaluation of the system using fivefold cross validation, the results showed the efficient capability of wavelets for analysis of speech and music.

- **Gopala K. Koduri, *et al.*, 2012. [10]** worked on 'Caranatic' music. The authors discuss that the characteristics of ragas depends upon the expressions of the performer, and hence suggest an approach describing the intonation from a computational perspective, to obtain a compact representation of the pitch track of recording. As the first step, they extracted the pitch contours from automatically selected voice segments. Then, full pitch ranged pitch histogram is obtained, and further normalized by the tonic frequency, and each peak is formally labeled and parameterized. The authors confirm such parameterization by considering an explorative classification task: three ragas are disambiguated using the characterization of a single track. In this paper, the authors also mentioned their shortcoming that peak of few of the ragas were not identified with the using the algorithm.

- **Joe Cheri Ross, *et al*. 2012. [11]** worked on Indian classical music**.** They considered the segmentation of selected melodies from audio signals by calculating matching measures on time series of automatically predicted pitch values. The methods were examined regarding detection of signature phrase of Hindustani vocal musical compositions (bandish). Exploitation of musical knowledge regarding the metrical relations between the *mukhdamotif* and the underlying rhythmic structure was done in

order to reduce the search space, by utilizing available similarity measures and more efficient detection methods.

- **Meinard Müller, *et al*, 2011. [12]** gave an introductory paper describing all the techniques that are currently being used to study various psychoacoustic parameters related to music signal processing. The authors signaled that music signal processing may be wrongly alluded as to be the subset of the larger field of speech signal processing as various methods are originally modeled for speech have been used in music, mostly with better results. Although, music signals have some unique acoustic and structural features that make them different from spoken language or other non-musical signals. Melody, harmony, rhythm, and timbre etc are discussed at greater length in this paper. The authors also demonstrate that, to be successful, deep and thorough insight in the behavior of music itself is a prerequisite to understand music audio signal processing techniques.

- **Matthias Mauch, *et al.* 2010. [14]** successfully proposed a novel technique for chord and local key prediction by analyzing chord sequences and change in key factors is performed simultaneously. He also proposed a multi-scale approach for chroma vectors and with an enhancement in performance is presented when the selection of chords is made out from different sized chromas. Although the key estimation showed better performance in compare to direct template based method, the chord accuracy shows improved results.

H. Papadopoulos, *et al.* 2008. [15] presented that in western tonal music harmony and metrical structures parameters of most importance. Paper proposed a new scheme for simulataneous estimation of chord progression and downbeats based on an audio file. To accomplish this a uniue topology of Markov models is utilized that allowed to model chord dependency on metrical structure. This model is checked on sample data set of sixty six famous music songs from the Beatles and showed improvement when compare to state of art technique.

Andre Holzapfel, *et al.* 2008. [16] proposed a novel method to predict onsets in musical signals on the basis of the phase spectrum and the mean of group delay function used. Analysis based on frame basis for music signal give evolution of group delay over the time, called as phase slope function. Detection of onsets is done by finding the positive zero crossings for the phase slope function. The comparison is made out between the proposed approach and amplitude based onset detection approach in state of art framework for beat tracking. Beat tracking accuracy for data set with less percussive measure obtained had an improvement of 82%, when a phased onset approach for detection is utilized rather a amplitude based scheme.

- **Khalid Youssef, *et al.* 2008. [17]** the authors presented us with a latest musical note recognizer system based on time delayed neural approach, self organizing maps and linear vector quantization. Illustration of 2 different application on the basis of this system is done.

Olivier Lartillot, *et al.* 2007. [18] the readers are introduced to MIRtoolbox, which is an combined set of functions in MATLAB software. This toolbox is deicated to extract musical features from audio files. This design is made out on the basis of distinct modular framework, with distinct algorithms divided into multiple stages. Formalization using a minimal set of elementary techniques, and combining different variants given by other approaches – considering new developed mechanisms, which can be selected by users and parameterized. An overview regarding set of features such as timbre, rhythm or totality is given that can be obtained with application of MIR toolbox.

- **Daniel P.W. Ellis, 2007. [19]** The classification of musical audios is addressed by modeling the statistics of broad spectral features which exclude pitch information and go on to reflect mainly instrumentation. Beat-synchronous has been used to investigate chroma features, which are modeled to show melodic and harmonic content and be not variant to instrumentation. Chroma features have a drawback in such that these do not contain information about classes such as artist, but also

contain data that is almost entirely independent of the spectral features, and hence the two can be combined to our benefit.

- **Aliaksandr Paradzinets, *et al.* 2007. [20]** High level musical features are proposed in this paper which can be of use for automating the music navigation system and its further classification. It studies the Continuous Wavelet Transform (CWT) based approach for the same. The similarities in the rhythm and melody have been extensively evaluated. Also, a comparison has been outlined of the proposed algorithm with other similar algorithms. The most important results has been the FFT transform showing improvement.

- **Bin Li, *et al*. 2007. [21]** described a methodology based on online signature verification utilizing null component analysis (NCA) and principal component analysis (PCA). K-L transform was used to transform the designate set of feature vectors and separate into null components (NCs) and principal components (PCs). To examine stable and unstable components of reference set both NCA and PCA approach were used. Work on data samples for 1,410 signatures of 94 signers showed that the online signature verification based on NCA/PCA method obtained best results with an equal EER of 1.9%.

- **Slim Essid, *et al.* 2006. [23]** The importance of musical instrument recognition is highlighted in this paper. It focuses on 10 major instrument types, mapped with a large database of sounds and studied with the help of over 150 signal processing features, which also have the inclusion of new descriptors. GMM and SVM classifiers were studied and studied in a one versus one scheme. The best results were achieved by SVM with RBF kernel, giving a 12% average improvement from the baseline system. In future, the work would consider techniques suitable to the SVM classification and the instrument recognition would be introduced at higher levels of taxonomy.

- **Douglas Turnbull, *et al.* 2005. [24]** The paper extensively deals with audio tracks being automatically classified into their respective genres of music. It aims to achieve the level of accuracy of humans using RBF networks. The initialization methods deliver accurate results and are trained with a hundred times slower gradient descent. The method successfully maps the accuracy of genre classification as attained by humans by using the RBF network.

- **Miguel Alonso, *et al.* 2004. [25]** an efficient beat tracking algorithm that processes audio recordings has been presented in this paper by the authors. The concept of spectral energy flux is also defined by them and later used to derive a new and effective onset detector based on the STFT. High performance for a large range of audio signals is displayed by these detectors. In addition, it is a very simple system that is proposed for tempo tracking and it is quite easy to implement and uses very little computational power.

- **Gianpaolo Evangelista, *et al.* 1998. [29]** This paper expands the definitions of dyadic wavelets for the inclusion of frequency warped wavelets, which are generated in discrete time by Laguerre transform alterations using the reconstruction of filterbanks. The definitive inclusion of pitch-synchronous wavelets vastly enhances the analysis of transients and denoisation of inharmonic pseudo-periodic signals. A unique orthogonal warping is obtained from the discrete Laguerre transform, which is computed by rational transfer functions. Frequency warped wavelets possess an arbitrary structure of band that can be used to adapt to perceptual characters or to signals.

- **Peter De Gersem, *et al.* 1997. [30]** The similarities of time-frequency descriptions and constant relative bandwidth between music and wavelet transformations have been depicted in this paper. For the extensive analysis, processing and synthesis of music, the Continuous Wavelet Transform (CWT) forms an interesting tool. However, it offers a high computational cost at the moment, leading to researchers looking into more efficient schemes. Other techniques result in more cost-effective

analysis in the field of music, but they lose some of the interesting properties used in the CWT synthesis.

- **Benyamin Matityaho, *et al.* 1995. [31]** a novel approach to recognize musical types using multi layered neural network as the decision making system is proposed. The model was presented for classify two types of music and shows success in both the cases. Thus one is able to tell which music genre is being played even if by looking at the signal description no distinction is seen.

2.2 Gaps in Study

- Some of the psychoacoustic parameters are not defined. Exact mathematical definition of timbre is not available.
- Pre-processing techniques are to be analyzed.
- No attempt has been made to combine PCA with music signal processing. PCA can be combined with music samples as pre-processing techniques.
- Study done on Indian musical genres like raagas and folk song, and Indian music instruments like tumbi, sarangi, algoze etc is inadequate.
- Multitrack editor of most of the Indian folk instruments is not commercially available. Adequate study is to be taken up as music world is progressing at very fast pace.
- No commercial modeling tool exits in the market. There is almost no efficient tool available for writing MP3 audio tracks into sheets

CHAPTER 3- ARTIFICIAL NEURAL NETWORK CLASSIFIER

3.1 Introduction to Neural Networks

Artificial neural network has been motivated from human brain, but it works in an entirely different manner. The brain is a highly complex, nonlinear and parallel computer. The structural constituents of brain, called neuron are capable of performing most of the computations much quicker than any of the computer known. When a baby is born, its brain has immense structure and it grows up it learns about its surroundings with time. This learning is known as "experience". Experience is constructed over time, most dramatic experience is built in two years of birth; but the development continue well beyond that stage too.

Artificial Neural Network is a pseudo network, constructed to replicate the human neural network in order to performs the tasks or functions of interest. The network is simulated in software or is implemented using electronic components on a digital computer.

3.2 Human Brain

The human nervous system can be seen as a three-stage system, as described in the block diagram of Fig. 3.1, middle to the system is the brain, neural (nerve) net, which continuously receives information, depicts it, and make appropriate decisions. Two types of arrows are shown in the figure. Black pointing from left to right represent the forward transmission of information-bearing signal through the system. Blue arrows pointing from left to right represent the presence of feedback in the system. The receptor converts stimulus from the human body or any external environment into electrical impulses those the convey information to the neural net.

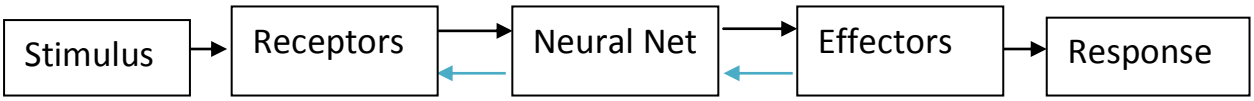

Figure 3.1 Block diagram of human nervous system

The difficulty to understand the brain has been made convenient as a result of the pioneering work of Ramon y Cajal (1911), who gave the idea of neurons as structural constituent of the brain. Usually, neurons are five to six orders of magnitude slower than the silicon logic gates; events in a silicon chip happen in the range of nanoseconds, whereas neural events happen in the range of milliseconds. However, the brain makes up for the relatively slower operational rate of a neuron processing by having a huge number of neurons (nerve cells) with massive interconnections between them. It is expected that there are almost 10 billion neurons in the human cortex, and 60 trillion connections. The end result is that the brain is very efficient structure. To be specific, the energetic efficiency of brain is nearly 10-16 J (joules) per operation per second, whereas the equivalent value for the finest computers is orders of magnitude larger. Nerve endings, or synapses, are elementary structural and functional units those mediate the interactions between neurons.

3.3 Models of Neuron

A neuron is an information-processing unit which is fundamental for the operation of a neural network. We describe three basic elements of the neural model as:

1. A set of synapses, connecting links, each of which is defined by its own weight or strength. To be specific, a signal $x_{j,}$ at the input of synapse 'j' is connected to neuron k and is multiplied by the synaptic weight w_{kj}. It is very important to know the meaning in which the subscript of the synaptic weight w_{kj} is written. The first subscript, k, in w_{kj} refers to the neuron being used, and the second subscript, j, refers to the input end of the connecting neuron to which the weight refers. Contrary to the weight of a synapse in the brain, the synaptic weight of the artificial neuron may also lie in a range that includes negative along with positive values.

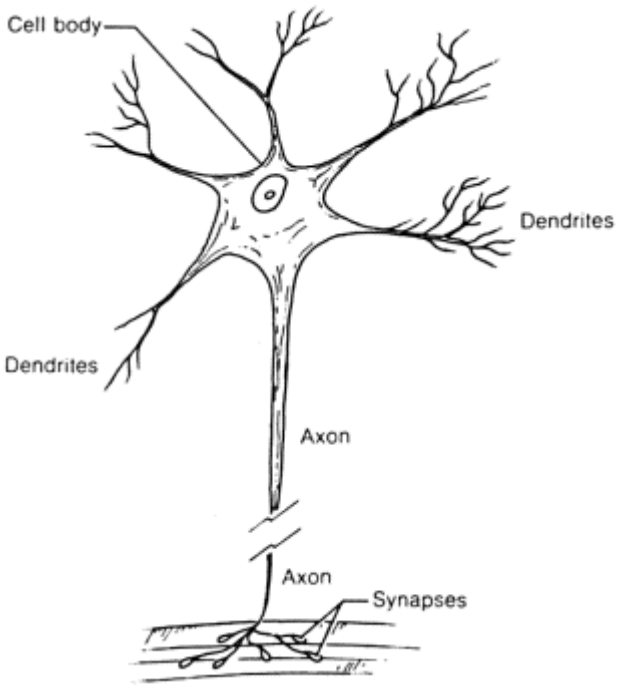

Figure 3.2 Structure of Neuron [27]

2. An adder, for summing of individually weighted input signals; the operation described here is a linear combiner.
3. An activation function, is for the controlling amplitude of the output of the neuron. The activation function is also named as a squashing function, in that it limits (squashes) the allowable amplitude range of the output to some specific finite value.

The neural model of Fig. 3.3, there is an extra element, an externally applied bias, represented by b_k. The bias, b_k, effects the increasing or lowering of the net input of the activation function, depending on whether the value is positive or negative, respectively.

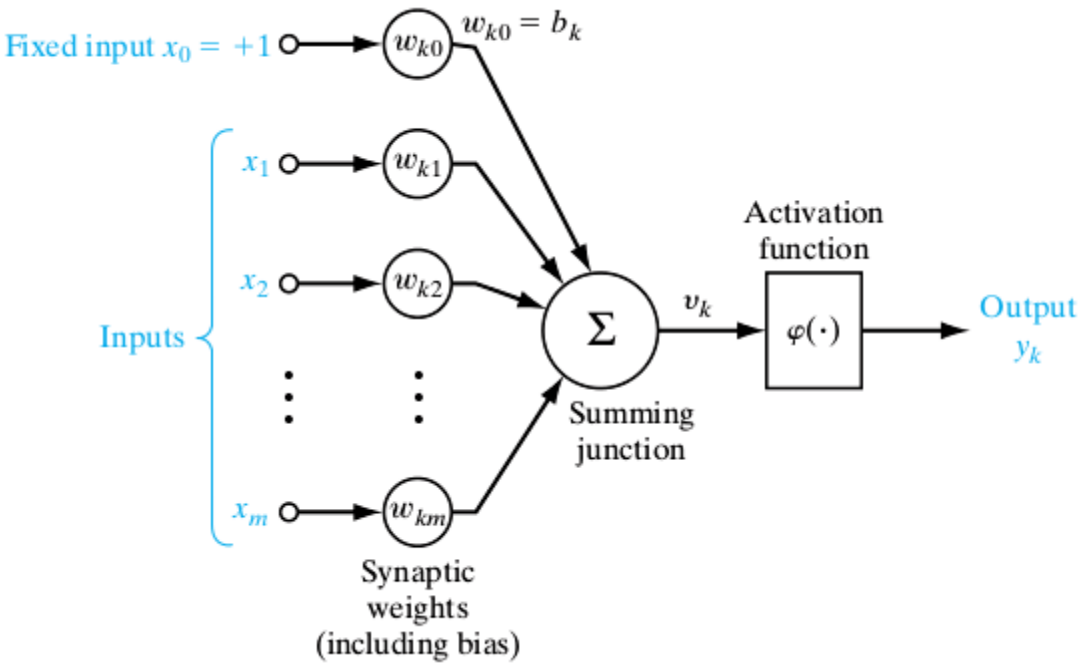

Figure 3.3 Nonlinear model of a Neuron

3.4 Perceptron

The perceptron has played an important role development of neural networks: It was the first algorithmically designed neural network. Its invention was done by Rosenblatt, a psychologist. This discovery inspired engineers, physicists, and mathematicians to devote their research effort towards the different aspects of neural networks in the 1960s and 70s. In the years, 1943-1958, several researchers have contributed alot:

- McCulloch and Pitts, in 1943, introduced the idea of neural networks as computing machines.
- Hebb, in 1949, postulated the first rule for self-organized learning.
- Rosenblatt, in 1958, proposed the perceptron as first model for learning along a teacher (i.e., supervised learning).

Linearly separable types of patterns are classified using the simplest form of neural network, i.e. Perceptron. It consists of single McCulloch-Pitts neuron with an adjustable synaptic either 1 or 0 along with weights and bias (threshold).

Perceptron consists of single neuron with variable synaptic weights and bias. The algorithm which is used to adjust the free parameters of the neural network first appeared

in a learning procedure developed by Rosenblatt (1958, 1962) for his perceptron brain model.

The considering the figure, equations given below describe the way computations are done in a perceptron.

$$\sum_{k=0}^{m} w_k(i)x_k(i) = v(i) \quad (3.1)$$

$$y(i) = \varphi(v(i)) \quad (3.2)$$

$$e(i) = d(i) - y(i) \quad (3.3)$$

The manner in which error signal is used to control the adjustment to the neuron's synaptic weights is determined by the cost function.

Using single layer perceptron we can only classify the linearly separable patterns like OR, AND logic but we cannot classify non-linear patterns like EXOR using single layer perceptron. The perceptron is an uni-layer neural network, its operation is based on error-correlation learning. Term "single layer" is used here to signify the fact that the computation layers of the network consist of one neuron for the case of binary classification. Finite number of iterations are performed while learning process. the patterns has to be linearly separable for the classification to be successful.

3.4.1 Multilayer Perceptron

We define at a neural network structure as the multilayer perceptron, to overcome the practical limitations of the perceptron.

The following three points are for highlighting the basic features of multilayer perceptron:

- There is an unique nonlinear and differentiable activation function in model of each neuron .
- Few of the hidden layers are present in the network which are hidden from both input and output nodes.
- The network has a high degree of connectivity, the amount of which is determined by synaptic weights of the network.

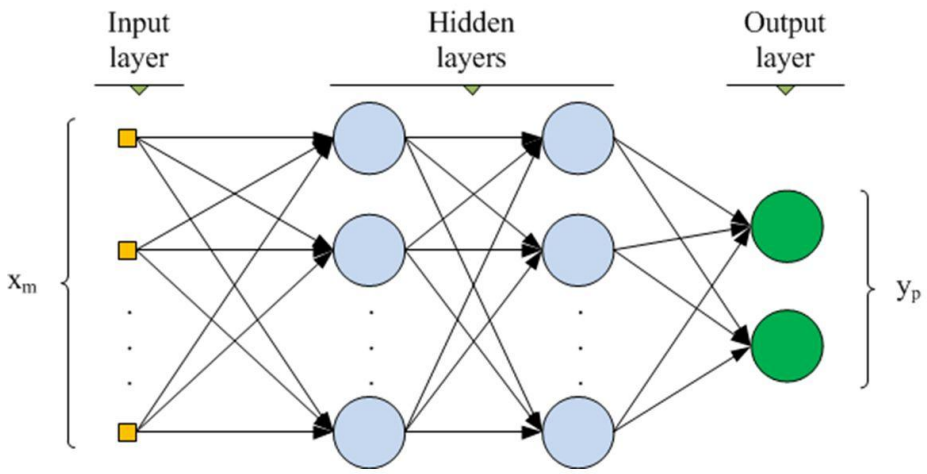

Figure 3.4 Architectural Graph of Multilayer Perceptron with only two hidden layers

In figure 3.4, there are two hidden layers and an output layer. The network shown here is fully connected in order to set the stage for the description of multilayer perceptron in its general form. This shows that any neuron in a specific layer of the network is well connected to all the nodes (neurons) in the previous layer. From left to right, signal flow in the network progresses in a forward direction on a layer-by-layer basis.

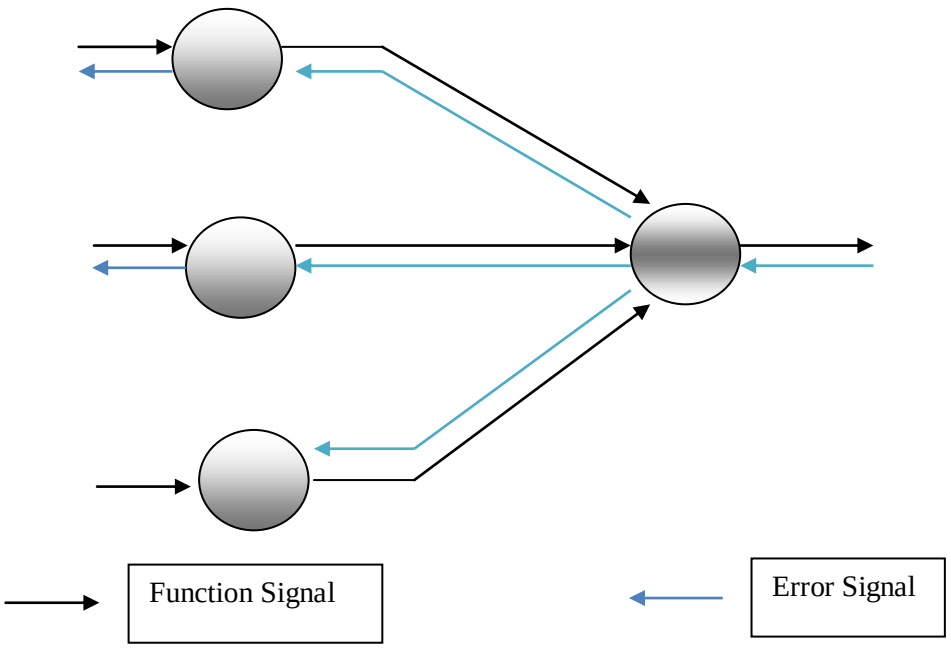

Figure 3.5 Representation of the directions of two basic signal flows in the multilayer perceptron: forward propagation of the function signals and back propagation of the error signals.

Figure 3.5 represents a portion of the multilayer perceptron. Two kinds of signals are seen in this network:

1. **Function Signals**

   A function signal is stimulus (input signal) that is given at the input end of the network, transfers forward (neuron by neuron) in the network, and output is recieved at the output end of the network. Such a signal is referred a "function signal" for two main reasons. First one, it is assumed it will perform useful function at the output of network. Second is, at each and every neuron of the network via which a function signal passes through, the signal is tested as a function of the inputs and related weights applied to the specific neuron. That function signal is also known as the input signal for network.

2. **Error Signals**

   An error signal, which originates at the output neuron of the network, goes backward (layer by layer) throughout the network. It is known an "error signal" as its computation at every neuron of network involves an error-dependent function in either one form or another.

Output neurons comprise only output layer of the network. The remaining neurons comprise all hidden layers of network. Therefore, the hidden units are not considered as part of output or input nodes of the network—hence these are named as "hidden". The very first hidden layer is given input from the input layer, which is made up of sensory units, the outputs resulting from the first hidden layer are applied as input to next hidden layer, and so on for the rest of network.

Every hidden or output neuron of a multilayer perceptron has to perform two necessary calculation

1. The calculation related to the function signal evolving at the output of each neuron, which is represented as a continuous nonlinear function of input signal and the weights linked with that neuron.

2. The calculation of an approximation of the gradient vector (i.e., the gradient of error surface in relation with the weights connected to the inputs of the neuron), which is required for the backward propagation through the network.

3.5 Learning Process

There are many different ways in which we learn from our own surroundings and environments, same happens with the neural networks. In broader sense, we may classify the learning processes in the method in which neural networks function as follows: learning along the teacher and learning without the teacher. Second form of learning may be named as unsupervised learning and reinforcement learning. when compared to brain, the learning techniques of neural network are same as of learning techniques of brain, and these are:

a) Error-Correction Learning
b) Memory Based Learning
c) Hebbian Learning
d) Competitive Learning

Algorithm

Artificial neural Network is family of automatic learning algorithms based on biological nervous system of animals, specifically the brain. ANN is used to estimate the functions which are the expected result for the unknown inputs which are large in numbers[22] [31]. ANN is highly adaptive in nature, thus these are very capable of machine learning along with pattern recognition. To compute the value from input it is represented by system of interconnected "neurons". It is a technique in which machine learns from data provided to it. Network can be trained to the best approximations even if the function is non-linear. The parallel computations are performed which reduces the computation time to the great extent. In this paper, we have used back propagation method.

At the output of n^{th} iteration for j number of neurons, the error signal is defined by

$$e_j(n) = d_j(n) - y_j(n) \qquad (3.4)$$

Corresponding total instantaneous energy Ɛ(n) is given by

$$Ɛ(n) = \frac{1}{2}\sum_{j\in C} e_j^2(n) \qquad (3.5)$$

Where, C is the set which includes all the neurons in output layer of the network. On summing of Ɛ(n) over all n average squared energy is received, and it is normalized with respect to set size N, it is given by,

$$\mathcal{E}_{av} = \frac{1}{N}\sum_{n=1}^{N} Ɛ(n) \qquad (3.6)$$

For any given training set, $\mathcal{E}_{av}$, represents the cost function, which deals with the learning performance. It is needed because it helps to adjust the parameters of the network to have minimum value of $\mathcal{E}_{av}$.

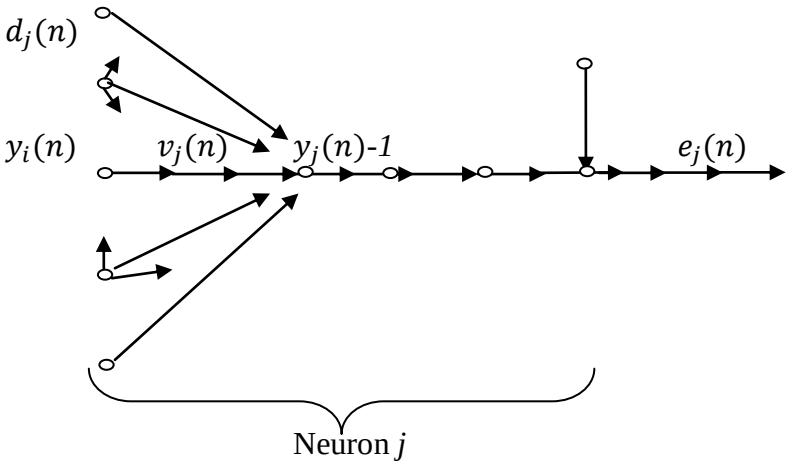

Figure 3.6 Signal flow graph representing the details of output j^{th} neuron.

figure 3.6, represents the neuron j which is fed by a set of signals which are produced by the neuron which are on its left. The induced local field, $v_j(n)$, is produced at the activating function associated with neuron j,

$$v_j(n) = \sum_{i=1}^{m} w_{ji}(n)y_j(n) \qquad (3.7)$$

Where, m is the total number of inputs.

3.6 Back Propagation Algorithm

Assume the three-layer neural network with N input neurons, H hidden neurons and M output neurons respectively. The output of the m^{th} output node due to the p^{th} input pattern is given by o_{pm} , whereas the output of the k^{th} hidden node for the p^{th} input pattern is given by o_{pk}. Let w_{km} be the weight between the m-th output neuron and the k-th hidden neuron, and w_{nk} be the weight between the k^{th} hidden neuron and the n^{th} input neuron. The desired output for the m^{th} output neuron due to the p^{th} input pattern is given by t_{pm} .The input for the n^{th} input neuron due to the p^{th} input pattern is denoted by x_{pn}.

Using this definition, the output of the k-th node in the hidden layer is given by:

$$o_{pk} = f\left(\sum_{n=1}^{N} w_{nk} x_{pn}\right) \tag{3.8}$$

Where f is the activation (sigmoid) function defined as

$$f(x) = 1/(1 + e^{-x}) \tag{3.9}$$

Similarly, the output of the m-th node in the output layer is given

$$o_{pn} = f\left(\sum_{k-1}^{k} w_{kn} o_{pk}\right) \tag{3.10}$$

We define the sum of squared error of the system to be:

$$E = 1/2 \sum_{p=1}^{P} \sum_{m=1}^{M} \left(t_{pm} - o_{pm}\right) \tag{3.11}$$

The back-propagation learning algorithm is to change the current weights w_{km} and w_{nk} iteratively such that the system error function E is minimized. The weight updates are proportional to the partial derivative of E.

The partial derivative of *E* with respect to $w_{km,}$ is:

$$\frac{\partial E}{\partial w_{km}} = \frac{\partial E}{\partial o_{pn}} \cdot \frac{\partial o_{pm}}{\partial w_{km}} \quad (3.12)$$

Where $\frac{\partial E}{\partial o_{pm}} = o_{pm} - t_{pm}$ and $\frac{\partial o_{pm}}{\partial w_{km}} = o_{pm}(1-m)o_{pk}$ (3.13)

And the partial derivative of Φ with respect to w_{nk} is

$$\frac{\partial E}{\partial w_{nk}} = \sum_{m=1}^{M} \frac{\partial E}{\partial o_{pm}} \cdot \frac{\partial o_{pm}}{\partial o_{pk}} \cdot \frac{\partial o_{pk}}{\partial w_{nk}} \quad (3.14)$$

$$\frac{\partial o_{pm}}{\partial o_{pk}} = o_{pm}(1 - o_{pm})w_{km} \quad (3.15)$$

$$\frac{\partial o_{pk}}{\partial w_{nk}} = o_{pk}(1 - o_{pk})x_{pn} \quad (3.16)$$

The weight change for the (n+l)-th iteration can be expressed as follows, where μ is the learning rate of the gradient method.

$$\Delta w_{km}(n+1) = \mu \sum_{p=1}^{P} \delta_{pm}\, o_{pk} \quad (3.17)$$

Where,

$$\delta_{pm} = (t_{pm} - o_{pm})o_{pm}(1 - o_{pm})\Delta w_{nk}(n+1) = \mu \sum_{p=1}^{P} \delta_{pk}\, o_{pn} \quad (3.18)$$

Where $\delta_{pk} = o_{pk}(1 - o_{pk}) \sum_{m=1}^{M} \delta_{pm}\, w_{km}$ (3.19)

The conventional back-propagation algorithm is basically a gradient-descent method; it has the problem of getting trapped in local minima, by which back-propagation may lead to failure in finding a global optimal solution**.** Besides, the convergence rate of back-propagation is still too slow even if learning can be achieved. Attempts to improve the performance of the original back-propagation algorithm have concentrated on:

i) selection of better activation function

ii) selection of dynamic learning rate and momentum

3.7 Radial basis Function

Radial Basis function neural network is a type of artificial neural network mostly used for regression, function approximation, classification and data clustering problems. A radial basis function neural network has three basic layers named as input layer, output layer

and in between them is hidden layer. The input vector of each unit reaches hidden layer when given as input to input layer. Using radial basis function in the layer each hidden layer produces an activation in the layer. Ultimately, linear combination of the activation is computed in each of the hidden layer unit and thus classified output is produced in the output layer. The use of activation function produced in hidden layer effects the output and the weights related with the connections between the hidden layers and output layers.

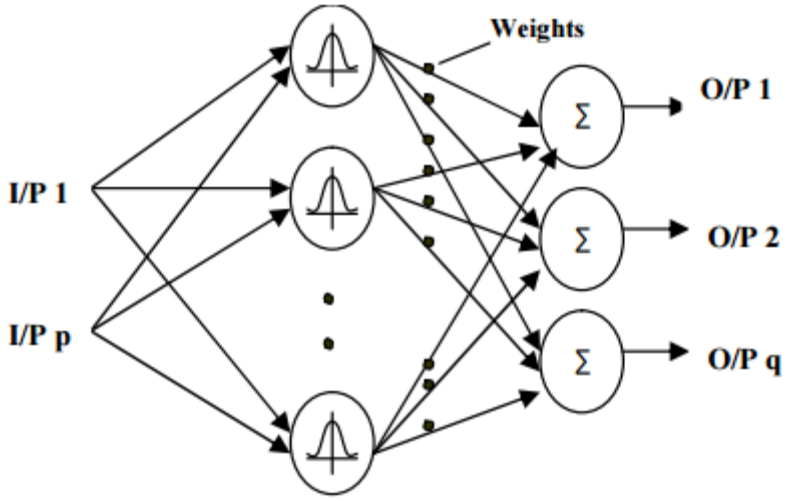

Figure 3.7 Radial basis Function Artificial neural Network Architecture

The optimization criteria used is different depending on the problem where radial basis function is to be used and corresponding learning algorithm. Data classification is one of the main applications of RBF.[6] The number of hidden units, weights and activation function associated with hidden layer and output layer is determined by the most learning algorithms. The mathematical form of output unit in radial basis functions if given as:

$$f_j(x) = \sum_{i=1}^{h} w_{i,j} r_i(x) \qquad (3.20)$$

here, f_j represents the function which is linear a combination of 'h' radial basis functions $r_1, r_2 \ldots , r_h$ corresponding to the j^{th} output.

3.8 The Neural Network toolbox in MATLAB:

MATLAB 7.11.0 includes in its collection of toolboxes a comprehensive Application program Interface (API) for developing neural networks. Along with console-based

programming facilities, **MATLAB 7.11.0** includes an easy-to-use Graphic User Interface (GUI) that encapsulates all functions available in the **Neural Networks Toolbox**.

In MATLAB's command window type nntool

>>nntool

A GUI of NNTOOL **Network/Data Manager GUI** opens up **:**

In this figure 3.8, it can be seen there are few block like input data, target data, network, to which data is imported using import and new clicks. It is an option to import input ant target data from MATLAB workspace..

In figure 3.9, it can be seen that, it is outcome of clicking at new click in fig 3.8. In this we need to select the input and target that would be available after importing them to nntool. Then create the network. Network would be created. Imported data and network would be available in input, target and network respectively.

Figure 3.10 is outcome of clicking at networks created after figure 3.9. The created network is saved in network block available in figure 3.8. Clicking on desired network would open circuit diagram as in form of figure 3.10. Click at train tab of block. The training Input and Target are given as input and target to train the network. In training performance tab epoch is set to desired number. Epoch is the maximum number of iterations one wishes to allow in the training. Goal is set as per requirement.

On clicking the train network, figure 3.11 gets open. In this figure some epoch values are seen. It tells how many iteration have the tool gone through for the specific training. Time tells how much time it has taken for the same. Performance tells what the performance is for the specific case.

The minimum performance after changing the weights is calculated for many networks with different number of neurons those can be changed in network creation.

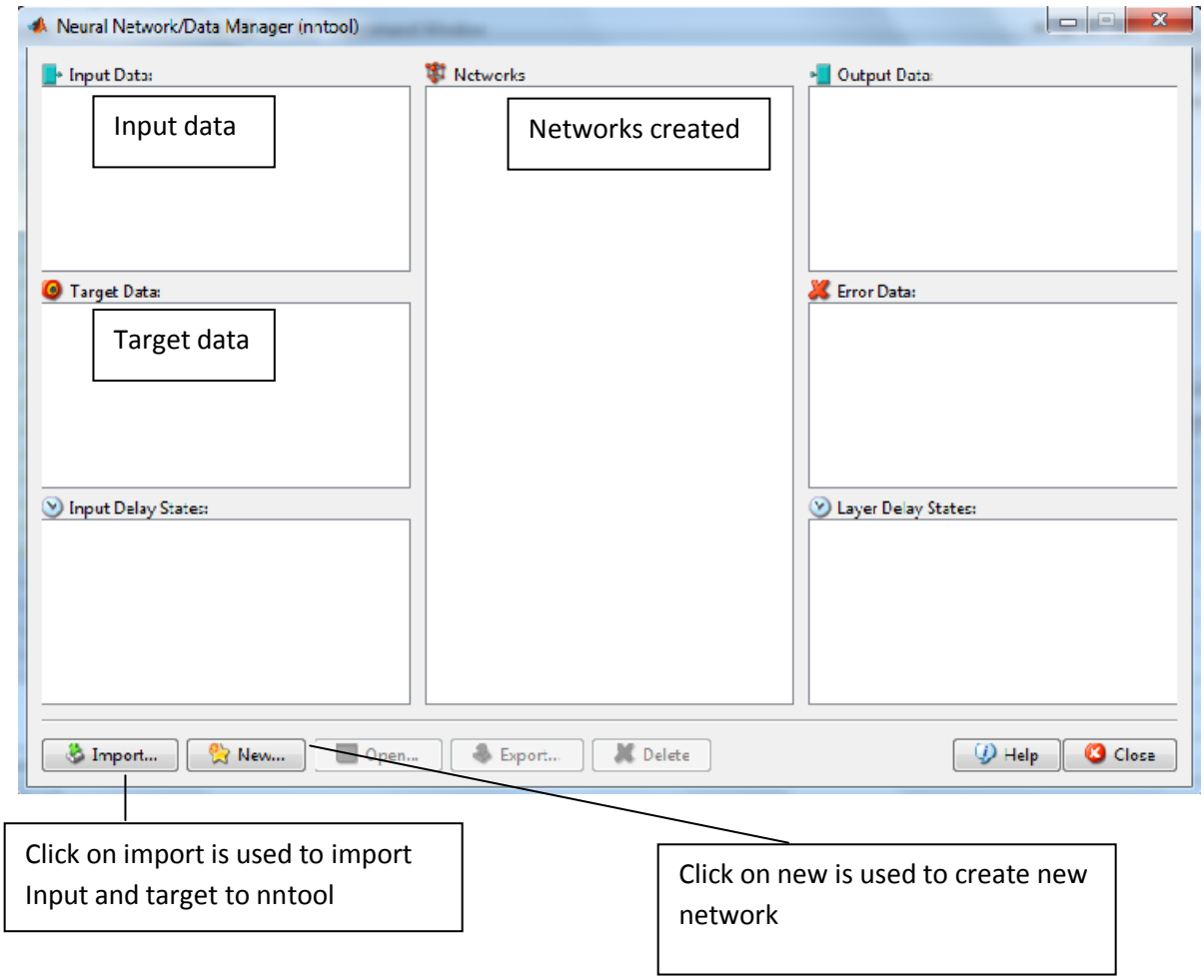

Figure 3.8 Neural network/data manager (nntool)

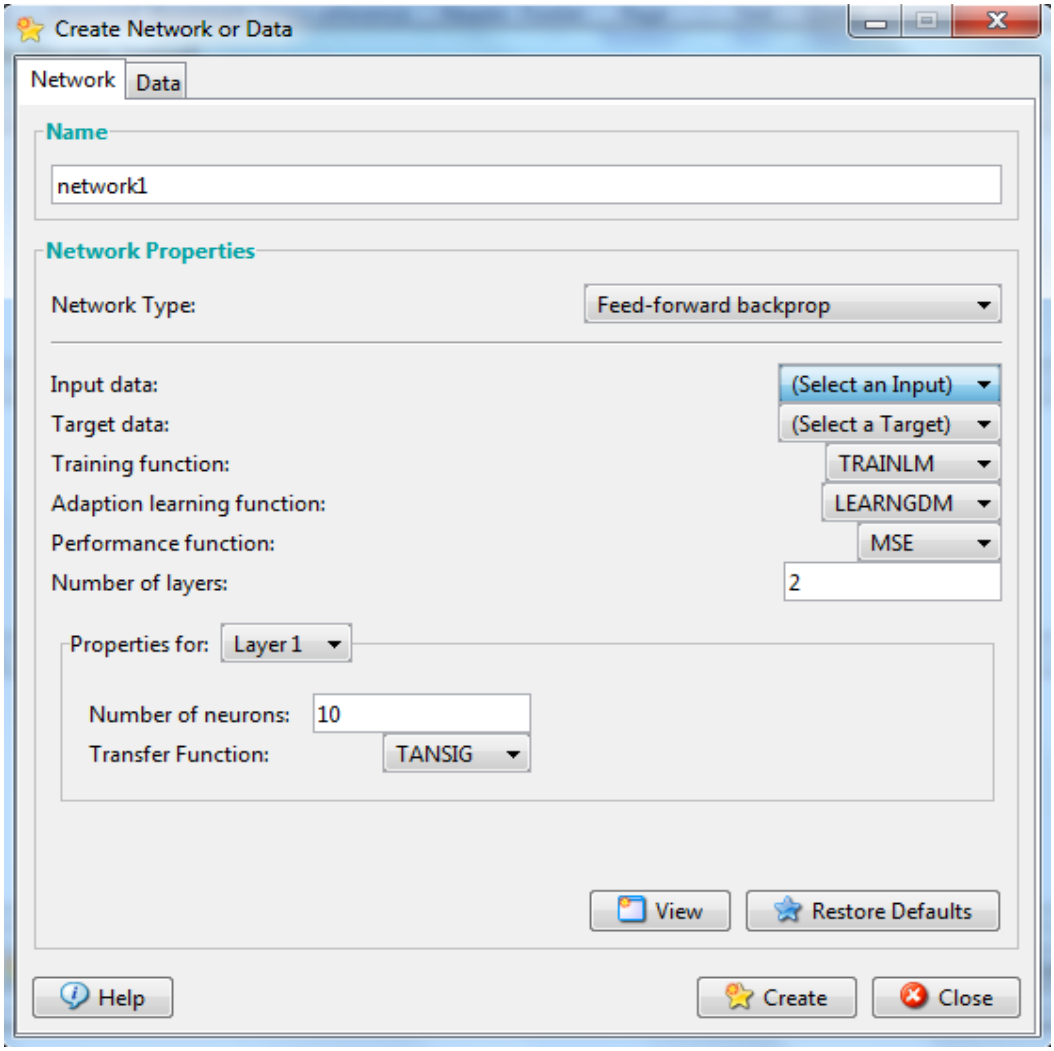

Figure 3.9 Create Network using data

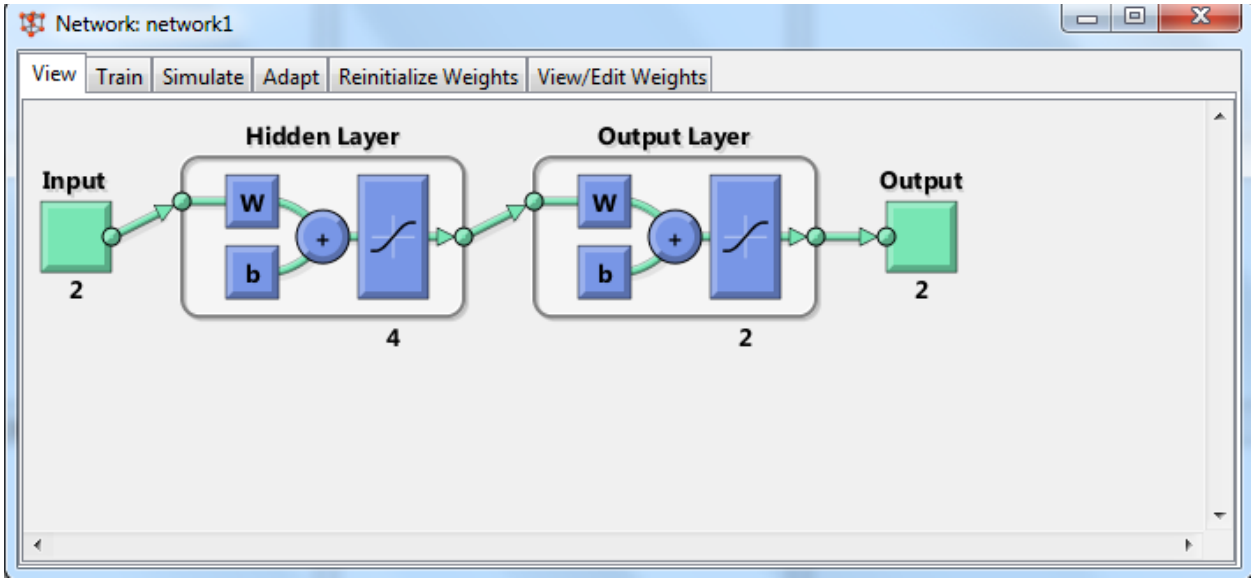

Figure 3.10 Block diagram of Network 1 created

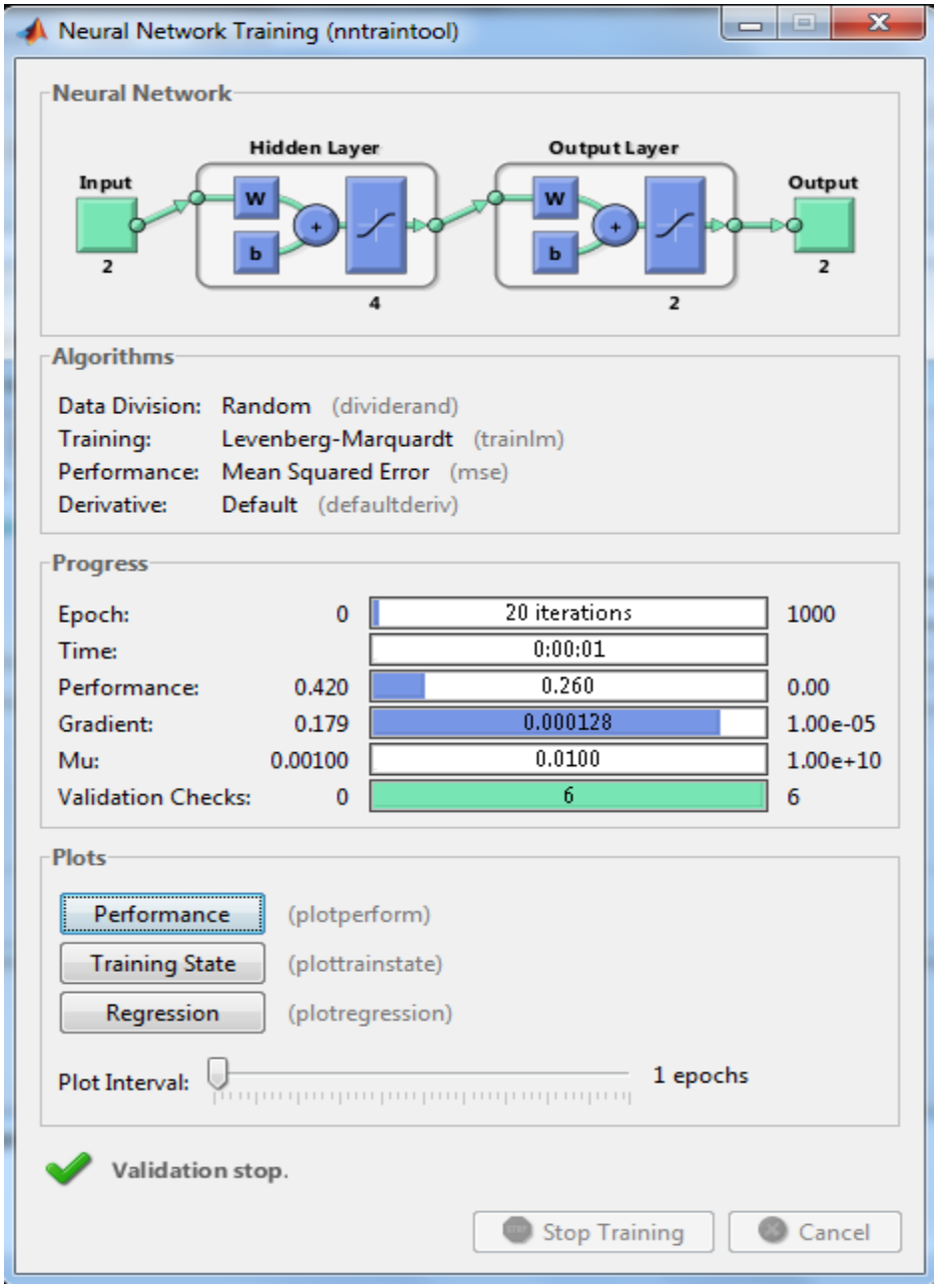

Figure 3.11 Training of Network

The performances for all the iterations are noted and finally minimum of all is selected. For each architecture corresponding to each simulation run, the initial weights are noted. Using the test data the architecture is tested. Thus the performance level of the nntool for test samples is noted.

CHAPTER 4- DATA and FEATURE EXTRACTION

4.1 Data description

The music signal to be analyzed is extracted from the original music sample. In this work various raagas have been analyzed. 4 Raagas namely Asa Raag, Bhairav Raag, Gauri Raag and Malar Raag are studied.

The downloaded files are in the form of MP3 format. But MATLAB doesn't support big MP3 files, thus these files are required to be converted into WAV (Waveform Audio**)** file format. For this, software is downloaded using[26]. This software is convertor and cutter too. It converts MP3 to WAV format and also cuts it to desired timing. Like, I have used the sample to be of 10 seconds each. These signals were one by one imported to MATLAB. From workspace the data was copied and saved as variables in editor window. After having desired number of samples the processing was started, using Neural Network. In total, 80 samples are used, 20 each from Asa Raag, Bhairav Raag, Gauri Raag and Malar Raag. 75 % of the total samples are for training the neural network and remaining 25% are used for testing.

4.1.1 Data set

To acquire the first set of data, the music files are read directly in MATLAB from their WAV file format, using the command wavread. The data, ehich is then extracted, has 444,000 total values, corresponding to each music sample. Out of those value, the top 256 values are used from each sample. In total, 80 samples are used, all from different ragas.

For the second set of data, the values are taken from the utput of the chromogram, after taking the spectrum of the raw data. This is done by a specific procedure. The data set which is being used is basically processed data. This processed data is the output for each sample. The chromagram data has 61 values in each sample.

Thereafter, the next set of data is used from the first and second data set. Then, data extraction techniques are applied to both the data sets. Therefore, the resulting data set that is formed is raw, with wavelet transform being applied to each of its samples, and the

other of chromogram data with wavelet transform. Finally, PCA is applied on the chromogrammed data, and on wavelet transformed data. In total, 8 data sets are created.

4.2 Feature processing and extraction

4.2.1 Chromagram

The chromagram, also referred to as Harmonic Pitch Class profile, is the display or description of energy distributions along the pitch classes.

Firstly, we need to define the chroma spectrum, X(n), it is the measure of signal's strength with respect to a given value of chroma C. Chroma C, is just the fractional part of the based 2 logarithmic output of frequencies, that is concluded with using Shepard's helix of pitch perception[18][8]

$$C = \log_2 f - \lceil \log_2 f \rceil \qquad (4.1)$$

In equation (1), $\lceil \rceil$ represents the greater integer function. The chroma spectrum represents the standard Fourier power spectrum. On frame fragmenting before calculating fourier spectrum, we can also create time-frequency distribution, that is denoted by X(t, f). We can define a "time-chroma" distribution, X(t ,C) as

$$X(t, C) = \text{fun}(X(t, f)) \qquad (4.2)$$

Here, fun is an aggregation function, which is used in the remapping of the time frequency distribution. The distribution is named as 'Chromagram'. 'f' is given by 2^{C+n} which is obtained from equation (1) and the aggregation function we use here is summing function. Elements of chroma features for t^{th} frame of V (t, k) can be obtained by using the formula

$$V(t, k) = \sum_n \left(\frac{X_t(n)}{N_k}\right) \qquad (4.3)$$

Where $k \in \{0, 1, 2, \ldots, 11\}$

$n \in S_k$

$X_t(n)$ is the logarithmic result of the DFT (Discrete Fourier Transform) of the t^{th} frame. N_k is number of elements in S_k, which is subset of the discrete frequency region for each pitch class. The arithmetic mean of log magnitude DFT values belonging to set S_k is taken and then each feature vector is normalized by subtracting scalar mean of the vector's feature set. The set S_k has 12 features generated by association of each DFT bin corresponding to each of the 12 pitch classes. The associated frequency of DFT bin is calculated and chroma value is obtained using formula (4.1). The result saved the bin belongs to the pitch class which has the nearest chroma values. The chroma values are reassigned such that pitch class B has centre at chroma values 1, pitch class C has centre around value 0 an other pitch classes have their centre at chroma values k/12[13]. The range of spectrum is limited. The lower bound is chosen at 20 Hz and upper bound at 2000 Hz. The chromagram of audio signal is shown as:

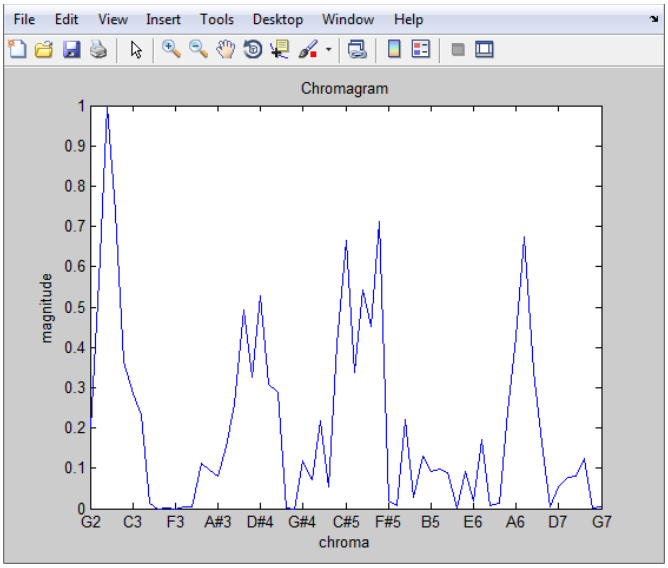

Figure 4.1 The chromagram features of audio

Figure 4.1, presents the chromagram features of audio of 'Asa Raag' from classical Punjabi folk music. The corresponding pitch class is labeled across each chroma feature. One can the amplitude corresponding the each pitch (such as C, C#, D, D#, .. A, B)[16].

4.2.2 Principal Component Analysis

Principal component analysis (PCA) is a technique which is applied for finding maximum contributing features. PCA has now been widely investigated and has been successfully applied to image recognition tasks. It is also a common technique in extracting features from data in a high dimensional space. This quality makes it an interesting tool for our study. [22]

Principal component analysis (PCA) involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. It decreases the vector's dimension with assurance that the data loss would be lowest. New orthogonal and uncorrelated vectors are made by PCA. Each basis vector is chosen so that the variance of the projection along it is maximized.

Assuming there are m observations of n variables. The data are available as a matrix X sized (m x n), where m is the number of samples and n is the number of variables. Principal components are the projections of original variables along directions determined by k eigenvectors ($k < n$) corresponding to the k largest eigenvalues of a covariant matrix X. These k principal components will determine the subspace with largest variations among all possible k dimensional subspace projected from X, while keeping to be orthonormal with other score variables. By discarding those noisy principal components that do not contribute significantly to overall variation, dimensionality can be significantly reduced.

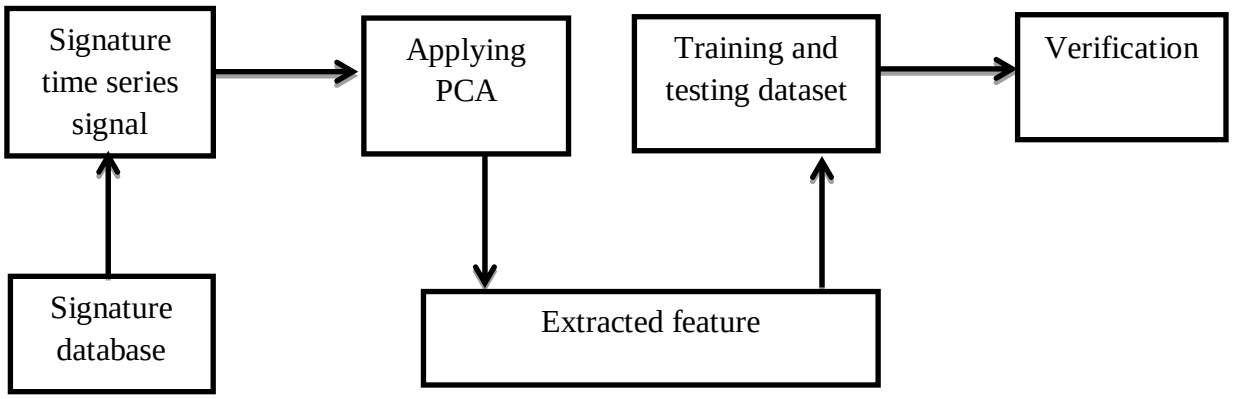

Figure 4.2 Online Verification System with PCA

However, in order to represent the feature space of each signature in a lower dimension, we consider six fundamental steps for computing PCA, before performing feature selection. The procedural steps are as follows. [15]

Step 1. Find the mean value of dataset X using (1) on each variable (x, y, p) where N is the number of available samples.

$$\overline{X}=\frac{\sum_{i=1}^{n} X_i}{N} \tag{4.4}$$

Step 2. Subtract the mean value (X) from each sample value (X) as shown in the following equation to have a new matrix (data adjust) with the same dimension, ($N * M$).

$$\Phi_i = X_i - X \tag{4.5}$$

Step 3. Compute the covariance of any two variables, (x, y), (x, p), and (y, p), separately using (3) on the previous matrix ($N * M$).

$$\text{Cov}\,(M)\,\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{N-1} \tag{4.6}$$

Step 4. Using the following equation, compute the Eigen values from covariance matrix.

$$|M-\lambda I| = 0 \tag{4.7}$$

Step 5. Also, calculate the eigenvectors from the covariance matrix using the following equation:

$$(M-\lambda_j I)\, e_j = 0 \tag{4.8}$$

Step 6. Finally, retain the largest eigenvectors K as the principal components with respect to the Eigen values.

After PCA transforms the data, the result obtained is composed of as many numbers of components as features as there are dimensions in dataset. We could reconstruct the original data by these components. The information that is not going to be explained by the components in original data is called the residual. The number of components is dependent on the value of the residual information.

The values in the score matrix are ranked based on their variance in a decreasing order, which also corresponds to the arrangement of the principal components. For instance, the first component has the highest variance value with respect to its score compared to the other two components. Likewise, the second component has the second highest variance and so on the last component has the least variance value.

Advantages of PCA

- Smaller representation of database because we only store the training images in the form of their projections on the reduced basis.
- Noise is reduced because we choose the maximum variation basis and hence features like background with small variation are automatically ignored.

4.2.3 Wavelet Transform

The wavelet transform is very much similar to windowed Fourier Transform. Wavelet transform has completely different basis function from that of Fourier Transform. The Fourier transform fragments the signal into sines and cosines. The decomposed signal is represented in Fourier space where as signal decomposed by wavelet transform is represented in both real and Fourier space. Wavelet transform is expressed by the following equation

$$F(a,b) = \int_{-\infty}^{\infty} f(x)\psi^{*}_{(a,b)}(x)dx \qquad (4.9)$$

where * is the symbol of complex conjugate, and ψ is the basis function. This function is chosen according to specific requirement.

Depending on the basis functions to be used, the wavelet transform is an infinite set of many transforms. Thus, tremendous applications of wavelet transform have been reported. There are various criteria on whose basis we can sort different types of wavelet transform, like on the basis of orthogonality, wavelets are divided in orthogonal for discrete wavelet transform and non-orthogonal for continuous wavelet transform. The properties of these categories are described below.

a) The output on using discrete wavelet transform is of the same length as of the input. Sometimes, in the resultant output vector the data values are almost equal to zero. This output vector represents the fact that the input is decomposed into wavelets those are orthogonal to its scaling and translations. Therefore, to decompose such signal same or lesser number of wavelet spectrum as in the number of signal data point are used. For signal processing and compression, this form of wavelet transform is used.
b) In contrast to discrete wavelet transform, continuous transform gives an array which is one dimension larger than the input as output. for any 1 D data, we get an image which is potted in image-frequency plane. The signal frequencies can be seen during the signal evolution and then compared with the spectra of other signals. The resultant vector values are highly correlated to each other which represent that the used transform is non-orthogonal to each other thus there is high redundancy.

Continuous wavelet transform

Continuous wavelet transform (CWT) is a form of wavelet transform using several scales and wavelets. The data obtained as the output is highly correlated and wavelets used in this case are not orthogonal. This transform can also be used for discrete time series, with a condition that the length of smallest wavelet translation should be equal to the data samples. Thus sometimes it is also called as Discrete Time Continuous Wavelet Transform (DT-CWT).

The principle of working of continuous wavelet transform is the convolution of signal directly with the scaled transform. For each and every scale, we obtain an array of the length N, same as that of signal using this technique. By using randomly chosen M

scales, we get a field of N x M dimensions in time-frequency domain. the algorithm used for this calculation is based either on the convolution by mean of multiplication or by direct convolution in Fourier space. This transform is sometimes referred to as fast wavelet transform.

The selection of wavelet to be used for a particular task is most challenging problem. Careful selection can influence the frequency and time resolution of the result. The key feature of wavelet transform is that, high frequencies have bad frequency but good time resolution whereas low frequencies have bad time and good frequency resolution. by selecting the WT carefully we can't change the key feature but as a total result total frequency of total time resolution can be increased. The result corresponds directly to the width of wavelet used in real and Fourier space. If Morlet wavelet is used, whose real part is damped cosine function, in frequency such wavelets are very well localized thus high frequency resolution is expected. In contrast to this wavelet transform, there is Derivative of Gaussian (DOG) wavelet whose frequency result is poor but time localization is good.

CHAPTER 5- RESULTS

Music signal processing and analysis is a rapidly emerging field of research, which can benefit the signal processing community by enriching it with appealing applications and novel solution.

Music is imaginably the most difficult form of art so is its processing. Data extraction and processing of music should be done carefully. For music signal processing, there is need to extract relevant information from the samples to be tested therefore, it requires the specially designed methods so that music's characteristics like pitches, rhythm, harmony, timbre, and instrumentation can be taken care of without any kind of hassle.

5.1 Back propagation Algorithm Neural Network

As described above total of 80 samples were taken for the experiment, from the available dataset 75% i.e. 64 samples were used for training and rest for test. As back propagation method is used for network training and testing, learning rate and moment constants are varied continuously along with variation in number of neurons used. The number of neurons used is varied from 2 to 9. Best result in terms of performance is achieved using 5 neurons. The values of learning rate and moment constant are: 0.1, 0.4, 0.7 and 1 each. Total of 16 combinations are formed out of the moment constant and learning rate. Thus the total performance outcomes for each case of learning rate and moment constant are compared. Performance for different values of moment constant and learning constant for all the possible data set is given in table 5.1. After calculating the performance of training, simulation output is also calculated. 25% of dataset is used for the purpose of testing. The percentage classification is done by comparing the simulation output with the required output. A table 5.2 is created using percentage simulation output for all the input data sets. The weights are saved and checked for each corresponding value of learning rate and moment constant.

Sr.no.	Value of learning rate and moment constant	Performance in Back propagation neural network							
		Raw Data	Raw+ CWT	Raw+ PCA	Raw + CWT+ PCA	Chroma-gramed Data	Chroma + CWT	Chroma + PCA	Chroma + CWT + PCA
1.	LR=0.1, MC=0.1	0	0.0742	0.1318	0.1699	0.0317	0.1380	0.1255	0.1238
2.	LR=0.1, MC=0.4	0.0232	0.0787	0.1330	0.1586	0.0231	0.1393	0.1012	0.1441
3.	LR=0.1, MC=0.7	0.0553	0.1161	0.1276	0.1750	0.0187	0.1559	0.1181	0.1503
4.	LR=0.1, MC=1	0.1972	0.3193	0.2097	0.2338	0.4086	0.2212	0.2139	0.2097
5.	LR=0.4, MC=0.1	0.0032	0.1579	0.0598	0.1422	0.0036	0.1101	0.0458	0.0862
6.	LR=0.4, MC=0.4	0.0062	0.0976	0.0831	0.1296	0.016	0.0896	0.0636	0.0383
7.	LR=0.4, MC=0.7	0.0070	0.1019	0.1028	0.1382	0.0062	0.1600	0.0951	0.0648
8.	LR=0.4, MC=1	0.1924	0.4718	0.2066	0.2361	0.7559	0.2052	0.2252	0.2127
9.	LR=0.7, MC=0.1	0.0067	0.1335	0.0640	0.1192	0.0018	0.1410	0.0176	0.0767
10.	LR=0.7, MC=0.4	0.0023	0.1668	0.0725	0.1268	0.0011	0.0513	0.0530	0.0615
11.	LR=0.7, MC=0.7	0.0023	0.0947	0.0429	0.1177	0.0015	0.1039	0.0586	0.0638
12.	LR=0.7, MC=1	0.2125	0.3641	0.1937	0.2557	0.4111	0.2720	0.2649	0.2535
13.	LR=1, MC=0.1	0.0015	0.1335	0.0293	0.1355	0.0104	0.0979	0.0275	0.0423
14.	LR=1., MC=0.4	0.0007	0.1610	0.0485	0.1106	0.0007	0.1007	0.0076	0.0422
15.	LR=1, MC=0.7	0.0024	0.1663	0.0728	0.1157	0.0010	0.1070	0.0162	0.0361
16.	LR=1, MC=1	0.2842	0.3951	0.2322	0.2249	0.3118	0.3275	0.2373	0.2131

Table 5.1 Performance in Back Propagation Neural Network

Sr.no.	Value of learning rate and moment constant	Simulation output for Back propagation neural network							
		Raw Data	Raw+ CWT	Raw+ PCA	Raw + CWT+ PCA	Chroma-gramed Data	Chroma + CWT	Chroma + PCA	Chroma + CWT + PCA
1.	LR=0.1, MC=0.1	12.50	0	31.25	18.75	50	56.25	0	37.5
2.	LR=0.1, MC=0.4	18.75	0	6.25	37.5	6.25	43.75	37.5	31.25
3.	LR=0.1, MC=0.7	31.25	25	25	37.5	43.75	43.75	37.5	18.75
4.	LR=0.1, MC=1	37.50	18.75	25	12.5	43.75	31.25	12.5	37.50
5.	LR=0.4, MC=0.1	43.75	6.25	56.25	25	37.5	25	18.75	50
6.	LR=0.4, MC=0.4	50	56.25	18.75	56.25	43.75	0	50	43.75
7.	LR=0.4, MC=0.7	50	0	31.25	43.75	25	43.75	25	31.25
8.	LR=0.4, MC=1	37.50	0	18.75	18.75	37.5	0	25	6.25
9.	LR=0.7, MC=0.1	56.25	62.5	25	68.75	12.5	18.75	25	6.25
10.	LR=0.7, MC=0.4	12.50	43.75	50	25	25	56.25	18.75	31.25
11.	LR=0.7, MC=0.7	18.75	50	18.75	31.25	37.5	56.25	37.5	37.50
12.	LR=0.7, MC=1	62.50	31.25	31.25	31.25	37.5	25	31.25	18.75
13.	LR=1, MC=0.1	25	12	50	12.5	12.5	12.5	12.5	31.25
14.	LR=1., MC=0.4	37.50	5	25	12.5	0	18.75	18.75	37.50
15.	LR=1, MC=0.7	18.75	0	25	25	12.5	25	12.5	31.25
16.	LR=1, MC=1	37.5	0	56.25	37.5	56.25	25	50	18.75

Table 5.2 Simulated Output in Back Propagation Neural Network

Using the simulated output table, table 5.2, data box plots are plotted for each data set. The total of 8 data sets are used as input for neural network the box plots are plotted keeping in mind the different values of learning rate. The box plot has 4 learning values written on their x-axis. Box plot explains the spread of simulated output. Red line in each plot tells us the mean values of the scattered data.

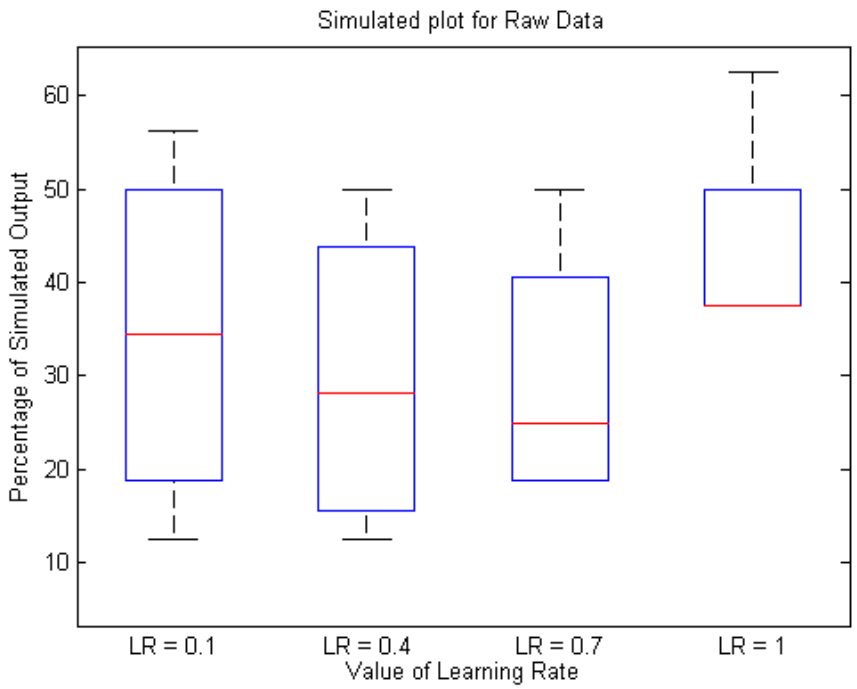

Figure 5.1 Simulated plot for Raw data

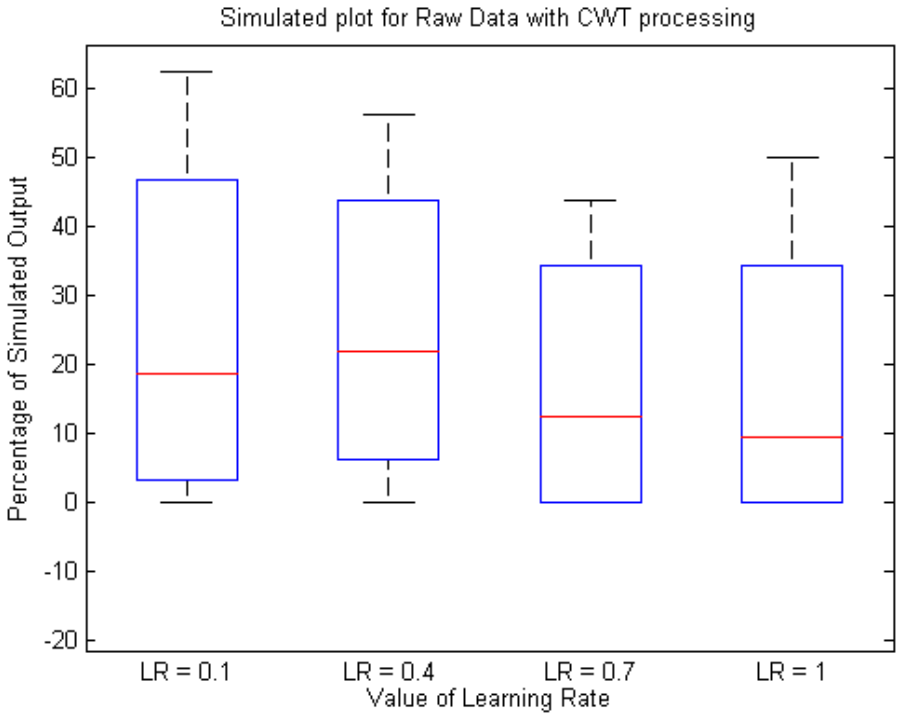

Figure 5.2 Simulated plot for Raw data with CWT Processing

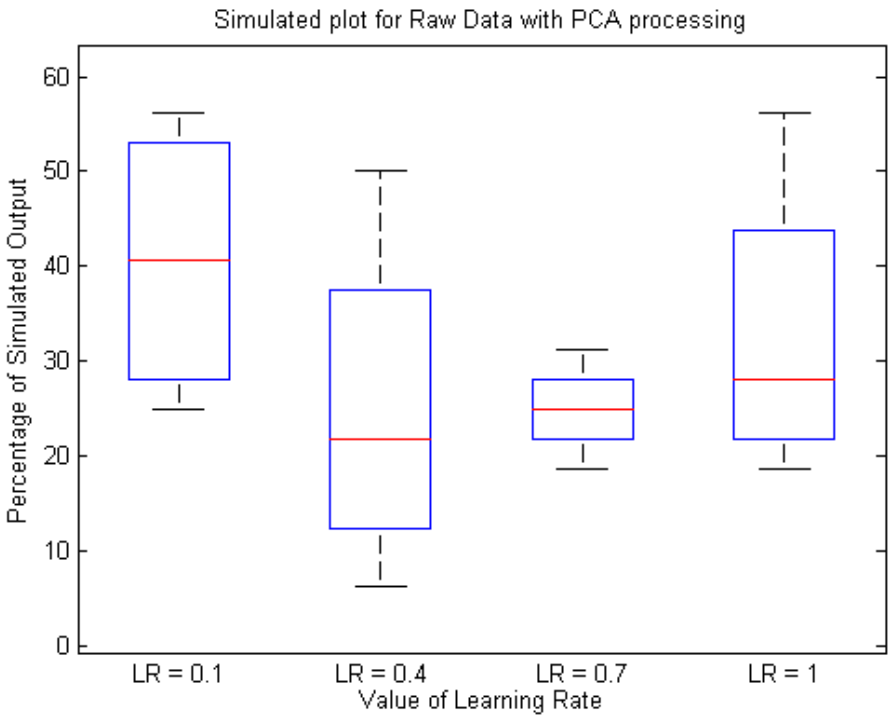

Figure 5.3 Simulated plot for Raw data with PCA processing

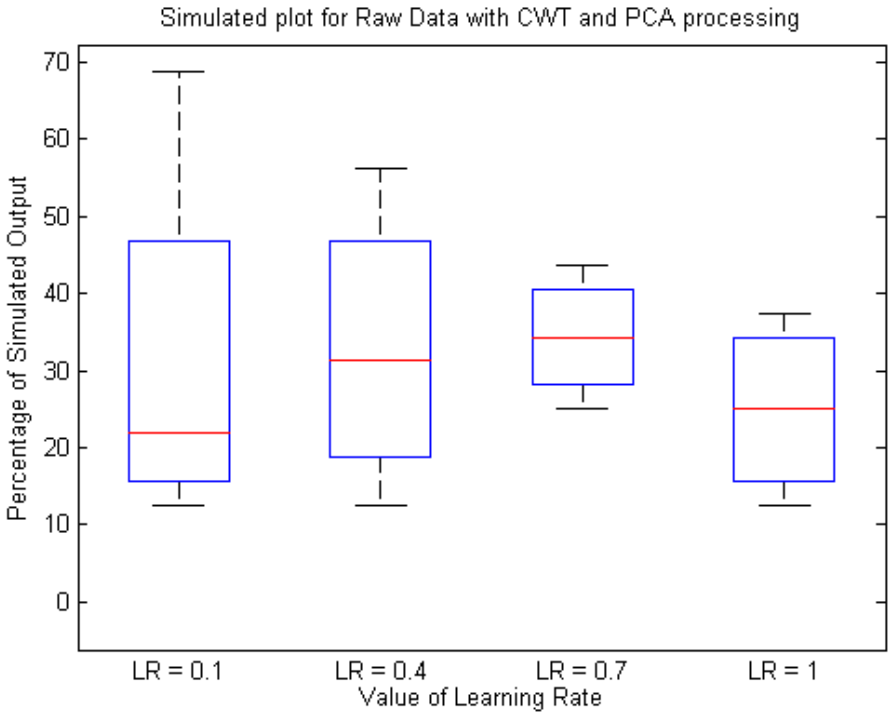

Figure 5.4 Simulated plot for Raw data with CWT and PCA processing

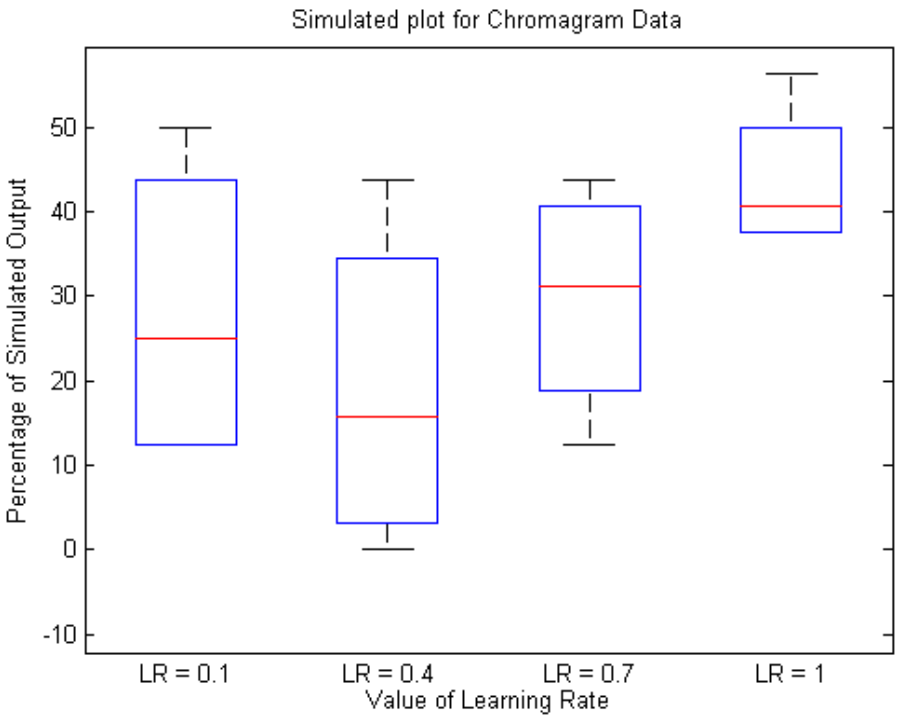

Figure 5.5 Simulated plot for Chromagramed data

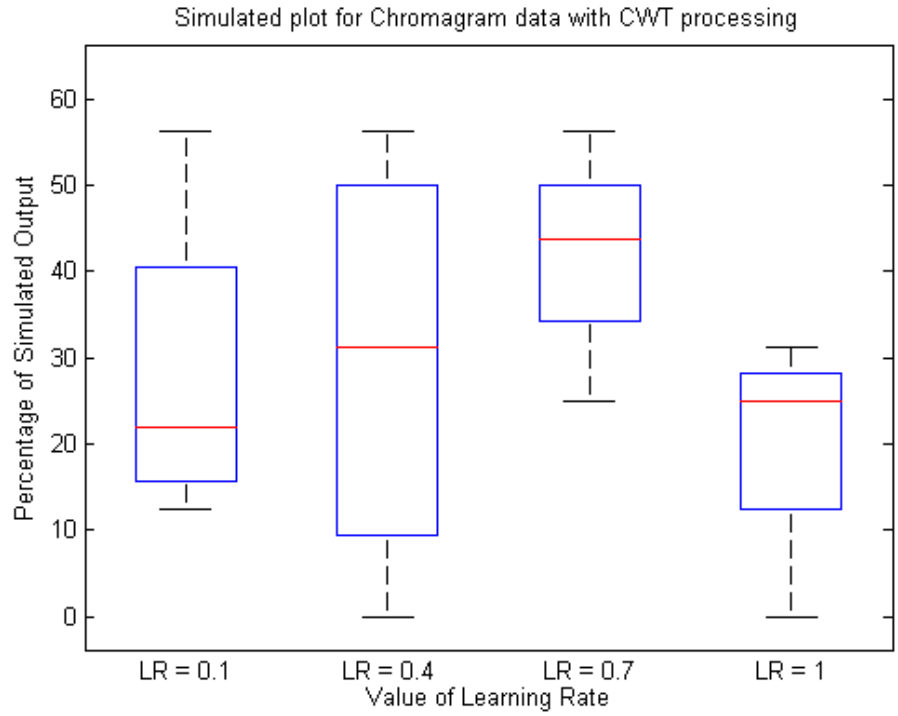

Figure 5.6 Simulated plot for Chromagram Data with CWT processing

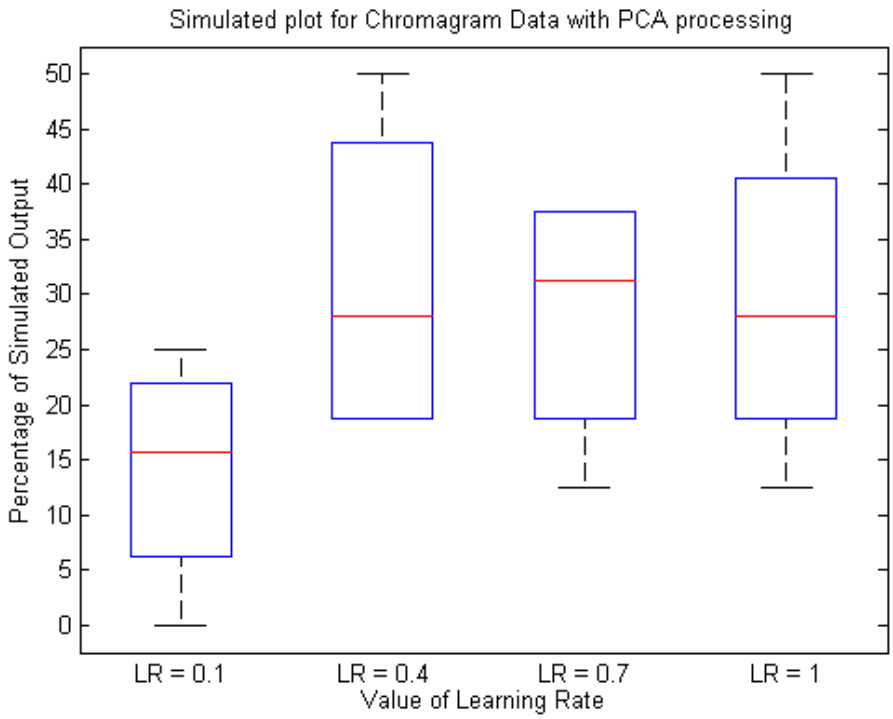

Figure 5.7 Simulated plot for Chromagram data with PCA processing

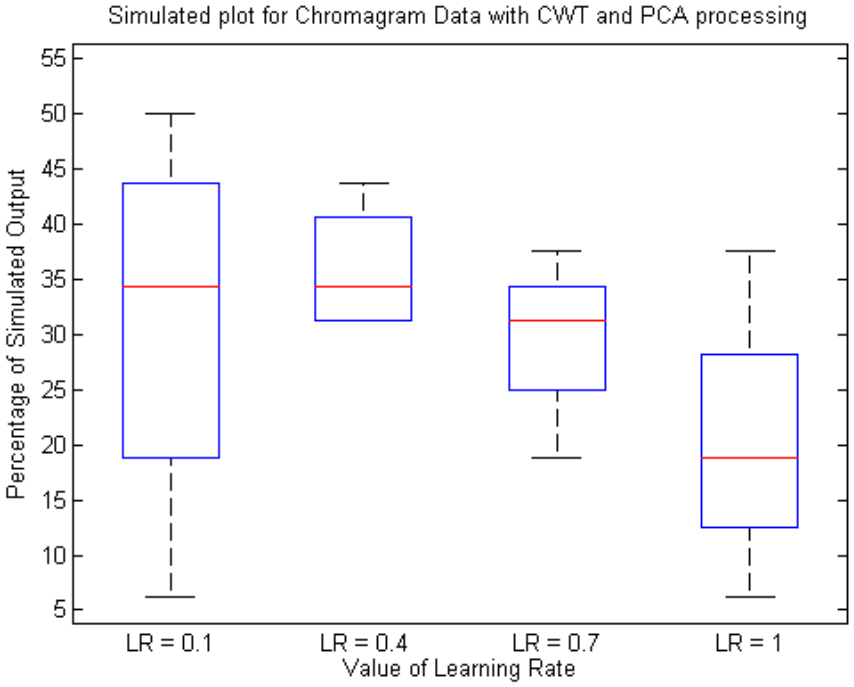

Figure 5.8 Simulated plot for Chromagram data with CWT and PCA processing

Sr.no.	Result	Result of Simulation output for Back propagation neural network							
		Raw Data	Raw+ CWT	Raw+ PCA	Raw + CWT+ PCA	Chroma-gramed Data	Chroma + CWT	Chroma + PCA	Chroma + CWT + PCA
1.	Percentage classification	34.375	21.09	31.64	30.85	30.07	37.11	25.78	31.64

Table 5.3 Result of simulated output for Back Propagation neural network

Table 5.3 gives us the result of simulated output. The maximum result is achieved for Chromagramed data with CWT processing i.e. 37.11% classification. Maximum of 37.11 percent data can be classified on the basis of chromagram data and then CWT processing applied to data.

The original chromagram data is plotted in scatter form in figure 5.9. All the data is scattered and mixed. Unless the data is plotted in different colour and shapes we can't differentiate between the different Raags. In the plot all the samples are intermixed.

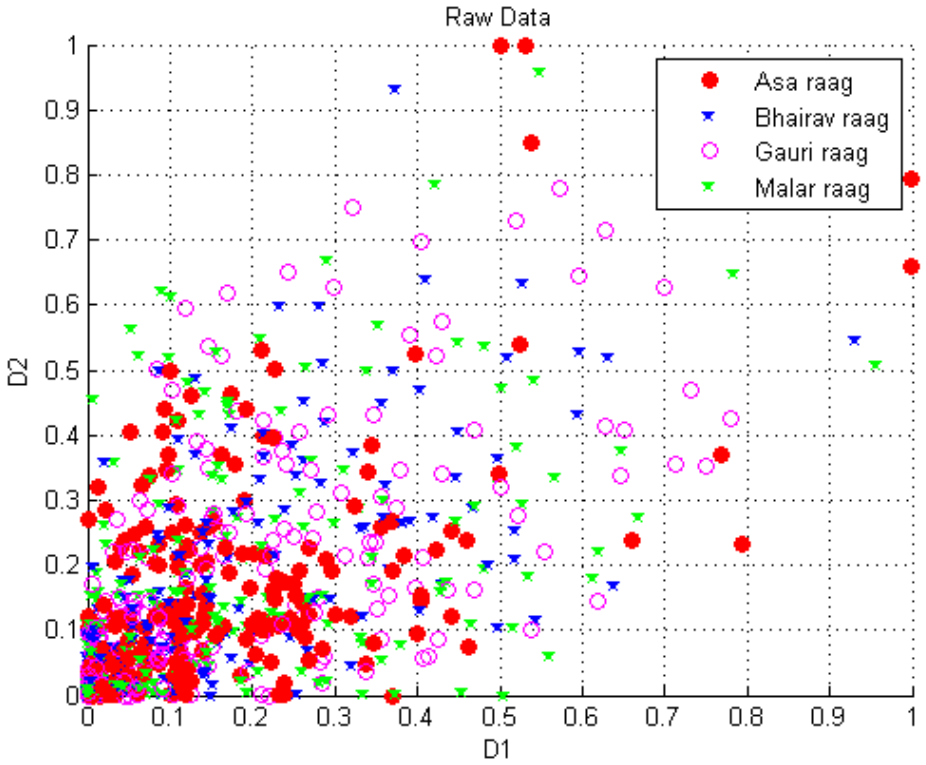

Figure 5.9 Scatter plot of Chromagramed raw data

Based on the mean values, the raw Chromagramed data is plotted again in figure 5.10. As the plot is on the basis of mean values, the plot is scattered around the mean of the data available. Thus the data is aggregated from being segregated in previous case. The output visible in figure 5.10 is more clear than the output in figure 5.9

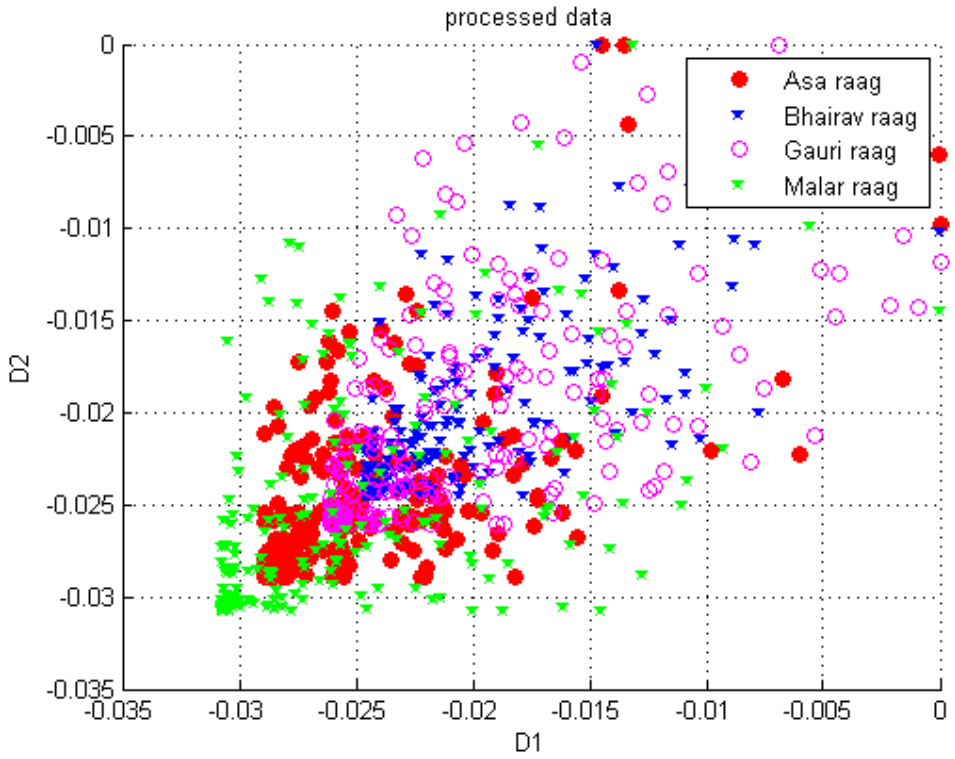

Figure 5.10 Scatter plot of Chromagramed processed data

5.2 Radial Basis Function Neural Network

As same as above total of 80 samples were taken for the experiment, from the available dataset 75% i.e. 64 samples were used for training and rest for test. As radial basis function method is used for network training and testing, spreading constant is varied continuously along with variation in number of neurons used. The number of neurons used is varied from 2 to 9. Best performance of 100% is achieved in each case of training of network. The values of spreading constant are 0.1, 0.2, 0.3., ... , 1 each. Total of 10 values are used. As the best performance in previous case was achieved with 5 neurons. Simulation output is also calculated for 5th neuron classification. 25% of dataset is used for the purpose of testing. The percentage classification is done by comparing the simulation output with the required output. A table 5.4 is created using percentage

simulation output for all the input data sets. The weights are saved and checked for each corresponding value of spread constant.

Sr.no.	Value of Spread	Simulation output for Radial Basis Function neural network							
		Raw Data	Raw+ CWT	Raw+ PCA	Raw + CWT+ PCA	Chroma-gramed Data	Chroma + CWT	Chroma + PCA	Chroma + CWT + PCA
1.	Spread= 0.1	25	25	25	25	25	56.25	25	25
2.	Spread= 0.2	50	25	25	25	25	31.25	25	25
3.	Spread= 0.3	50	25	25	25	50	31.25	25	25
4.	Spread= 0.4	50	43.75	50	43.75	50	31.25	50	31.25
5.	Spread= 0.5	56.25	25	25	31.25	50	25	25	43.75
6.	Spread= 0.6	56.25	31.25	18.75	25	31.25	62.50	18.75	43.75
7.	Spread= 0.7	43.75	37.5	18.75	18.75	56.25	50	12.5	37.5
8.	Spread= 0.8	25	31.25	12.5	18.75	56.25	25	6.25	37.5
9.	Spread= 0.9	25	37.5	12.5	31.25	56.25	31.25	12.5	31.25
10.	Spread= 1.0	25	31.25	6.25	25	7556.25	25	025	37.5

Table 5.4 Simulated Output in Radial Basis Function Neural Network

Using the simulated output table, table 5.3, data box plots are plotted for each data set. The total of 8 data sets are used as input for neural network the box plots are plotted for each of the data set. Original Raw data and its processed data's simulated output's box plots are given in figure 5.11, with x-axis the types of data used for neural network training and simulation. Chromagramed data and its processed data's box plot is plotted in figure 5.12.

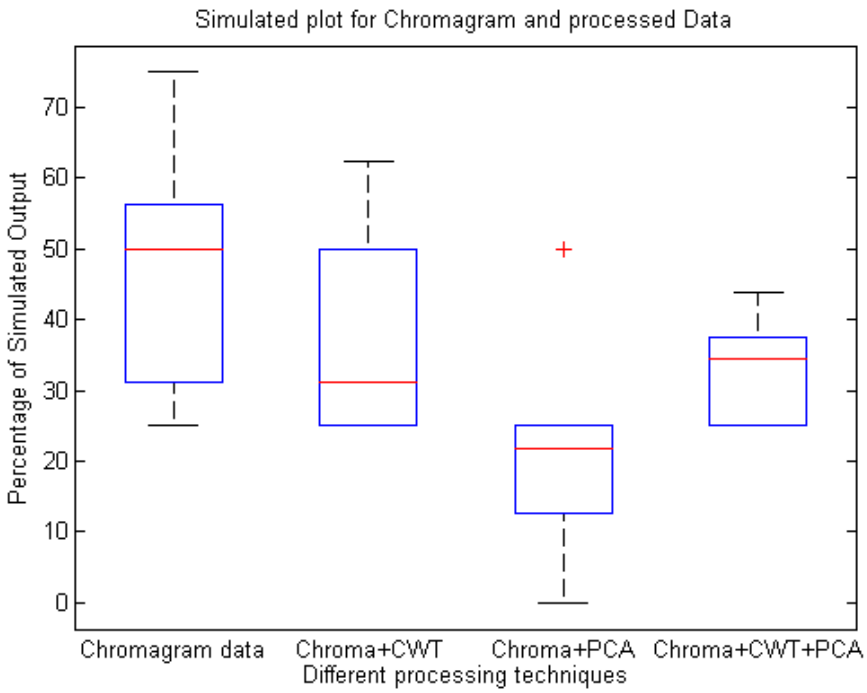

Figure 5.11 Plot for simulated Raw and raw processed data

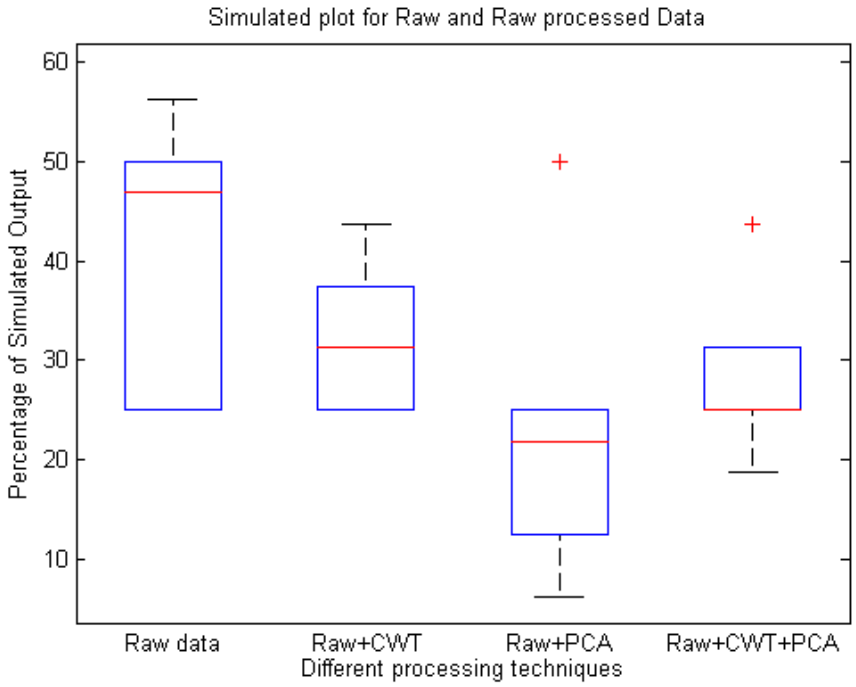

Figure 5.12 Plot for simulated chromagramed and chromagramed processed data

Sr.no.	Result	Result of Simulation output for Radial Basis function neural network							
		Raw Data	Raw+ CWT	Raw+ PCA	Raw + CWT+ PCA	Chroma-gramed Data	Chroma + CWT	Chroma + PCA	Chroma + CWT + PCA
1.	Percentage classification	40.625	30.625	21.875	26.875	47.5	36.875	20	33.75

Table 5.5 Result of Simulated Output in Radial Basis Function Neural Network

Table 5.5 gives us the result of simulated output of radial basis function neural network. The maximum result is achieved for Chromagramed data 47.5% classification. Maximum of 47.5 percent data can be classified on the basis of chromagram data applied to data.

CHAPTER 6- CONCLUSION AND FUTURE SCOPE

6.1 Conclusion- It can thus be concluded that MIR poses different sets of challenges compared to ordinary speech processing. Therefore, for the music samples to get distinguished identity in feature space, it is required that a proper set of parameters is selected.

The work presented in this report is intended to enhance the class separation of individual musical genres may be enhanced.

In the case of Back propagation Neural Network the chromagramed data set gave the maximum simulated output of 37.11%. In back propagation technique best error performance of 0.0007 is seen at raw data's training and Chromagramed data's training.

Similarly, in case of Radial basis function the Chromagramed data set gave the maximum simulated output of 47.5%. On taking deeper consideration it can be concluded that, chroma features are the best ones to classify the different raagas.

When CWT and PCA are used as pre-processing techniques CWT is the better pre-processing technique than PCA as the data classification shown with CWT pre-processing of raw data in Back propagation algorithm and RBF is 21.09% and 30.625% respectively. Whereas, in case of PCA pre-processed Chromagramed data the result is 25.78% for back propagation algorithm and 20% for RBF neural network. With the use of PCA the number of samples reduces significantly with which the result at output is not up to the mark.

6.2 Future Scope

though the techniques employed in the course of this study have yielded improved classification results, there is a lot of scope to further improve the classification performance by extracting polyspectral features from the music signals.

Furthermore, the final objective of making an automatic transcripter can only be accomplished when an unsupervised classifier, such as a gaussian mixture model is able to distinguish between individual melodies with a high success rate. This leaves a lot of scope for the future workers to try building such a system and also identify features specific to indian system of melodies.

LIST OF PUBLICATIONS

- Priya, Ravi Kumar," Identification of individual ragas using artificial neural network classifier*", Journal of Artificial Intelligence Research & Advances*.2015; Vol 2(2): pp. 1-5

REFRENCES

[1] Wasiq Khan, Ping Jiang, Rob Holton, "Word Spotting in Continuous Speech Using Wavelet Transform", *IEEE International Conference On Electro/Information Technology (EIT)*, pp. 275 - 279, 2014.

[2] Vahab Iranmanesh, S. M. S. Ahmad, W. A. W. Adnan, S. Yussof, O. Ayodeji Arigbabu, and F. L. Malallah, "Online Handwritten Signature Verification Using Neural Network Classifier Based on Principal Component Analysis," *The Scientific World Journal*, Vol. 2014, 2014

[3] Siddharth Sigtia, Simon Dixon, "Improved Music Feature Learning With Deep Neural Networks", *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP),*pp. 6959-6963,2014.

[4] Franz A. de Leon, Kirk Martinez, "Music Genre Classification Using Polyphonic Timbre Models", *Proceedings of the 19th IEEE International Conference on Digital Signal Processing,* pp. 415-420 , 2014.

[5] Bilal Hadjadji and Youcef Chibani, "Optimized Selection of Training Samples for One- Class Neural Network Classifier", *International Joint Conference on Neural Networks (IJCNN),* pp.345 – 349, 2014.

[6] Alex Alexandridis, Eva Chondrodima, Georgia Paivana, Marios Stogiannos, Elias Zois, Haralambos Sarimveis, "Music Genre Classification Using Radial Basis Function Networks and Particle Swarm Optimization", *6th IEEE Computer Science and Electronic Engineering Conference (CEEC),* pp. 35-40, 2014.

[7] Hideyuki Tachibana, Nobutaka Ono, Hirokazu Kameoka, Shigeki Sagayama, "Harmonic/Percussive Sound Separation Based on Anisotropic Smoothness of

Spectrograms", *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* Vol. 22, No. 12, pp- 2059 – 2073, December 2014.

[8] Guangzhao Bao, Zhongfu Ye, Xu Xu, and Yingyue Zhou, "A Compressed Sensing Approach to Blind Separation of Speech Mixture Based on a Two-Layer Sparsity Model", *IEEE Transactions on audio, speech, and language processing*, vol. 21, no. 5, pp. 899 - 906, 2013.

[9] Ching-Hua Chuan, Susan Vasanaand AsaiAsaithambi, "Using Wavelets and Gaussian Mixture Models for Audio Classification", *IEEE International Symposium on Multimedia*, pp. 421-426, 2012.

[10] Gopala K. Koduri, Joan Serr`a, and Xavier Serra, *"*Characterization of intonation in carnatic music by parameterizing pitch histograms*"*, *Proceedings ISMIR conference*,Porto, pp. 199-204, 2012.

[11] Joe Cheri Ross, Vinutha, T. P.and Preeti Rao, "Detecting melodic motifs from audio for Hindustani classical music", *Proceedings ISMIR conference*,Porto , pp.193-198, 2012.

[12] Meinard Müller, Daniel P. W. Ellis, Anssi, and Gaël Richard, "Signal Processing for Music Analysis", *IEEE Journal of Selected Topics in Signal Processing*, Vol. 5, Issue. 6,pp.1088– 1110, Oct,2011.

[13] Laurent Oudre, Cédric Févotte, Yves Grenier, "Probabilistic Template-Based Chord Recognition", *IEEE Transactions on Audio, Speech, and Language Processing,* Vol. 19, No. 8, pp- 2249 – 2259, November 2011.

[14] M. Mauch and S. Dixon, "Approximate note transcription for the improved identification of difficult chords", *Proceedings Conference ISMIR*, pp. 135–140, 2010.

[15] H.Papadopoulos and G. Peeters, "Simultaneous estimation of chord progression and downbeats from an audio file", *Proceedings IEEE International Conference Acoustic Speech and Signal Process (ICASSP)*, pp. 121–124, 2008.

[16] Andre Holzapfel and Yannis Stylianou, "Beat tracking using group delay based onset detection", *Proceedings ConferenceISMIR*, pp. 653-658, 2008.

[17] Khalid Youssef and Peng-Yung Woo, "Music Note Recognition Based on Neural Networks", *ICNC '08 Fourth International Conference on Natural Computation*, pp. 18-20, Oct. 2008.

[18] Olivier Lartillot, Petri, "MIR in MATLAB (ii): a toolbox for musical feature extraction from audio", *Proceedings Conference ISMIR*,pp. 127-13, 2007.

[19] Daniel P.W. Ellis, LabROSA, "Classifying music audio with timbral and chroma features", *Proceedings Conference ISMIR* ,2007.

[20] Aliaksandr Paradzinets, Oleg Kotov, Hadi, Liming Chen, "Continuous Wavelet-Like Transform Based Music Similarity Features For Intelligent Music Navigation", *International Workshop on Content-Based Multimedia Indexing*, pp. 165 - 172, 2007.

[21] Bin Li, D. Zhang and K. Wang, "Online signature verification based on null component analysis and principal component analysis", *Pattern Analysis and Applications*, Vol. 8, pp. 345-356, 2007.

[22] Z. Ye, Y. Ye and H. Mohamadian, "Biometric Identification via PCA and ICA Based Pattern Recognition", *IEEE International Conference on Control and Automation*, pp. 1600 – 1604, 2007.

[23] Slim Essid, Gaël Richard, Bertrand David, " Musical Instrument Recognition by Pairwise Classification Strategies", *IEEE Transactions On Audio, Speech, And Language Processing*, Vol. 14, No. 4, pp. 1401-1412, 2006.

[24] Douglas Turnbull and Charles Elkan, "Fast Recognition of Musical Genres Using RBF Networks", *IEEE Transactions On Knowledge And Data Engineering*, vol. 17, no. 4, pp. 580-584, 2005.

[25] Miguel Alonso, Bertrand David, Ga¨el Richard, "Tempo and beat estimation of musical signals", *Proceedings Conference ISMIR*, 2004.

[26] Software used to convert files and cut them is downloaded from, "http://download.cnet.com/Free-Convert-MP3-To-WAV/3000-2140_4-75984980.html".

[27] http://pespmc1.vub.ac.be/pos/turchap1.html

[28] J. Stephen Downie, "The Scientific Evaluation of Music Information Retrieval Systems: Foundations and Future" *Computer Music Journal* (MIT journal), Vol. 28, No. 2 , pp. 12-23, 2004,.

[29] Gianpaolo Evangelista, "Discrete Frequency Warped Wavelets: Theory and Applications", *IEEE Transactions On Signal Processing*, Vol. 46, No. 4,pp. 874-886, 1998.

[30] Peter De Gersem, Bart De Moor, Marc Moonen, "Applications of the Continuous Wavelet Transform in the Processing of Musical Signals", *13th International Conference On Digital Signal Processing Proceedings*, pp.563-566, 1997.

[31] Benyamin Matityaho, and Miriam Furst, "Neural network based model for classification of music type", *Eighteenth Convention of Electrical and Electronics Engineers in Israel*, pp- 4.3.4/1 - 4.3.4/5, March 1995.