Multi-Pitch Detection with Gender Identification and Emotion Detection

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

Master of Engineering
in
Information Security

Submitted By
Navdeep Kumar
801233010

Under the supervision of:
Mr. Ravinder Kumar
Assistant Professor
CSED

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004

Dec 2014

Certificate

I hereby certify that the work which is being presented in the thesis entitled, **"Multi-Pitch Detection with Gender Identification and Emotion Detection"** in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Information Security* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

(Navdeep Kumar)
801233010

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Mr. Ravinder Kumar)
Assistant Professor
Computer Science and Engineering Department
Thapar University
Patiala

Countersigned by

(Dr. Deepak Garg)
Head
Computer Science and Engineering Department
Thapar University
Patiala

30/12/14
(Dr. S. S. Bhatia)
Dean (Academic Affairs)
Thapar University
Patiala

Acknowledgement

It is a great privilege to express my gratitude and admiration towards my respected supervisor **Mr. Ravinder Kumar** Assistant Professor Computer Science & Engineering Department. He has been an esteemed guide and great support behind achieving this task. This work would not have been possible without the encouragement and guidance of my supervisor. His enthusiasm and optimism made this experience both rewarding and enjoyable. I am truly grateful to him for extending his total co-operation and understanding whenever I needed help and guidance from him.

I am also heartily thankful to **Dr. Deepak Garg**, Associate Professor and Head, Computer Science & Engineering Department and **Dr. Jhilik Bhattacharya**, PG coordinator, for motivation and providing uncanny guidance and support throughout the preparation of the thesis.

I would also like to thankful to entire faculty and staff members of computer science & Engineering Department for their direct and indirect help, co-operation and affection which made my stay at Thapar University memorable.

Most Importantly, I would like to thanks my parents, my brother and my friends for showing me right direction and help me stay calm in the oddest of the times and keep moving even at time when there was no hope.

Last but not least, I would like to thankful to God for his blessing, showing me the right direction, not letting me down at the time of crises and with his mercy, it has been made possible for me to reach so far.

Navdeep Kumar

(801233010)

Abstract

Pitch is a perceptive attribute of a sound. Pitch has a very important acoustical feature in speech analysis play a vital role in many applications like speech recognition, prosody analysis, speaker identification, emotion recognition and computational auditory scene analysis (CASA). Pitch perception is very complex process. Pitch determination of single source is easy as compare to multi-Sources or of polyphonic signals. The problem of estimating the fundamental frequency or pitch of periodic waveforms occurs in various form of application, and has received notable interest over the recent years, for example, several speech and audio problems notably depend on the initial forming of an estimate on the pitch or pitches including problems. Fundamental frequency plays a vital role in Gender identification also. There are many effective and accurate proposed algorithms on single source pitch determination and detection. But the multi-pitch real life scenario occurs regularly than single pitch case, and often also in speech processing. It is difficult to accurately estimate the multiple pitches of a mixed signal; multi-pitch estimation has potential application in speech sources separation, speech enhancement, speech recognition.

We study the problem of estimating the fundamental frequencies of a signal containing multiple harmonically related sinusoidal signals along with this emotion detection is also taken into consideration. We propose a system which is used to detect single as well as multi-F0 and detect emotions from signals. Multi-pitch system performance is evaluated on 135 mixture audio signals which compromise of human speech of different genders and emotion detection system is evaluated on 150 single person audio signals. We applied our system on pitch application field in gender identification. Our system evaluates number of human in complex mixture speech signal and their respective genders. It also evaluates emotion of audio signals. Multi pitch system is based on iterative approach in which frame by frame analysis is taken in order to the detect multi-F0. Effective implementation of algorithm and its high accuracy rate helps in classifying gender identification and person calculation. On the other hand Emotion Detection System is based on GTCC feature extraction technique which helps in recognizing emotions with high accuracy. Use of Genetic algorithm plays a vital role in optimizing the results. This optimization leads to high efficiency and better performance of the algorithm

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1. Natural Language Processing (NLP)

Natural processing language mainly deals with concept of interaction between human and computer. Computer science, artificial intelligence and linguistics are used in NLP in order to make the interaction possible between computer and human (natural) languages. Main hurdle which is faced in NLP is enabling the computer to understand and derive the meaning from human natural language or any other natural language generation. There are various researched tasks in NLP. Some of them are used as sub-tasks for solving larger tasks, while others have direct real-world applications. These tasks differ from actual NLP tasks in terms of various parameters such as volume of research devoted, well-defined problem setting, standard metric for evaluating the task and competition to each task. Some of the tasks are listed below:

- Automatic summarization
- Co-reference resolution
- Discourse analysis
- Machine translation
- Morphological segmentation
- Named entity recognition (NER)
- Natural language generation
- Natural language understanding
- Optical character recognition (OCR)
- Part-of-speech tagging
- Parsing
- Question answering
- Relationship extraction Sentence breaking
- Sentiment analysis
- Speech recognition
- Speech segmentation
- Topic segmentation and recognition

1.2. Signal

This term is mainly used in communication systems, signal processing, and electrical engineering. It is defines as a function that provides information about various phenomenon's, their behaviour and attributes. A message conveyed among observers, that provide information about the physical system in terms of variation in time or variation in space (such as an image) is potentially a signal.

1.3. Speech

Syntactic combination of lexical and names that are drawn from very large number of words is known as speech. It is basically the vocalized form of human language. Phonetic combination of a vowel and consonant of speech creates the spoken words.

1.4. Speech processing

Speech processing is the approach of studying speech signals and learns the various methods of processing these signals. Speech processing is a vast approach, including various areas of study as follows:

Speech Recognition: In this area, voice is recognized. This can be done by analyzing the linguistics of speech signal. After this, they are converted into a computer-readable format.

Speaker Recognition: Target of this method is to identify the speaker of speech signal.

Speech Coding: This study is mainly beneficial in telecommunication. It is a way of doing data compression.

Voice Analysis: This area mainly helps in medical treatments. Functioning of vocal cords and analysis of vocal loading can be done through this method.

Speech synthesis: Various ways of making computer-generated speech are described in this. It is also referred as artificial synthesis of speech. This technology enhances the capability of using computer in this field.

Speech enhancement: This area deals with removing noise from speech signals which further leads to high intelligibility of a speech signal.

Speech compression: It deals with various ways to increase the amount of information which can be transferred, stored, or heard. This is done for a limited period of time and space as defined initially or predefined.

Speaker diarization: It is a process of dividing the audio input stream into homogeneous segments depending upon speaker identity.

1.5. Voice

Every human being has a distinct voice which can be used to identify the person. This identification is similar with the concept of fingerprints. Voice mainly refers to the sound made by human being for talking, screaming, laughing etc. These sounds are produced using the vocal folds (Vocal Cords) which are the primary sound source. Various components at different levels of multitude are mixed together for composition of human voice. Due to this voice of each human being is different and unique. These components are named as pitch, tone, and rate. The rate at which the vocal cords vibrate is referred as pitch. Higher the number of vibration per second, higher will be the pitch which results in high sound of voice.

1.6. Generating human voice

Human voice is generated in three main parts; the lungs, the vocal folds within the larynx, and the articulators. The lungs are used to produce the required air pressure or air flow for vibrating the vocal cords. This air pressure then flows towards vocal cords where it is converted into pulses that are audible. These pulses form the laryngeal sound source. The muscles of the larynx came into action and set the length and tension to 'fine tune' pitch and tone. Tongue, palate, cheek and lips combine form the articulators. It lies above the larynx. It articulate and filter the sound exhaled from the larynx and interaction with the laryngeal airflow is done in order to strengthen it or weaken it as a sound source.

Vocal cords and articulators (in combination) can produce complex arrays of sound. The tone of voice changes in order to show different emotions like anger, happiness.
Adult men and women can be differentiated on the basis of various parameters in case of voice. Adult males have large vocal cords as compare to adult women. Voice of adult men's has usually low pitch as compare to women. Length of vocal cords of

male lies between 17 mm and 25 mm and those of female lies between 12.5 mm and 17.5 mm in length. Due to this difference in vocal cord size, the pitch of voice of men and women is different. There are many other ways to differentiate between male and female voice. Men's usually have large vocal tract due to which the men's have a lower-sounding timbre in their voice. This is not dependent on vocal cords. Figure below shows the various different parts of the vocal folds and cords.

.

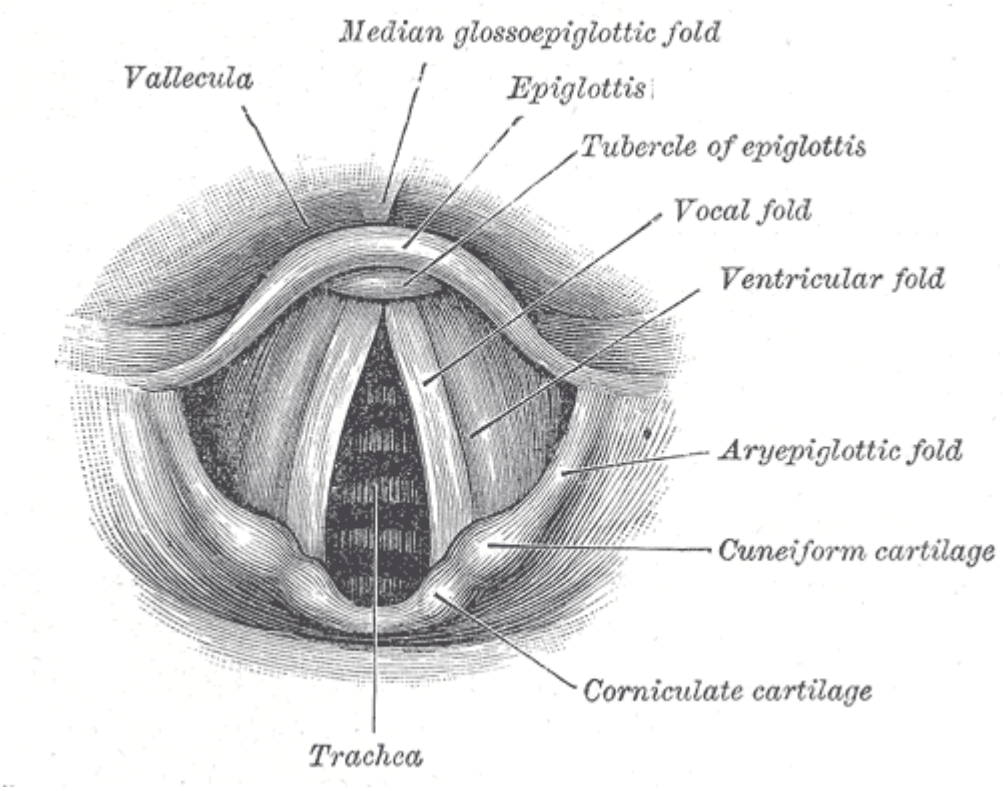

Figure 1.1: anatomical diagram of vocal cord [1].

Vocal cords in both sexes lie within the larynx. Vocal cord is connected to the arytenoids cartilages (near spinal code) from back and to the thyroid cartilage (under the chin) from front. They do not have any outer edge and the inner edges are free to vibrate. Vocal cord can be shorten and bulge due to an epithelium, vocal ligament and vocal is muscle. Vocal cord is covered with vestibular fold or false vocal cord from both above sides.

The ability of every person to dynamically modulate certain parameters of the laryngeal voice is considered is taken into account by human spoken languages. The main parameters considered for differentiating between the voice of men and women

is voice pitch and the degree of separation of the vocal cords. If vocal cords are coming together it is referred as adduction and if separating then it is referred as abduction.

Humans can hear the sound from air if the frequency of vibrations lies between 20 Hz and 20,000 Hz. It is normally believed that the fundamental frequency of male and female lies in midrange but it is not an accurate answer. But is reality, the frequency of male and female is much lower than the mid range. Average fundamental frequency of men lies between 85 Hz and 155 Hz. A women's voice ranges from 165 Hz to 255 Hz. In case of children this range varies from 250 Hz to 300 Hz and higher.

1.7. Essential Terminology

Acoustics: In this field of science, the mechanical waves in solids, liquids and gases are studied. Along with mechanical waves the vibration, sound, ultra-sound and infrasound are also taken into consideration.

Acoustic phonetics: It deals acoustic aspects of speech signals. Various properties of waveforms such as mean squared amplitude of a waveform, its duration, its fundamental frequency are analyzed in this. It also deals with relation between these properties and other branches of phonetics such as auditory phonetics. This comparison helps in abstracting phones, phrases, or utterances.

Multi-pitch analysis: In this the fundamental frequencies (F0s) of polyphonic audio are analysed. The frequency and number of pitches in each time frame is estimated during multi-pitch analysis. After this estimation, the pitches are organized according to sources. This is one of the most challenging tasks in the field of speech signals.

Formants: Formats are measured as the point of highest amplitude in the frequency spectrum of the sound. This measure g is done using spectrogram this concept is defined by Gunnar Fant as "the spectral peaks of the sound spectrum of the voice". It can also be used to find mean of acoustic resonance of human vocal tract.

Fundamental frequency: The fundamental frequency of voiced speech of adult male lies between 85 to 180 Hz and that of women lies between 165 to 255 Hz. Therefore most of the speeches have their fundamental frequencies below the voice frequency

band. The ranges of signals which are heard by human beings have the frequency range from 20 Hz to 20 kHz.

The signals mainly lie in either time domain or frequency domain. Change of signal with respect to time is shown in time domain. Signal is presented by waveform in case of time domain.

On the other hand n the frequency domain the signal is represented by a spectrum. Performing measurements in time domain is easy but very less information is presented in time domain. In order to get the information such as: *spectral* information, frequency content and behaviour of the audio signals and of complete systems, the signal must be changed to frequency-domain. There are various methods to convert the signal from time domain to frequency domain. Some of them are as follows:

1. Fourier transform→ no repetitive signals, transients
2. Z-transform→discrete signals, digital signal processing
3. Hilbert transform

Discrete Fourier Transform: Discreate Fourier Transform (DFT) of ther length –N sequence x[n] is defined by:

$X[k]=\sum_{n=0}^{N-1} x[n]\, e^{-j2\pi kn/N}$, $0 \leq k \leq N-1$ ..... [2]

Sampling Rate(SR): It is defines as number of samples per second. It is also known as sampling frequency. Measuring unit of SR is Hz. It sets the maximum frequency that can be analysed. There is a relation between highest frequency analyzed in computer (Nyquist limit frequency (f_{max})) and the sampling rate. The relation is defined as follows:

f_{max}= sampling Rate/2

Sampling period (T): Sampling period is inversely related to sampling rate. Sampling period is defined as length of time (in seconds) between samples. Time resolution is set to T, which means no details can be gathered from time signal whose duration lies below this value (T).

Word Length: It defines the number of bits that are used by sound card in analog–to-digital and digital-to-analog convertors. It varies according to hardware. It is

advisable to use the maximum values compatible with the hardware. In figure 1.2 sampling rate of a signal is defined and derived.

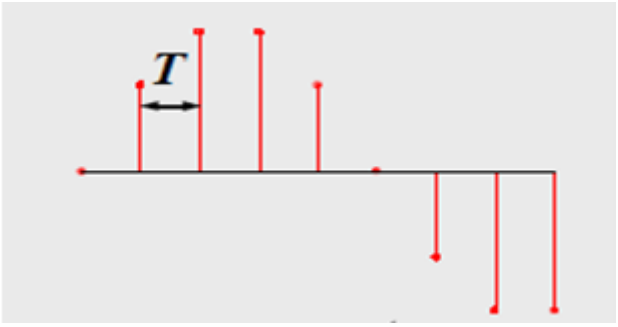

Figure 1.2: Defination of Sampling Rate [3].

Sampling rate=1/(Sampling Time)

Measuring unit of Sampling rate is samples/sec and Sampling time is sec/samples.

Frequency Domain: It is referred as analysis of mathematical signals (Functions) with respect to frequency. Signal in time and frequency domain is quite different. How a signal changes over time is shown in time-domain graph. On the other hand how the signal lies over a range of frequencies is described in frequency-domain graph.Information about phase shift that must be applied to each sinusoid is also present in frequency-domain representation.This information is essential for recombining the frequency components so that the original time signal can be recovered.

A signal can be converted from time to frequency domain and vice-versa through various mathematical operators called a transform as discussed above. The inverse Fourier transform perform the task opposite to Fourier transform that is convert the frequency domain f back to a time domain. In order to visualize real signals in frequency domain, spectrum analyzer is used.

1.8. FFT Analysis: Viewing Frequency Domain Information

As discussed above, Fourier Transform is used for converting time domain into frequency domain. Frequency domain represents both the amplitude and phase of the sinusoidal components that are essential part of the signal.

According to Fourier theory, any complex time signal is composed of various sinusoidal waves. These waves have changing frequency, amplitude, and phase. This concept is used for transforming signal from time to frequency domain and vice-versa. To convert a continuous signal x(t) to its frequency domain counterpart X(jω), the forward Fourier Transform can be used. FFT is described as follows:

$$X(j\omega)=\int_{-\infty}^{+\infty} x(t)\, e^{-j\omega t} dt. \quad \text{.....[3]}$$

Complete time history of a signal is required by the Fourier Transform to describe the signal. Along with this an infinite number of sinusoidal frequency components are also needed for full fledged description of signal. Initially this is of no use because signal can be observed for a finite amount of time. So, time windowing must be used to limit the view of signal to a finite time frame. After this the Fourier Transform can be implemented on signal. The figure 1.3 shows that the signal conversion from time domain to frequency domain has been done using the Fourier transformation and inverse Fourier transformation method has been used to convert the signal from frequency domain to time domain.

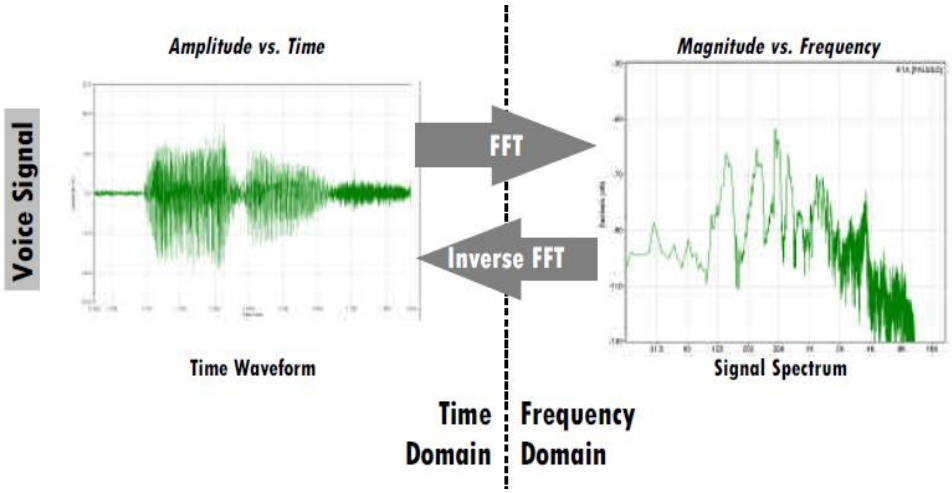

Figure 1.3: Fourier transformation: signal conversion from time domain to frequency domain [3].

Discrete Fourier Transform (DFT) can be used for this purpose as it deals with sampled-data signals. Fast Fourier Transform (FFT), is further used for accelerating the computational power of DFT. Input to FFT is finite-length blocks of sampled data. These blocks are known as FFT frames. Length of these FFT data frames is represented by NFFT. The length of FFT frame can be easily computed in units of time using the sampling rate.

1.9. Emotion Detection System

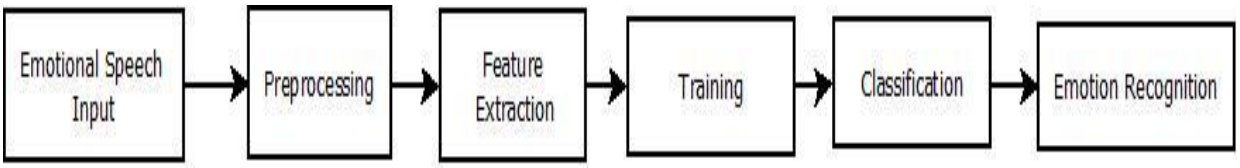

Figure 1.4: Block Diagram of Emotion Detection System

Combination of various stages forms an emotion recognition system as shown in above figure. The first basic step of this system contains the task of collecting the speech samples which contain emotions. These samples play a vital role in creating the database of speech .This database is further utilize as an input to the emotion recognition system.

After the first step, the second phase is pre-processing of the input signal. The pre-processing of speech signal is done so that the corresponding emotions in it can be identified. This phase mainly focuses on improving the quality of input signal. This signal is then further used to collect the various features. After this system needs to use the classifier so that the speech signals can be classify according to the emotions. These extracted features are further used to provide training to the classifier. This training results in a system which is familiar with various emotions and their features. After this training now when a speech signal is taken as input, this trained system can properly classify it.

1.10. Gammatone Cepstral Coefficient

In order to model the filter responses of human auditory, a special function is used named as Gammatone function. This function is named on an impulse response which

is the results of a gamma distribution and a sinusoidal tone centered at the frequency, being computed as:

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

Where:

$g(t)$ represents the Impulse response of Gammatone filter.

K represents amplitude factor.

n represents filter order.

f_c represents central frequency and units are hertz.

ϕ shows the phase shift.

B is the duration of impulse response.

Results of comparison between the impulse response given by the Gammatone filter and by the human, shows that there is a convergence between the frequencies measured by human beings and the frequencies of the cochlea. Following are some reasons which explain this convergence:

a) There is very much similarity between the magnitude response of reox function and the GT filter.
b) The bandwidth of filter is corresponding to the fixed distance on the basilar membrane.

Chapter 2
Literature Review

Recent work has aimed instead forming the pitch estimate using second order statistics [4] [7]. For multi-F0 source signals, containing several harmonically related signals, these method estimate each of the present pitch signals separately forming different forms of iterative estimation schemes, typically requiring a priori knowledge of both the number of sources and the model order of each of the sources. In this work, a novel method is proposed which estimates the fundamental frequencies of a signal with multiple pitches, without assigning any prior knowledge of either the number of sources present or their number of harmonics.

In 1964, A. M. Noll [8] introduced the cepstral analyser. This has been through short time power spectra and band-pass filter bank process. This cepstral is used to determine the fundamental frequency and determine the voiced-unvoiced signals in audio wave. It has been designed on IBM digital computer. However, he in 1966 [9] introduced the advanced technique of computing power spectrum of log of power spectrum to obtain the better peak. The estimation of pitch was much better than the previous experiment. This experiment has been performed at Bells Laboratory and cepstrum of the signal was calculated and automatically gets plotted on microfilm. In 1976, L R Rabiner [10] performed analysis of eight pitch detection algorithms. This experiment has been performed on considerable size of speech corpus. The errors are also measured and relative deviation is being observed for the same.

In 1999, M. Karjalainen *et al.* [11] present a more efficient way of analysing the multi-pitch and periodicity of complex signals. This model is much more efficient and practical than the Meddis and O'Mard's model of pitch perception model. Although this proposed model is more efficient but still it contains various features of previous model. This model describes the way of separating sources of complex audio signals. This model is mainly useful when transcription of music is need to be represented automatically and when representation of audio signals is to be done in a structural manner. In the previous approach Meddis-O'Mard used the gamma tone filter bank for simulating the selectivity of frequency during peripheral hearing. After this

simulation splitting of signal into channels such as ERB (equivalent rectangular bandwidth) is done. These channels are rectified with half wave and filtered by low pass, so that the activity of the hair cells can be simulated. After this autocorrelation function (ACF) of each channel is calculated to extract signal periodicity. At lasts the summation of all the ACFs results in summary autocorrelation. Over all periodicity properties of the incoming signal is shown by SACF [12] [13]. Details of the unitary multi-channel pitch analysis model have been shown in figure 2.1.

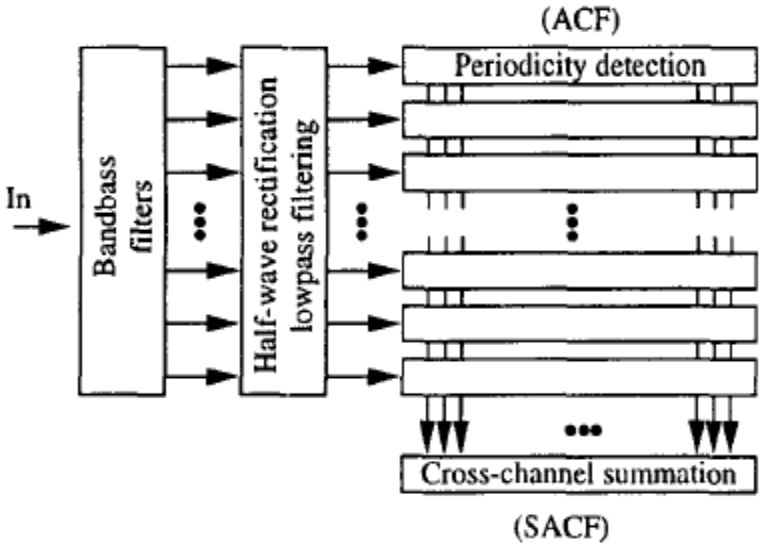

Figure 2.1: A block diagram of the unitary multi-channel pitch analysis model [11].

In the proposed simplified pitch analysis model, the Meddis-O'Mard model is represented by model's middle part (x2→x3). The lower part of this model represents the extension proposed by new model. In this extension, the main concept is that the audio frequency range is divided into two sub channels. Different methods are followed for low and high frequencies. Frequencies below 1 Hz are analyzed by autocorrelation while those above 1 kHz are taken through three steps. First they are rectified and then they are low-pass filtered. At the end there autocorrelation is calculated. Comparison between these two approaches shows that this two-channel analyzer is much more computationally efficient than a multi-channel pitch analyser. The figure 2.2 shows the analysis of simplified pitch model proposed by T. Tolonen and M. Karjalainen.

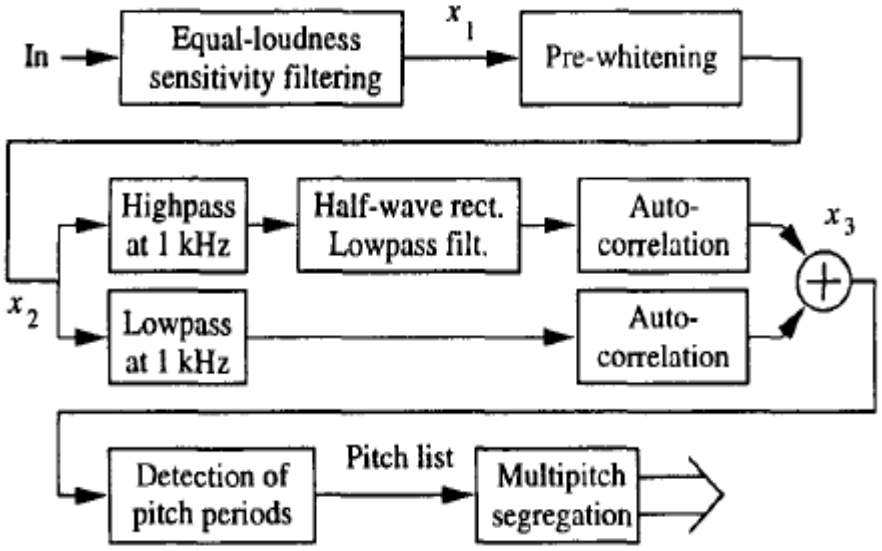

Figure 2.2: A block diagram of the simplified pitch analysis model [11].

Limitation of this approach as compare to human perception is in the case where relative amplitude of sub-signals differs substantially. Another point mentioned by author is that this proposed algorithm can be compared in a more efficient way in aspect of temporal resolution of pitch detection.

In 1999, P. J. Wulmsley *et al.* [14] described an approach for pitch estimation and note detection. Along with this the number of sounding notes happening concurrently and the number of harmonics is determined. This approach is mainly implemented for polyphonic audio signals. Bayesian probabilistic framework and harmonic signal model is used in this approach due to which prior knowledge about the nature of musical data can be determined. Frequency variation over time is explicitly modelled using latent variables. This modelling is done to get the high correlation between model parameters during adjacent frames of data. Joint estimation of these Parameters is done so that the robustness of estimation can be increased against transient events. Markov chain Monte Carlo (MCMC) methods are used for performing this estimation. Individual frames of data are modelled as the sum of harmonic sinusoids. In this whole concept the data is considered on frame-by-frame basis. The main point which is exploited in this approach is that those frequencies are considered or are useful which are longer than a single frame. Main advantage of this approach is that the incidence of spurious frequency from transient events is greatly reduced.

In 1999, W.W. ZHAO *et al.* [15] proposed a new technique for formant and pitch estimation using Wigner-Ville distribution of a multi-component signal. This method results in more accurate, easy to recognize and better resolution formants and pitch estimation as compared to other methods. For estimating only formant, one pitch-period segment is sufficient while on the other hand minimum two pitch-period segment is required if both the pitch and formant detection need to be done. Proposed method is compared with various other methods such as spectrum method. The spectra of the vowels "e" and "a" are combined with their time-averaged WVD plots As the amplitude peaks produce by spectrum method are not related to formants therefore which ones corresponds to these frequencies are not easy to determined. On the other hand, in proposed method all the produced amplitude peaks occur at the formant frequencies and thus they can be determined easily. In the previous approach the vowels whose formant frequencies are close to each other, it is difficult to find their locations and to distinguish their differences. As two narrow amplitude peaks are produced by new proposed approach, it is easy to find the locations of these frequencies. Results shows that the formants determined from the WVD method and from spectrum method are quite same, but this estimation is easier in WVD method.

In 2000, R. Janc *et al.* [16] consider the fundamental frequency of snores and studied them with signal processing techniques. These frequencies are classified based on the site of obstruction in the upper airway. This classification is helpful in diagnosis of various diseases such as Multiple System Atrophy in the early stage. Pitch is the time domain part of frequencies and it can be defined for snores in the same manner as it is done for speech vowels. Most of the pitch detection algorithm is composed of three main blocks: a pre-processor; a basic extractor; and a post-processor. Second block estimates the pitch values and the third block do error correction and graphical presentation. A method named as autocorrelation detector with centre clipping of speech detection is selected for implementing it on snores. This method can measure the pitch presence and absence in the snores. The detector scheme is shown in Fig.2.3.

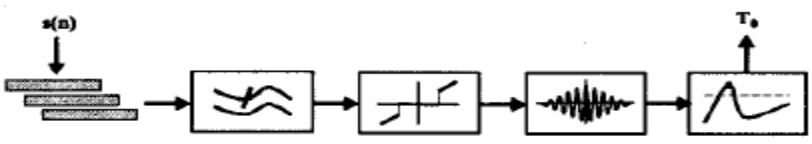

Figure 2.3: Scheme of the AC pitch detector with centre clipping [16].

Signal segmentation, low-pass filtering (LPF), centre clipping and autocorrelation is done in first block which is pre-processor. Peak detector acts as the basic extractor. Pitch samples are interpolated without error correction in third and last block.

Another method proposed for evaluating the performance of the automatic detector is manual pitch estimation. Simple snores and snores of QSAS patients are taken as input for these two methods. A smoothed version of the manual pitch estimation is produced by first method as it follows slow pitch variations. Three other parameters named as pitch mean value, pitch standard deviation and pitch density are also defined in order to analyze the pitch evolution into a snore. Both automatic and manual algorithms show similar values of these parameters. There is difference between the value of pitch density of n simple snorers and OSAS patients. These differences need to be checked again on much bigger database.

In 2001, A. P Klapuri [17] proposed a new model after implementation of three models for multi pitch detection in speech signals. This work contains spectral smoothness evaluation. The speech corpus being used varies from one to six speeches in an audio wave. The error rate reduces as we reduce the number of sounds in one audio wave.

In 2003, M Wu *et al* [18] used HMM for detection of pitch tracks being framed. The robust algorithm is proposed for speech recognition algorithm for noisy speech. A. P Klapuri [19] in his next paper performed another experiment in which he calculates harmonicity and spectral smoothness. This is repetitive process in which one sound is being detected first. This sound is then removed from complex mixture of sounds in signal. Now, the residual signal is undergone through same process.

In 2003, A. P. Klapuri *et al.* [19] proposed a new technique that estimates the fundamental frequencies of concurrent musical sounds. This algorithm works well in the presence of harmonic and noise sounds. This new method performs the multiple-F0 analysis for sound sources of diverse kinds. This technique works in iterations. In every iteration, most prominent sound is taken and its fundamental frequency is estimated. This stage of estimating the F0 of most prominent sound is known as predominant- F0 estimation. As this estimation is done in the presence of harmonic sounds, so the harmonic frequency relationships of concurrent spectral components are taken in order to add them to the sound sources. An algorithm is proposed which

can handle these inharmonic sounds. After this, the sound considered in first step is removed from the mixture and the remaining signal is considered for next iteration. An algorithm utilizing the relationship of concurrent spectral components is taken into account for estimation. This algorithm tells that the amplitude of a harmonic partial is usually close to the amplitudes of the nearby partials of the same sound. This relationship is in terms of frequency and it does not assume ideal harmonicity .On the other hand, for performing subtraction the spectral smoothness principle is used. Using these techniques estimation of multiple fundamental frequencies is done in a more accurate manner. These techniques enable to do this in a single time frame. Even no long-term temporal features are used in this estimation process. Samples are recorded from 30 musical instruments and four different sources are taken for collecting input for the system. Sound source and pitch is combined randomly to perform multiple fundamental frequency estimation. Mixtures containing one to six simultaneous sounds have different error rates described as 1.8%, 3.9%, 6.3%, 9.9%, 14%, and 18%, respectively. The proposed algorithm performed much better than the average of ten trained musicians in case of musical interval and chord identification. The proposed method works well in noise and is capable of handling sounds exhibiting in harmonicities. This algorithm can be applied in automatic transcription of music, implying wide pitch range, tone colours variations, and in conditions where there is need for robustness as noise is present around. The proposed algorithm works well in case of rich polyphonies, corrupted signals due to high additive noise or in case where the frequency bands are not present. Limitation of this proposed approach is that a long analysis frame is needed as compare to other methods. Long analysis frame is required so that the algorithm works well in case of low-pitched sounds as the execution is done in frequency domain which requires fine frequency resolution.

In 2004 , multiple pitch estimation using the bispectrum of the audio signal is proposed by S.S. Abeysekra *et al.* [20].Two dimensional distribution is consider for bipsectrum. This bispectrum is the Fourier transform of the third-order cumulant of the signal. This computed bi-spectrum is used for estimating pitch.2-dimensional frequency-lag domain based pitch separation is more flexible as filtering is done in the frequency lag domain. Power spectral density (PSD) based method is used for calculating the autocorrelation. Assumption about the position of harmonic components that are uncorrelated is done. Distribution of power between these

frequencies is also assumed. Although the estimation of pitch is done by using the linear mechanism governing the process but still the there is no accounting done for the phase relations between frequencies. It is known that the information about phase coupling can be extracted from phase relations and this information can help in improving the accuracy of estimates. As explained above after estimation, those pitch components are subtracted from the frequency lag distribution and the residual is considered for performing this estimation recursively. This proposed approach performs this task comparatively in a easy manner due to the use of two dimensions. Bayesian model is used in this estimation technique as in this model fundamental frequency, harmonics and amplitude is used to describe each component signal. Bayesian model is modified by Simon Godsill [21]. After this modification non-white residual spectrum along with time varying amplitudes is included in this model. This model also contains accounting for detuned partials due to inharmonicity. This means that there is no strict linear relationship in frequency for harmonics. Block diagram of bispectrum shown in Figure 2.4

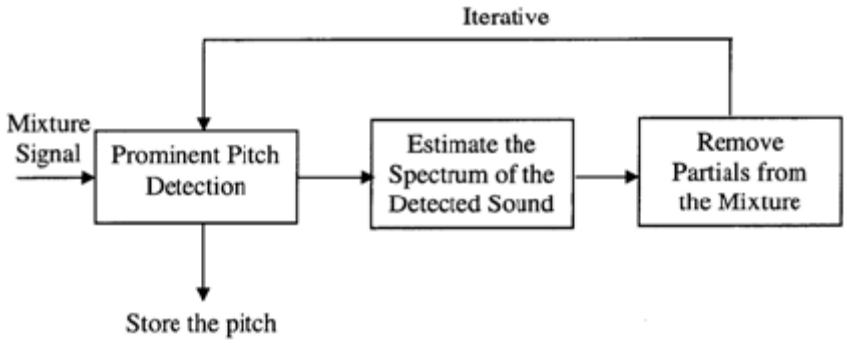

Figure 2.4: Block Diagram of bispectrum model [20].

This proposed method shows better multiple pitches for mixtures of real audio signals having both male and female speeches along with guitar and violin signals. Selecting the proper signal segmentation and appropriate dynamic 2-d filters is the crucial part of this method.

In 2005, a new technique describing the way of estimating multiple pitches (multi-FO) in complex audio signals in a computationally efficient manner is proposed by J. Wan *et al.* [22]. Somewhat similar to previous approaches, here also the fundamental frequency is estimated for the predominant pitch and then the detected sound's harmonic structure detected so far is weakened and the residual signal is considered

for next iteration. Block diagram showing a part of procedure of harmonic enhancement system in figure 2.5.

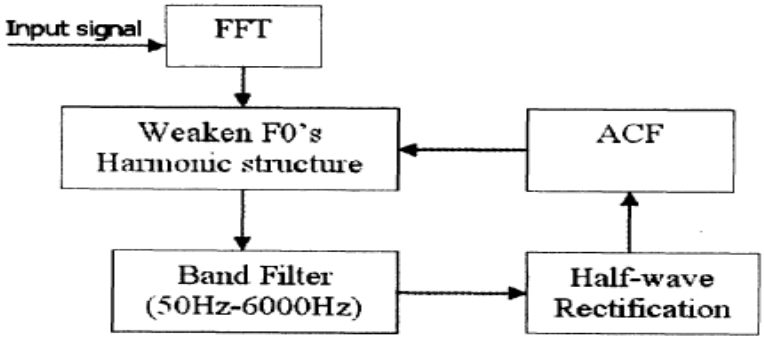

Figure 2.5: procedure of Harmonic Enhancement system [22].

Harmonic enhancement is proposed for estimation in this method. Due to the use of harmonic enhancements along with estimating predominant pitch for human sounds and instrumental sounds, the multiple pitches can be estimated as well. This means this algorithm works well for estimating multiple pitches robustly. Auditory system is not simulated in this algorithm but harmonic structure is enhanced using half-wave rectification. Human auditory system is taken for implementing the half-wave rectification. What happens in brain is very important in order to observe the pitches of several simultaneous sounds. Use of this approach makes the estimation more robust as compare to other methods. The basic idea of this proposed algorithm is that the F0's harmonic structure should be enhanced so that ACF can detect it easily. This proposed approach also decreases the computational complexity of performing multiple pitch estimation.

In 2007, some new methods are described by M. G. *Christensen et al.* [23] for estimating multi-pitch and fundamental frequency. Previous methods which use the nonlinear least-squares, Multiple Signal Classification (MUSIC) and the Capon principles estimate the fundamental frequency by one- dimensional search. Monte Carlo simulations are used by these methods for evaluating their statistical properties. Statistical signal processing is used in this new proposed estimator for finding fundamental frequencies $\{\omega_k\}$. Proposed estimators contain an approximate nonlinear least-squares (NLS) method; a Multiple Signal Classification (MUSIC) based method as well as a Capon-based method. All the proposed methods can be represented by following simple form:

$$\{\widehat{\omega_K}\} = arg\ max_{\{\omega_k\}} \sum_{K=1}^{K} J(\omega_k) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [23]$$

Cost function $J(\omega_k)$ for multiple ω_k are evaluated and then the highest peaks are picked for estimating the fundamental frequencies. Multiple dimensional searches are avoided as they cost a lot. Some Estimators presented:

1. Approximate NLS-based Method

Nonlinear least-squares method is used in this. Assumption of white Gaussian noise is used which says that the NLS method and the maximum likelihood method are equivalent. For attaining CRLB, high number of samples is taken. NLS method is also beneficial for sinusoidal estimation problem as asymptotic CRLB is achieved for large *N* in the colored Gaussian noise case. This method is robust to the color of noise. Single-pitch case and colored Gaussian noise are deal in a more computationally efficient manner by NLS method [24].

2. MUSIC-based Method

Orthogonally principle of MUSIC is used in this method to examine a subspace approach. This principle says that the signal and noise subspaces are orthogonal. Fundamental frequencies with high resolution and order can be estimated with this principle [4]. This approach is also implemented in case of multi-pitch estimation problem [25].

3. Capon-based Method

Capon approach is used in this estimator. This approach uses the designing of various filters that pass power undistorted at specific frequencies. Along with this they also decrease the power at all other frequencies. The filter design problem can be described as the optimization problem:

min_H Tr $[H^H\ RH]$ subject to $H^H\ Z_K = 1$ ..... [23]

Where

H^H→filter bank matrix

L→number of filters

M→length of filters

I is the $L \times L$ identity matrix.

As it is known that the asymptotic assumption is used by NLS method for finding fundamental frequencies independently no such assumption is used by MUSIC

approach. NLS approach is asymptotically efficient in case of colored noise; on the other hand the MUSIC approach used the covariance matrix decomposition which depends on the phase's distribution and noise's whiteness. Only MUSIC approach needs the prior knowledge of number of sources in order to evaluate the cost function. MUSIC- and Capon-based methods resulted in excellent statistical performance in case of multi- and single-pitch. On the other hand NLS perform well for single-pitch but in case of multi-pitch the algorithm does not perform well. Capon-based approach performs better than the MUSIC-based method in case of closely spaced fundamental frequencies.

In 2008, X. Zhang *et al.* [26] present novelty weighted SACF that uses the weighted summary correlogram for detected multi-pitch. This algorithm indicates the pitch period. Modelling of relationship between F0 and response frequency of channel is done by modified amplitude of ACF in this pitch period. Relationship between fundamental frequency (F0) of periodic sound and response frequency of its dominated channels is modelled by conditional probability. This conditional probability is taken as weight in this algorithm. Robustness to noise and to sub-harmonic error is achieved by SACF due to this weight. Post-processing in weighted SACF includes: pitch space measurement in a frame indicating number of perceived pitches; tracking of pitch contours which are further used to make the connection for pitches on individual frame into continuous tracks. This algorithm performs well for both single and multi-pitch. This algorithm can also perform in noisy environment. 10 voiced speeches and 10 different kinds of noises are mixed for testing the performance of this proposed algorithm. The proposed algorithm performs better than existing algorithms but it faces the problem of harmonic error. This problem is faced because along with pitch delay, multiple delays also have peaks in conventional SACF method. So, due to noise and pitch variation, having high peaks on multiple pitch delay is more risky as compare to true pitch delay. This arise the sub-harmonic error in the system. Multi-pitch detection faces many other problems; one of them is deciding the number of pitches on a frame. In [27], Hidden Markov Model (HMM).is used for modelling pitch number. Background noise coefficient is also proposed for handling problems of pitch number.

In 2009, R. Badeau *et al.*[28] Concept of overlapping partials of fundamental frequencies is used in this approach for multi-pitch estimation. Expectation maximization algorithm is used by this approach so that likelihood of the observed spectrum can be maximised. This is done by successively estimating the single-pitch and spectral envelope. This algorithm is mainly implemented in the field of musical chord identification. This approach performs the optimization of joint criterion by means of a recursive algorithm. In order to do this; expectation-maximization (EM) approach is implemented on spectrum model. This implementation resulted in proper statistical framework for the proposed method. EM algorithm has some advantages; some of them are low complexity and capability of handling spectral overlap between the harmonic components. First advantage is achieved because J successive single-pitch estimations are done for multi pitch estimation rather than exploring a vector space of dimension J. For second advantage the spectral envelope's smoothness is considered. As the log-likelihood function $Q_{\pi,s}^{n}$ which need to be optimized is not smooth so this may lead to trap the algorithm into local maxima1. So, methods for escaping from these maxima are required. This is done by testing multiples or sub-multiples of the current estimated frequencies. In this whole concept initialization also plays an important role. Previously wronged estimated itches can be correctly estimated after proper initialization. This method is reliable because of robustness to overlapping partials along with the capability to simplify the multi-pitch estimation task. This proposed method is evaluated by implementing it on audio-like synthetic signals.

In 2010, a new method for determining the number of speakers at frame level is proposed by P. Mowlaee *et al.* [29]. Author takes the inspiration from asymptotic maximum a posteriori rule for model selection. Multiple hypotheses tests based on various speaker models are done on observed signal for determining the number of speakers. The method of double-talk detection improves the performance of speech recognition in case of multiple speakers. In case where only one speaker is active, silence state is added to speaker code books. Model-based speaker identification (SID) module, called Iroquois [30] is used to find the number of speakers in mixture signal. Silence and mixture segment is excluded from procedure updating parameters. In this method that segment is selected where there is only one dominant speaker. Computationally auditory scene analysis (CASA) [31] is a source driven approach in

which time frequency segments from same source are combined and concatenate together to form a single stream. So, methods using CASA detects the number of persons speaking without requiring the prior knowledge of any speaker model [31]. However, these methods apply the multi-pitch estimator for estimating pitch trajectories. So, the accuracy of CASA depends on accuracy of multi-pitch estimator.
In proposed algorithm maximum a posteriori (MAP) criterion is combined with the SCSS to select the model for getting the number of persons. Multiple hypothesis tests are conducted so that the double-talk/single-talk regions in segments of the mixed signal can be determined. Firstly the person detection is explained for two speakers at different signal-to-signal ratio (SSR) levels, then explanation is extended for multiple persons in speech signal. Quality of the separated output signals is also considered by the proposed method. After finding the single-talk regions by using double-talk detector, the SCSS work is only to separate the mixture segments. Block diagram showing frame level speaker determination methodology in Fig. 2.6

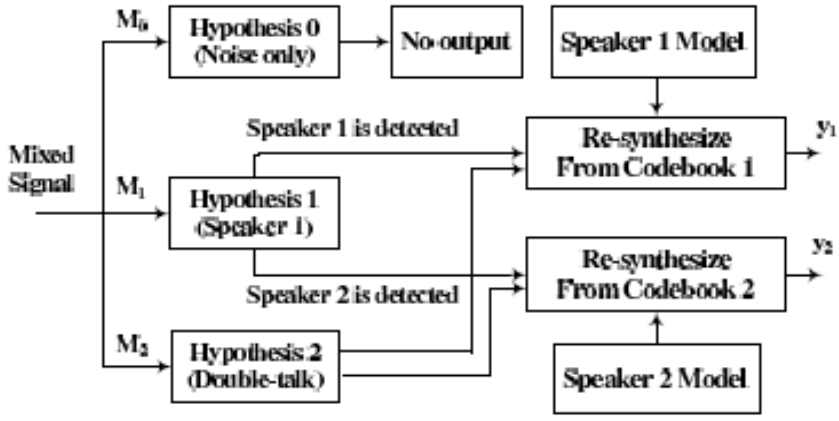

Figure 2.6: Block Diagram of frame level speaker Determination [29].

It is clear from results that the proposed method improves the signal quality as compared to methods where no detector was used. This method can easily find the single-talk and doubletalk regions in both cases of speaker and gender dependency.

In 2010, E. Vincent *et al.* [32] proposed models for time varying amplitude speech signals. They used their basic model as non-negative matrix factorization.

In 2011, a technique for automatic transcription of music signals is proposed E. Benetos *et al.* [33]. This model use the joint multiple-F0 estimation and subsequent

note tracking. The constant-Q resonator time–frequency image is used in time-frequency representation of music signals. Along with this a noise suppression technique is applied in the pre-processing step .In order to use this suppression technique, pink noise assumption and cepstral smoothing is considered. The optimal tuning and inharmonicity parameters are computed and a salience function is proposed in order to select pitch candidates. After this an overlapping partial treatment procedure is executed for each pitch candidate combination. An optimal pitch set is selected by using a scored function proposed in this approach. This function combines spectral and temporal features. Post-processing stage contains the Note smoothing, which lead to employing HMMs and conditional random fields (CRFs) [34]. Block diagram shows full system functionality in Fig. 2.7

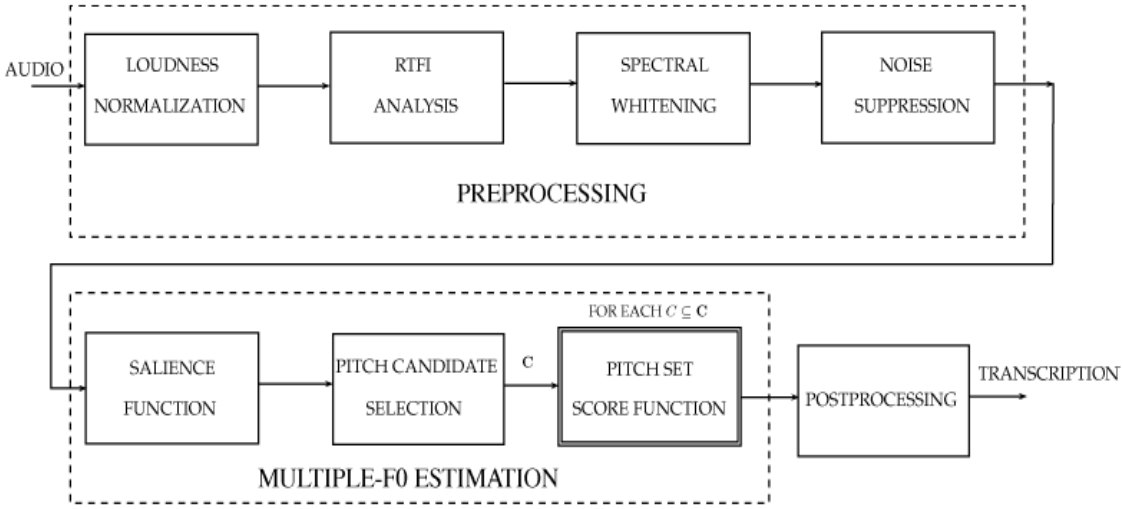

Figure 2.7: Block Diagram of HMM based system [33].

The proposed system shows the improved results in case of pitch combinations instead of iterative selection.

Koretz *et al.* [35] used Maximum A posteriori probability algorithm for multiple pitch detection.

In 2011, a multi-length windows (MLW) method is proposed by Q. Huang *et al.* [36] for estimating the multiple pitches of the single-channel mixed speech of two simultaneous speakers. This method used the harmonic model because overlapping between the harmonics of two concurrent speech signals can happen from the short time window analysis. In order to estimate the pitch and distinguish the potential harmonic peaks, the longest window is taken into consideration while the remaining

windows are used for rectifying the initial estimation. The proposed algorithm works in two steps: prominent pitch estimation and other pitch estimation. Prominent pitch means the pitch from the stronger source. The longest frame is considered for calculating initial pitch value. Then this pitch value is modified so that it fits well into short frame in order to assure the short time stationarity of speech. After this a specific spectral peak value is evaluated. Harmonics of other sources are yielded into shorter frame by this spectral peak value. Next step is to divide the harmonic frequency with the harmonic order for finding the modified pitch. Advantage of this proposed model is that multi-pitch estimation can be done in cases of both the short time stationarity of speech signal and high frequency resolution. This method results in more accurate pitch estimation .It also removes the problem of low frequency resolution of harmonic model. This algorithm performs much better than the MMSE method and SW harmonic method.

In 2013, Block sparsity for estimating the fundamental frequencies is proposed by S. I. Adalbjornsson *et al.* [37]. Signals considered as input in this approach contains the multiple harmonically related sinusoidal signals. Both the simulated and measured audio signals are used for confirming the accuracy and correctness of the proposed method. Previous estimation problem is considered and the block sparse formulation of that problem is introduced. This formulation says that the signal can be formed from a dictionary that contains set of signals that are harmonically related. This results in solving the convex optimization problem. Another algorithm derived by the author uses the alternating directions methods of multipliers (ADMM) technique. A signal containing two sources is formed .These sources have the fundamental frequency equals to ω_1 and ω_2, drawn uniformly on $[0.02\pi, 0.2\pi\ ]$. These frequencies have the minimum difference of 1/30 of the frequency range. In order to evaluate the performance of estimates of various algorithms, the 250 Monte-Carlo simulations are used. In this simulation number of harmonics is selected uniformly over [3, min(floor(π/ω_i),10)]. All the frequencies considered in this approach should be below the Nyquist limit and with amplitudes drawn as $a_{i,k}$~N(0,1). $10\log_{10}(||y||2/||w||2)$ represent the signal to noise ration and set it to15. These algorithms are compared with the previous proposed algorithms in terms of robustness and performance. All the previous algorithms works well when there is only one toe, but after mixing two

tones only the proposed algorithm and the ORTH algorithms estimates the pitch frequencies accurately.

In a method presented by Mirza Cilimkovic [38] for classification and clustering in data mining using Neural Networks (NN) as a classifier, the activities of brain were mimicked and ability to learn was discussed. The NN discussed had three basic layers namely input, output and hidden layers. The numerous nodes present in these layers which are attached to nodes present in other layers share data with each other. The nodes were provided with a weight component to show the connections. This arrangement was capable of learning from examples. The more the examples, the more such neural network could learn.

An improve genetic algorithm based approach was fiven by Hitesh Gupta, Deepinder Singh Wadhwa [39] for enhancing performance of speech signals under noisy conditions. The speech enhancement based algorithms basically have three categories namely filtering/estimation based noise reduction, beam forming and active noise cancellation (ANC) technique. This algorithm could easily discriminate among words and hence is responsible for evolutionary computation of genetic algorithm.

In a paper presented by Pan et al. [40] on Emotion Recognition (SER) a significant contribution to Human Machine Interface has been made. The three emotional states i.e. sad, happy and neutral were monitored under this research. The features which were explored: Mel-energy spectrum dynamic coefficients (MEDC), energy, Mel-frequency spectrum coefficients (MFCC), pitch and. linear predictive spectrum coding (LPCC). The two databases used for training using SVM classifier enabled authors to study multiple combinations of features over the different database are compared and explained. The best results were yielded by combining three features (Energy+MFCC+MEDC) in terms of accuracy rate of 91.3% on Chinese emotional database and 95.1% on Berlin emotional database.

Utane et al. [41] discussed the increasing role of human language in human machine interfaces. Human speech varies with different feelings and this variation can be judged easily. Studies were carried out on different emotional states like sad , anger , surprise , happy and neutral for differentiation of speakers emotional state. Hidden Markov model classifiers and Gaussian mixture model was proposed in this research for effectively judging the emotion of the speaker though speech. Different feature are extracted i.e. spectral features such as Mel frequency cepstrum coefficient and prosodic features like pitch, energy.

Sapra et al. [42] defined the relation of physiological and mental state with the emotions. Different physiological and mental state can result in different feelings, behavior and thoughts. Emotions are often related with personality, deposition, mood and temperament. The method of detection for of human emotions can be based on different acoustic features like pitch, energy etc. MFCC approach proposed under this paper uses nearest neighbor algorithm for classification. Emotions among males and females can be characterized differently as the acoustic features differ for both, and hence do the MFCC.

They introduced an alternative algorithm for the detection of multiple pitches in speech signal. The work is going on the similar concept and new techniques have been introduced. This is further carried on to better accuracy and size of dataset.

Table 2.1: Summary of Survey analysis

Year	Name of Author	Algorithms Used	Input	Output
1999	P. J. Walmsley *et al* [14]	Polyphonic pitch tracking using joint Bayesian estimation	Musical Signal	Parameters based on frequency variation
2001	A. P. Klapuri [17]	spectral smoothness principle	Polyphonic, multi-instrumental music and mixtures of simultaneous speakers	Segregate harmonic signals
2003	A. P. Klapuri [19]	Based on Harmonicity and Spectral Smoothness	Acoustic input signal sampled at 44.1 kHz rate and quantized to 16-bit precision.	Clean signal and noisy signal with different SNR's
2004	S. S. Abeysekera [20]	Using frequency-lag domain and Bispectrum	Multi-pitch input spectrum	Autocorrelation channels of 32~120 filterbank

				outputs.
2008	X. Zhang *et al* [26]	Based on Weighted Summary Correlogram	Multi-pitch input signal	Estimation of closest pitch frequency
2009	R. Bedeau *et al* [28]	Expectation Maximization algorithm	Musical chord	Estimate successive single - pitch and spectral envelope.
2010	E. Vincent *et al* [32]	Adaptive Harmonic Spectral Decomposition	Multi pitched musical audio, piano recordings, narrow band spectra	Estimates the model parameters and performance of pitch
2011	E. Benelos *et al* [33]	Using Harmonic Envelope Estimation	Musical signal and recordings with sampling rate 8kHz	Estimation of log-frequency spectral envelope.
2011	A Koretz *et al* [35]	Based on Maximum A Posteriori Probability	Monophonic and polyphonic signals, music and synthetic signals	Determine different harmonic sources present
2011	Q Huang *et al* [36]	Based on Multi-Length Windows Harmonic Model	Harmonics of two concurrent speech signals	Estimation of prominent pitch and autocorrelation
2013	S. I. Adalbjornsson [37]	Using block sparsity	Multiple harmonically related sinusoidal signals	Estimate pitch frequencies of two different signals.

Chapter 3
Problem Formulation

This chapter contains problem statement of this thesis which is much in discussion about multi pitch and their gender identification and emotion detection of a audio signal.

3.1. Problem Definition

Pitch perception is very complex process. Pitch determination of single source is easy as compare to multi-Sources or of polyphonic signals. The problem of estimating the fundamental frequency or pitch of periodic waveforms occurs in various form of application, and has received notable interest over the recent years, for example, several speech and audio problems notably depend on the initial forming of an estimate on the pitch or pitches including problems. Fundamental frequency plays a vital role in Gender identification also. There are many effective and accurate proposed algorithms on single source pitch determination and detection. But the multi-pitch real life scenario occurs regularly than single pitch case, and often also in speech processing. It is difficult to accurately estimate the multiple pitches of a mixed signal; multi-pitch estimation has potential application in speech sources separation, speech enhancement, speech recognition.

The important work done and algorithms proposed so far in pitch analysis mainly works on the concept of detecting pitches from a music corpus or mixture signals, containing various pitches of different instruments or humans. So far there is no such algorithm which itself identify number of person from the mixture signals with much of accuracy. The problem of estimating the fundamental frequencies of a signal containing multiple harmonically related sinusoidal signals motivated us to work on it. So, the problem here is to identify the number of pitches from the mixture signal. Along with this, the gender identification of those pitches is also taken into account. Here in this thesis work, an iterative approach for frame by frame analysis has been considered for multi-F0 detection.

This thesis work also focuses on detecting emotions from audio signals. Algorithms proposed till now are lagging behind in terms of performance. So the main target is to detect emotions in audio files without compromising the performance parameter.
Audio files containing voice of one person in one file is taken as input for this system. This proposed system mainly works on three types of emotions which are sad, joy and aggressiveness. In order to optimize the performance, genetic algorithm is taken into account for this.

Chapter 4
Implementation

4.1 Proposed System

There are previously many algorithms proposed in the field of multi-pitch detection and emotion detection. All the algorithms determine multi-F0 but most of them work on music notes or music transcription. We propose a recursive algorithm which is used for estimating fundamental frequencies and number of harmonics of a multi-pitch complex signal based on the harmonic summation method. Determining the number of harmonics is also done using the same algorithm and it continues to extract sources until the pitch detection criteria fails. We improve the performance and accuracy of emotion detection system by using genetic algorithm and GTCC feature.

4.1.1 Harmonic Model

The voiced speech signal is defined through various models; one of them is harmonic model. According to this model, the sum of sinusoids with time varying amplitudes, frequencies and phase is equal to the voices speech signal [18].

It can be represented as follows

$x(m)= \sum_{r=1}^{R} A_r \cos(2\pi F_r m + \theta_{0r})$ .....[18]

Where

R = harmonic order

A_R = instantaneous amplitude

F_r = frequency of the r^{th} harmonic component, approximates to rF_0.

θ_{0r} = initial phase

F_0 = fundamental frequency, also called pitch

4.1.2 Harmonic Structure

Initially the acoustic input signal is taken into account for calculating Fast Fourier Transform .This transformation used to convert time/space domain signal to complex frequency domain signal as output. Now this complex signal is divided into window frames of length 25ms.In our system sampling is done at 8000 KHz and quantized to

32 bit float precision. Width of each window function is considered to be 16384 samples.

Range of pitch filter considered in our algorithm is 100 Hz to 400 Hz. Control flow of multi-pitch detection has been shown below:

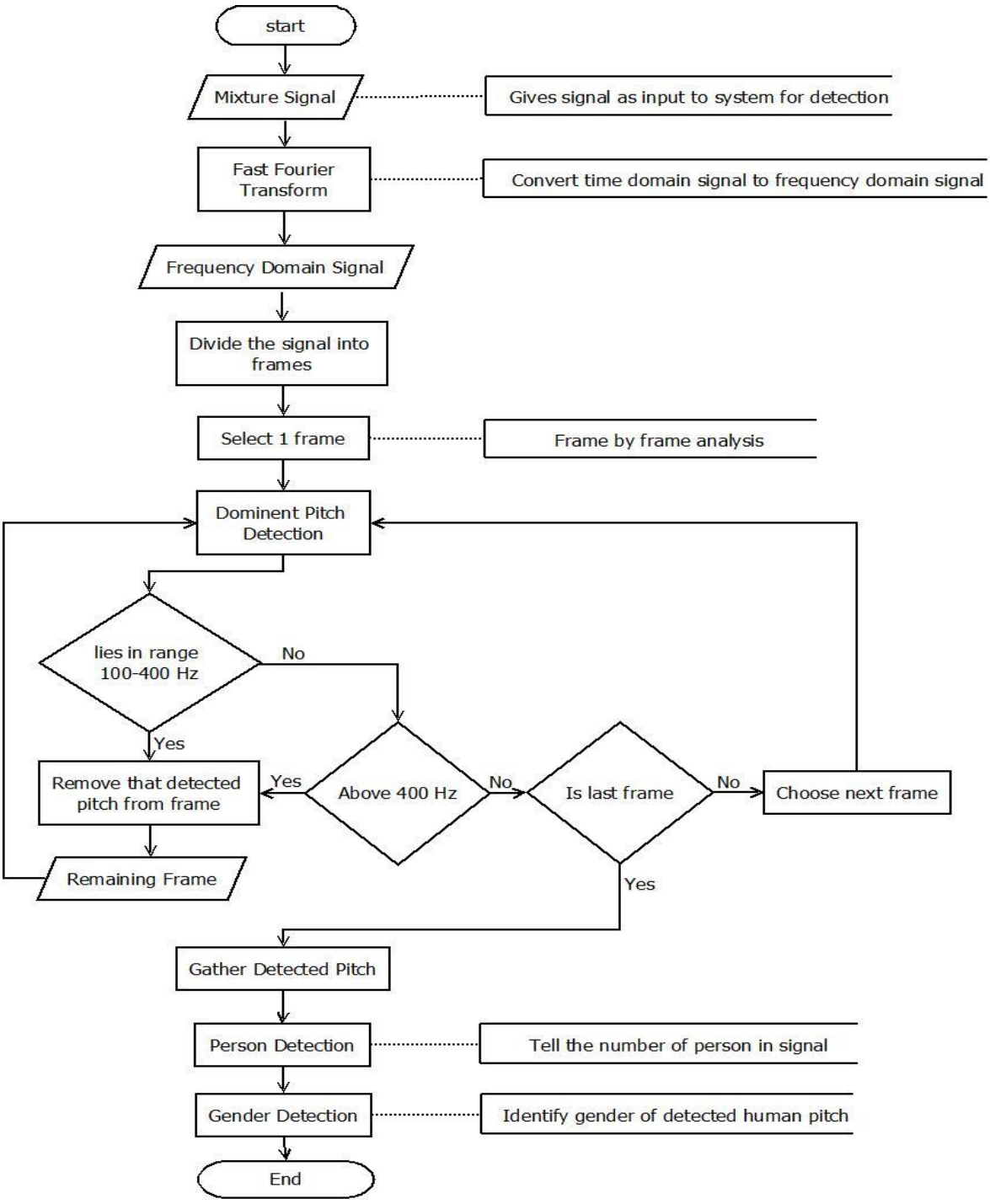

Figure 4.1: Flow chart of proposed system.

4.1.3 Harmonic Structure Subtraction

Proposed algorithm is applied on each and every frame in order to detect the dominant pitch in mixture signals. Amplitude of harmonic partials of the detected pitch is estimated. After this detected pitch is removed from the frequency domain of selected frame and the algorithm is iteratively applied on the residual signals.

4.1.4 When to Stop the algorithm

Iterative execution of harmonic subtraction step will leads to a stage where the pitch in residual signals is nearly same. So, no peak can be detected with the algorithm. At that time algorithm can stop working on that frame.

4.2. Algorithms used in proposed Multi-F0 Estimation model

Proposed model contain 7 algorithms/functions. Work together as a system.
Algorithm sequence below is as per the model working.

Function name is Main ( ). Purpose of this function is to read the wave file and
Then divide the signal into frame of 25 msec. After dividing signal into frames it supplies these frames to multi_pitch _detection ( ) function for detecting multiple pitches.

Main()

1. Filename = recorded speech file for multi-pitch test
2. [input , fs , nbits] = wavread (filename) // function used in MATLAB to read wav file input wav file and it output sampled data in input , sampling rate in fs and number of bits used to encode file in nbits.
3. Initialize parameters

   $F=2^{14}$

   N=200 //Frame length as per formula

   Frame_length = frame_duration x sampling rate.

   f_est = [ ] // matrix used to store frequency of single person

   f2_est = [ ] // matrix used to store frequency of two person

f3_est = [ ] // matrix used to store frequency of three person

pos = 1 to N

f_0_min = 100 // Minimum frequency of human being

f_0_max = 400 //Maximum frequency of human being

f_0_lim = 2 * π * [f_0_min f_0_max] / fs //establishing min and max frequency range

4. While pos <= length (input)
5. x = input (pos)
6. [w_0 , L] = multi_pitch_estimator (x, f_0_lim, F) //Call to multi_pitch_estimator function
7. f_0 = w_0 / 2 / π * fs // frequency in radian to frequency in Hz conversion formula
8. if size (w_0) = 1
9. f_est [ ] = f_0
10. else
11. if size(w_0) = 2
12. f2_est [ ] = f_0
13. else
14. f3_est [ ]=f_0
15. pos = pos + N

Function name is multi_pitch_detection (x, f_0_lim, F). Purpose of this function is to search for the multi fundamental frequencies and their order. It is recursive function. Input parameters are x input audio signal, f_0_lim is frequency range set according to human voice and F is FFT size. This function calls various different functions to search for multiple fundamental frequencies. It output fundamental frequencies and their corresponding model order.

multi_pitch_estimator (x, f_0_lim, F)

1. N = length (x)
2. r = x
3. R = fft (r , F) // used to do fft transformation and return F - point Discrete Fourier transformation(DFT)

4. [w_0 (1) , L (1)] = dominant_pitch_estimator (x , f_0_lim , F) // call to function dominant_pitch_estimator
5. a = find_amp (x , w_0 (1) * [1 : L (1)] // call to function find_amp used to calculate amplitudes of finded frequencies in previous step
6. Z = vandermonde (w_0 (1) * [1 : L ((1)] , N) //call vandermonde function used to store DFT matrix
7. k = 2
8. H = pitch_detect (r , w_0 (1) , L (1) , a) //call function pitch_detect used to detect pitch existence in frame
9. While pitch exist
10. r = r - Z * a //Dominant pitch substraction
11. R = fft (r , F) //Remain signal undergoes FFT transformation
12. [w_0 (k) , L (k)] = dominant_pitch_estimator(r , f_0_lim , F)
13. If L (k) > 0 //Harmonic Order is greater than zero
14. Z = vandermonde (w_0 (k) * [1 : L (k)] , N)
15. a = find_pitch (x , w_0 (k) * [1 : L (k)])
16. H = pitch_detect (r , w_0 (k) , L (k) , a)
17. Else
18. H = 0
19. k = k + 1
20. Return multi-pitch and harmonic order

Function name is dominan_pitch_detection (x, f_0_lim, F). Purpose of this function is to search for the dominant pitch in the frame and the harmonic order of that pitch. Input parameters are x input audio signal, f_0_lim is frequency range set according to human voice and F is FFT size. It output a fundamental frequency i.e. dominant pitch and its corresponding model order.

dominant_pitch_estimator (x, f_0_lim, F)

1. N = length (x)
2. X = fft (x , F) //FFT transformation
3. freq = [round (f_0_lim (1) / 2 / π * F) + 1: 1: round (f_0_lim (2) / 2 / π * F) + 1] // creates freq vector of range min frequency to max frequency i.e. 100 to 400
4. set w_0_set to 2 * π * (freq - 1) / F

5. Set parameters

   L_max = floor (N/ 4)

   J=zeros(size(w_0_set))

   L=zeros(size(w_0_set))

   h=1

6. for $w_0 = w_0$_set
7. L_w_0 = min ([floor (2 * π / w_0) - 1 L_max]) //assigning minimum value
8. P = zeros (L_w_0, 1)
9. L = zeros (L_w_0, 1)
10. len = 1
11. while len < L_w_0
12. f = round (w_0/ 2 / π * [1 **:** len] * F) + 1 // value where fundamental frequency found
13. c = X (f) / N
14. Z = vandermonde (w_0 * [1 **:** len] , N)
15. P (len) = var (x - Z * c)
16. c (len) = N * log(P (len)) + len * log(N) }+ 3/2 *log(N)
17. len = len + 1
18. [temp1 , temp2] = min(c) //c is matrix of n x 2
19. L(h) = temp2
20. J(h) = c (temp2)
21. h = h + 1
22. [temp1,temp2]=min(J)
23. Return harmonic order L(temp2)
24. Return fundamental frequency w_0_set(temp2)

Function name is find_amp (x, w). Purpose of this function is to calculate the amplitudes of the frequencies we find in multi_pitch_detection function. Input parameters are x input audio signal, w set of frequencies found. It returns a vector containing amplitudes of respected frequencies.

find_amp (x , w)

1. N = length (x) // taking full length of input file
2. Z = vandermonde (w , N)

3. a = Z / x
4. return a

Function name is vandermonde (w, N). Purpose of this function is to calculate the vandermonde matrix which is used to store the M-point DFT matrix for some set of frequencies. Input parameters are w set of frequencies found and N length of input file. It outputs a matrix of dimension N by number of frequencies .

vandermonde(w,N)

1. Transpose of w
2. Create a new matrix of z of dimention (w , length (N)) and initialize it with zero.
3. z = exponential (j*w)
4. for n = 1 to N
5. Z (n, each Coloumn) = z . ^ (n-1) // every element of z get multiplied by (n – 1)
6. Return Z

Function name is pitch_detect (x, w_0, L, a). Purpose of this function is to find that whether input file contain pitch or not. Criteria are based on the noise present in input signal or not. Input parameters are x input audio signal, w_0 set of frequency, and L is harmonic order of those frequencies. It returns either 1 or 0 depending upon the frequency found or not.

Pitch_detect (x, w_0, L, a)

1. N = length (x)
2. Z = Vandermonde (w_0 * [1 **:** L] , N)
3. If N * log (var (x – Z * a)) + L * log (N) + 3/2 * log (N) < N * log (var (x)) //Pitch detection creteeria
4. Return H=1 // pitch detected
5. Else
6. Return H=0 //pitch not detected

Function name is gender_identification (f_est, f2_est, f3_est). Purpose of this function is to find the number of person and their respective genders. We uses 5 thresholds each have different value set according to audio file length taken as input in main( ) function. Input parameters are f_est use to store frequency of single person, f2_est use to store frequencies of two person and f3_est use to store the frequency of three person depending upon the frame output. It Display final output by displaying number of person with their genders.

gender_identification (f_est, f2_est, f3_est)

1. Calculate number of rows and columns of f_est, f2_est and f3_est
   Assign them variables
   (r_1,c_1) = size (f_est)
   (r_2,c_2) = size (f2_est)
   (r_3,c_3) = size (f3_est)
2. If $c_1 < threshold_1$ and $c_2 < threshold_2$
3. Display 1 person
4. If mean(f_est) < male frequency range
5. Display Male
6. Else
7. Display Female
8. Else
9. If $c_1 > threshold_3$ and $c_2 > threshold_2$ and $c_3 < threshold_3$
10. Display 2 person
11. P = mean (f2_est) // p is a matrix of 2 rows
12. If P (1,1) < Male frequency range
13. Display Male
14. Else
15. Display Female
16. If P (2,1) < Male frequency range
17. Display Male
18. Else
19. Display Female
20. Else
21. If $c_2 < threshold_4$ and $c_3 > threshold_5$
22. Display 3 person

```
23.     P = mean( f3_est )                  // p is a matrix of 3 rows
24.     If P (1,1)  <  Male frequency range
25.         Display Male
26.     Else
27.         Display Female
28.     If P (2,1)  <  Male frequency range
29.         Display Male
30.     Else
31.         Display Female
32.     If P (3,1) <  Male frequency range
33.         Display Male
34.     Else
35.         Display Female
```

4.3. Flow of Multi-F0 Estimation System

1. The very first step of this proposed system is to get the mixture signal as input. For this Audacity tool is used to record the signal as shown in figure 4.10.

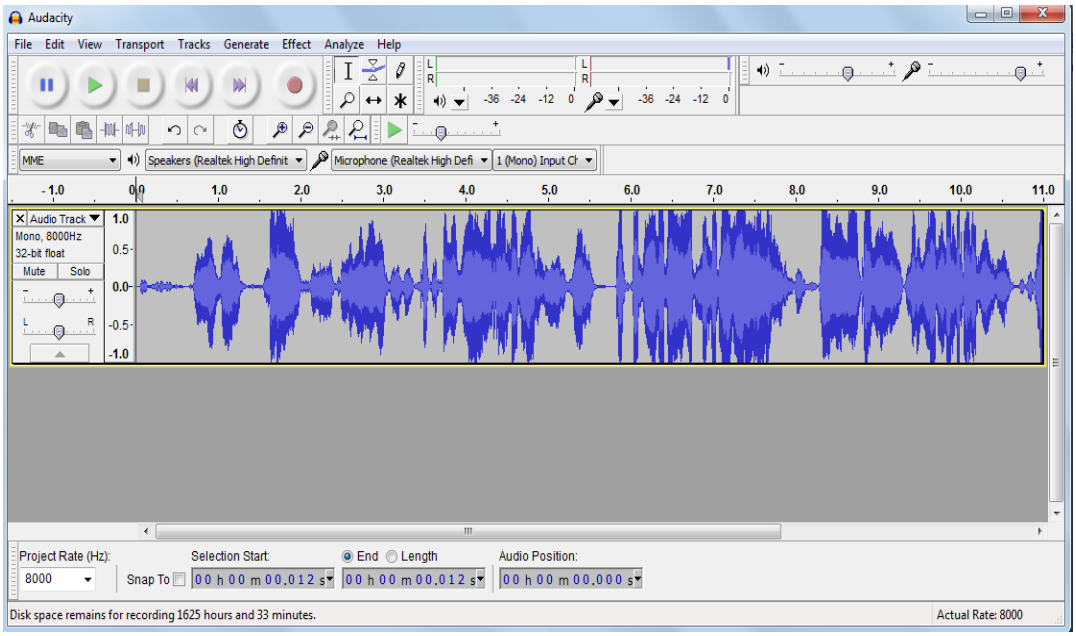

Figure 4.2: recording.

Then this recorded signal is cut to get the file having size of 1 second as shown in figure 4.11.

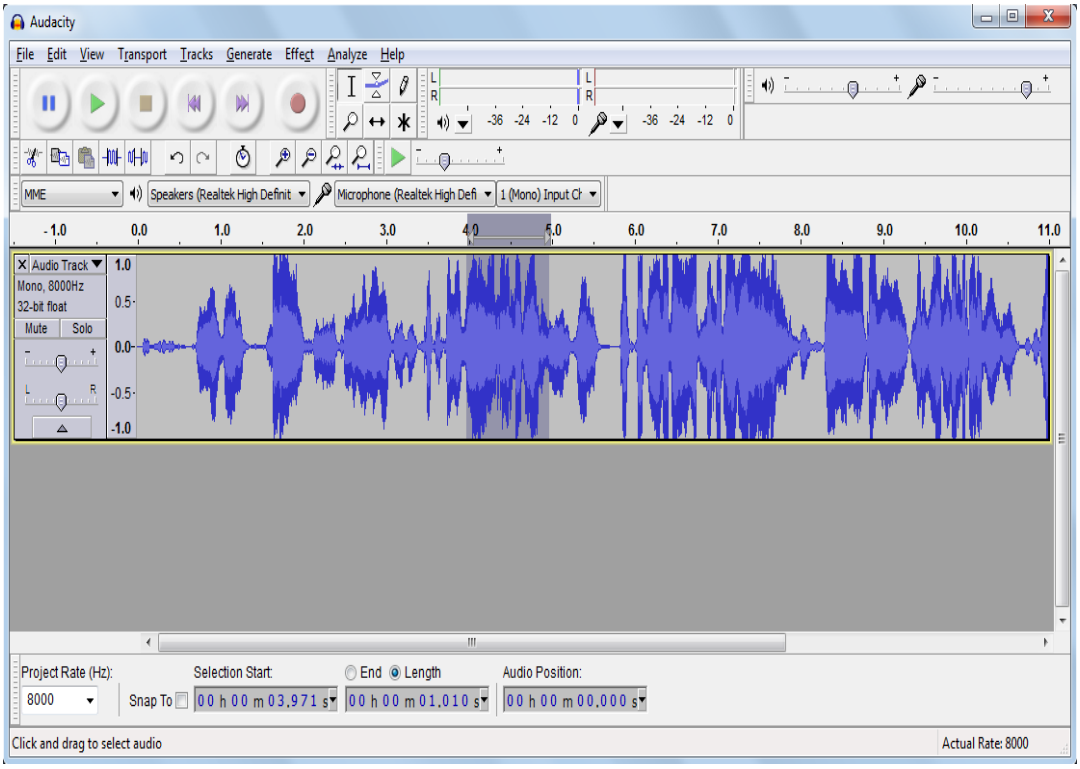

Figure 4.3: recording of cutting into 1 sec file.

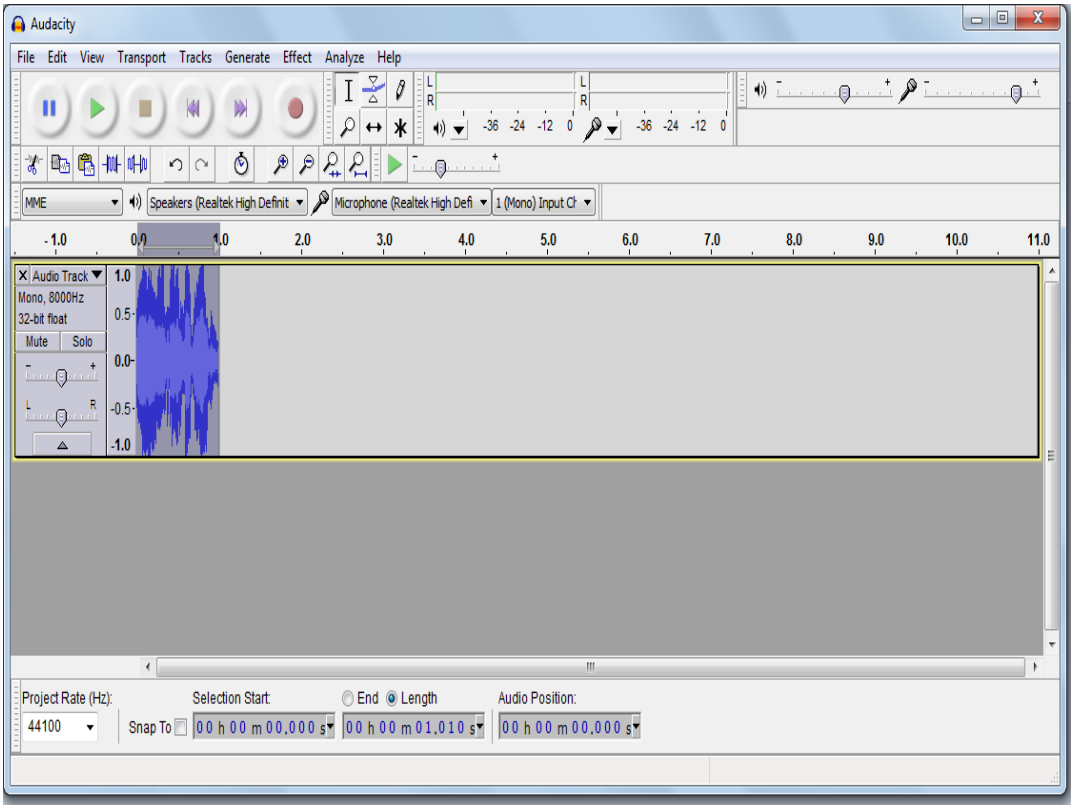

Figure 4.4: Save this 1 sec part as a new file.

This file is then converted into wave file having extension .wav. Wave read function is used to convert this file into wave file. For recording wave files various parameters such as sampling frequency, length of file etc need to be set first. So, after setting these parameters, wave files are recorded successfully. This wave file is referred as mixture signal in this thesis.

2. In second step, this mixture signal is taken as input. This mixture signal is currently in form of time domain. In order to convert it into frequency domain signal, fast Fourier transform is implemented. FFT function is used to perform this transformation.

3. In third step, the frequency domain signal is divided into frames for performing further operations on it. Signal is divided into frames on the basis of formulae described below:

Frame_length = frame_duration x sampling rate.

4. In next step, each frame is taken one by one for determining dominant pitches in that frame. For this, a range of pitches is predefined to the system. System detects one pitch from the frame and then confirms that whether that pitches lies in the predefined range or not. If it lies in the range than that pitch is removed from frame and saved in Vander mode matrix. If not, than next pitch is determined from same frame. In this way all the frames are analysed to gather the pitches that lies in the predefined range. So at the end of this step, system provides a Vander mode matrix that contains all the pitches that were removed from frames.

5. In fifth step, the number of persons is detected. Approach followed for this is described below: Firstly the person detection is explained for two speakers at different signal-to-signal ratio (SSR) levels, then explanation is extended for multiple persons in speech signal. Quality of the separated output signals is also considered by the proposed method. After finding the single-talk regions by using double-talk detector, the SCSS work is only to separate the mixture segments.

Some basic notation and definitions are now explained for making base of this approach. Mixed signal with N samples $y \in R^N$ is taken. These are composed of up to J speaker signals as $y = \sum_{j=1}^{J} s\left(\varphi_j\right) + e$. The matrix transpose is represented by T, $j \in [I, J]$ represents the number of signals in the mixed signal.

The j^{th} signal is represented with $s(\varphi_j) \epsilon R^N$. This jth signal is characterized by parameter vector φ_j and $e \epsilon R^N$ depicts the noise signal. All these are incorporated in the model. Sinusoidal modelling is used in this proposed approach for modelling the j^{th} speaker signal in the mixture as a parametric feature vector φ_j. This is composed of sinusoidal parameters such as amplitude, frequency and phase vectors. This system use K = 3 candidate models each denoted by M_k, for describing the mixed signal, y. M_0, $M_1$and M_2 are used to indicate noise-only, single-talk and double-talk respectively. Parameter vector $\emptyset_k$with L_k sinusoids are used to describe these models. So, basically the proposed approach addresses the following problem: given the mixed signal, select the model which is the most likely. Three models are considered for y, which are:

Case 1 M_0 : y = e,

Case 2 M_1 : y = s(ψ_j) + e for j ∈ [1, 2],

Case 3 M_2 : y = s(ψ_1) + s(ψ_2) + e,

Case 3 equation s(Ψ_1) + s(Ψ_2) shows estimation for the mixed signal, Case 2 s(Ψ_j) with $j \in [1, 2]$ Represents the j^{th} signal modelled. Ψ_j depicts the parameter set.

Now, the posterior probabilities of M_k with k ∈ Z_k = {0, 1, 2} is evaluated by system. The system estimate of the most likely hypothesis is denoted by M_k, and is obtained as:

$$\widehat{M}_k = \arg max_{M_k:\, k \in Z_K} \{\int_{\theta_k}^{k} p(y|\theta_k, M_k)\, p(\theta_k|M_k) d\theta_k\} \quad \ldots\ldots\ldots\ldots\ldots (1)$$

Equation (1) represents a complicated nonlinear maximization problem which occurs due to the used models. But in this proposed system the different criterion is used instead of numerical integration for the evaluation of marginal density in (1). Asymptotic criterion is defined as:

$$\widehat{M}_k = \arg max_{M_k:\, k \in Z_K} \{- \ln p(y|\hat{\theta}_k, M_k) + p_c\}, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

In equation (2), P_c is Model-dependent penalty of the criterion, $\widehat{\theta_k}$ is an estimate of θ_k for the k^{th} model M_k and - ln p(y |θ_k, M_k) Log-likelihood term obtained from an approximation of (1). Multiple-Hypothesis Algorithm when there are more than two persons speaking at one time.

Now the target is to determine - ln p(y $|\theta_k$, M_k) for each of the three underlying candidate models M_k with k ∈ Z_k = {0, 1, 2}. Again sinusoidal modelling is used for modelling the speaker signals in the mixture. Let s_i (Ψ_j) be the j^{th} speaker signal with j ∈ [1, 2] for the i^{th} frequency band which is modelled by the parametric vector Ψ_j. It is assumed in this system that the signal modelling error, e has a Gaussian distribution. Another assumption made by this system is that the modelling error sub band signal e_i is white in each i^{th} frequency band. Due to sub band decomposition and the independence assumption for all frequency bands, it is also assumed that e_i is independent from one band to another.

Based on this assumption it is clear that the likelihood function for all bands for each class M_k is given by

$P(e|\sigma^2) = \prod_{i=1}^{Q} p\,(e_i|\sigma_i^2)$

$$= \frac{1}{(2\pi)^{\frac{N}{2}}} * \frac{1}{\prod_{i=1}^{Q}\sigma_i} \exp\left(-\frac{1}{2}\sum_{i=1}^{Q}\frac{e_i^T e_i}{\sigma_i^2}\right), \quad \text{.....(3)}$$

In equation (3), total number of frequency bands represented by Q, variance due to the modelling error signal in i^{th} band represented by σ_i and band is represented by e_i.

For single speaker class, M_1, the modelling error at the *i*th frequency band, is given by $\hat{e_i}$ = y_i - s_i ($\hat{\Psi_j}$), for the mixed class, M_2, the estimated error is defined as $\hat{e_i}$ = y_i - s_i ($\hat{\Psi}_1$) - s_i ($\hat{\Psi}_2$).The noise estimated for the i^{th} frequency band as a coloured noise is not fitted by M_2. The criterion for sinusoids composed of unknown amplitudes and frequencies reduces to

$$\hat{M}_k = \arg max_{M_k:\, k \in Z_K} \left\{\frac{N}{2}\sum_{i=1}^{Q}\ln\sigma_i^2 + \frac{5L_k}{2}\ln N\right\} \quad \text{..... (4)}$$

Where

$\sigma_i^2 = \frac{1}{N}\widehat{e_i^T}\,\hat{e}_i$ Shows the estimated variance

i^{th} Represent the frequency band

Depicts the number of sinusoids

In the mixture class M_2, A mixture estimate is required to replace s($\hat{\Psi}_1$) + s($\hat{\Psi}_2$), so that the best pair of {$\hat{\Psi}_1$, $\hat{\Psi}_2$} can be extracted from the speaker models of the underlying speakers. The minimum mean square error (MMSE) estimator is used for the mixture magnitude spectrum in order to find the joint best states in the

speaker models. These models when combined, best describe the magnitude spectrum for the observed mixture, y. The noise model, M_0 is also included in this as one of the examined models by setting y = $\hat{e}$ and setting the number of sinusoids equal to zero (L_k= 0). The estimated noise variance is given by $\sigma_i^2 = \frac{1}{N} y_i^T y_i$. Finally, using the estimated value for σ_i depending on each possible class of M_k with k ∈ Z_k = {0, 1, 2}, the best model, as a result, is the one which yields high log-likelihood and low model order, which is achieved in (4).

6. In last step, gender of this identified number of persons is determined. For this a predefined range is given to the system. This range is decided after doing a survey on the range of pitches of male and female. This step tells the gender of number of persons identity.

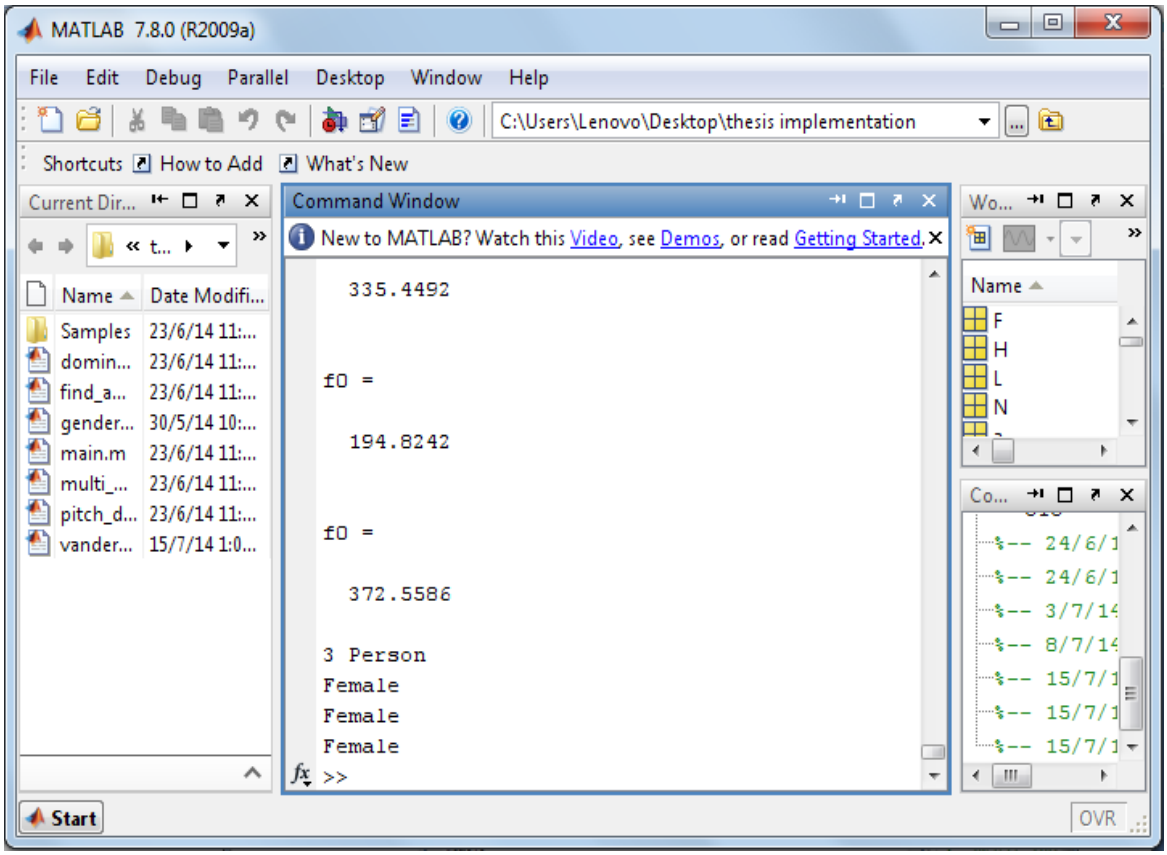

Figure 4.5: Final output of multi-f0.

4.4 Flow of Emotion Detection System

This flowchart depicts the overall working of system proposed in this thesis. As it is clear from the flowchart that a speech signal is taken as input for the system and then the system is initialized in order to perform the next step. Next phase is to apply the GTCC and GA algorithms. GTCC is used for feature extraction and GA is used for optimization. Both play a vital role in emotion recognition.

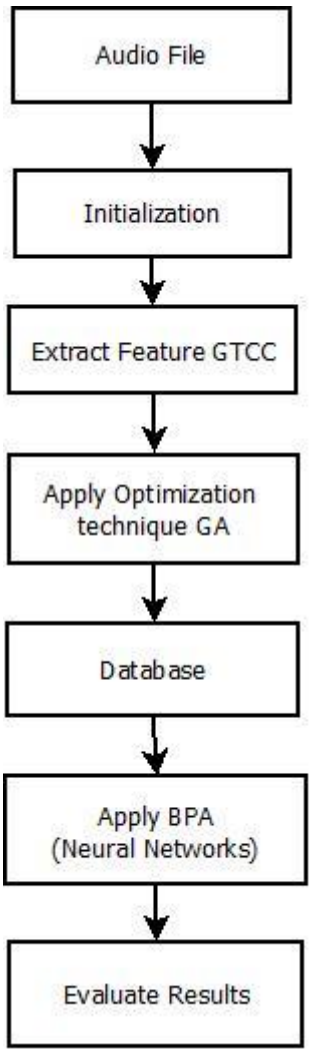

Figure 4.6: Proposed system of emotion Detection

Below snapshot show the main GUI of our Emotion detection System having different functionality like adding audio files to database, training neural network and then testing the system.

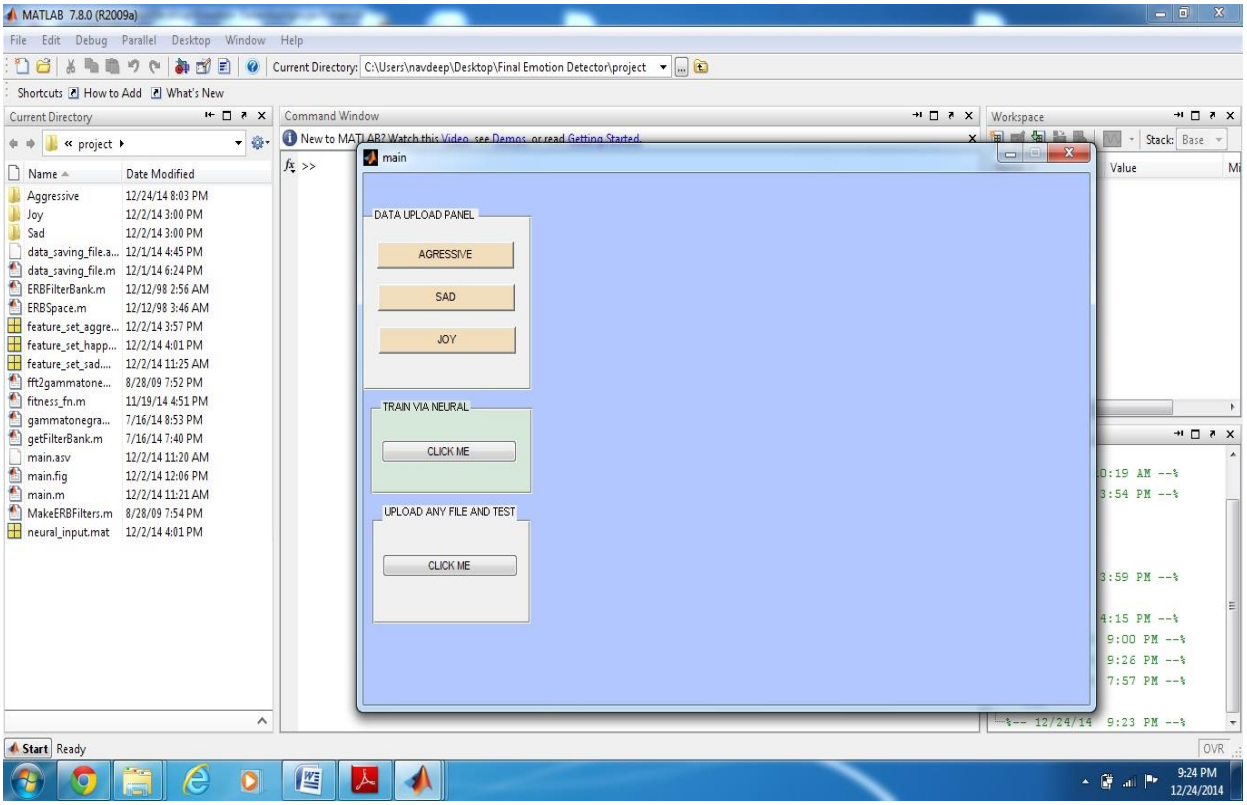

Figure 4.7: Main GUI of Emotion Detection

An input speech signal or voice is selected in order to train the neural network. This voice should be from any of three emotions.

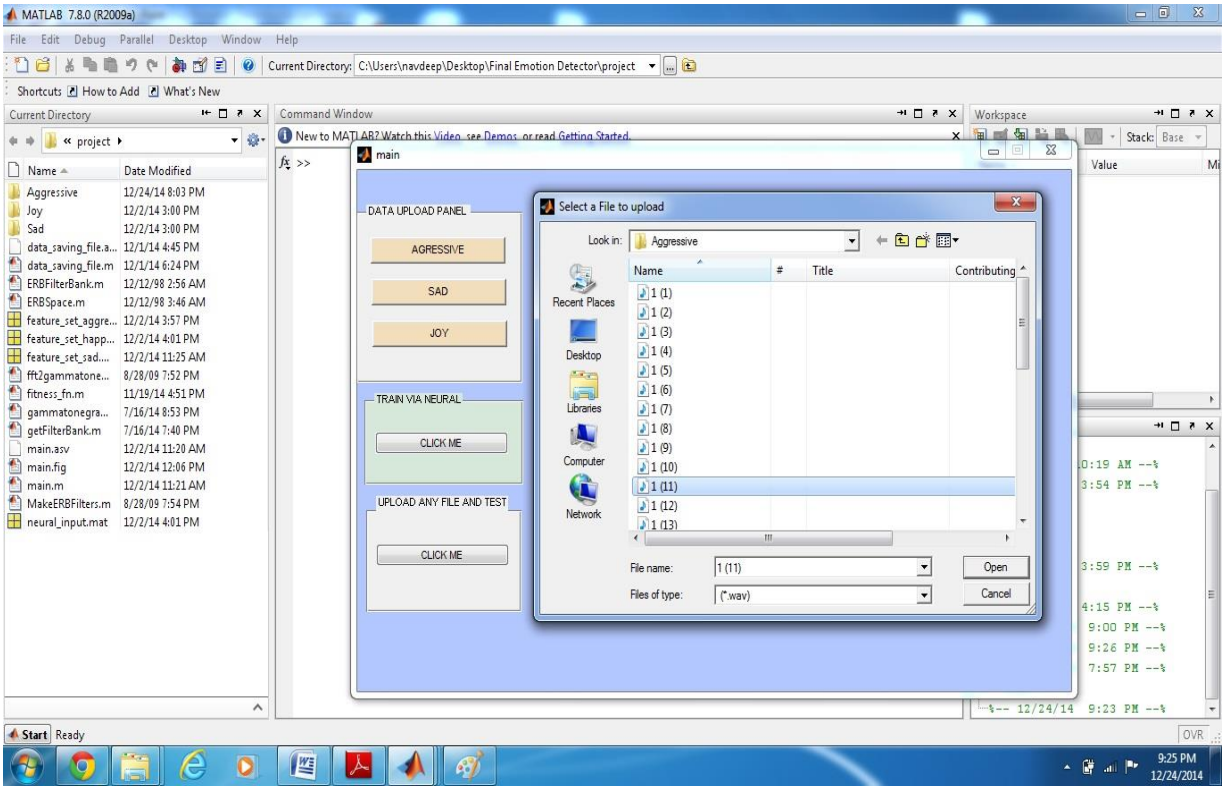

Figure 4.8: Selecting speech signal

This snapshot depicts the output value of audio sample which was added to the neural network in the previous step. This output value shows the time domain and frequency domain of the selected signal.

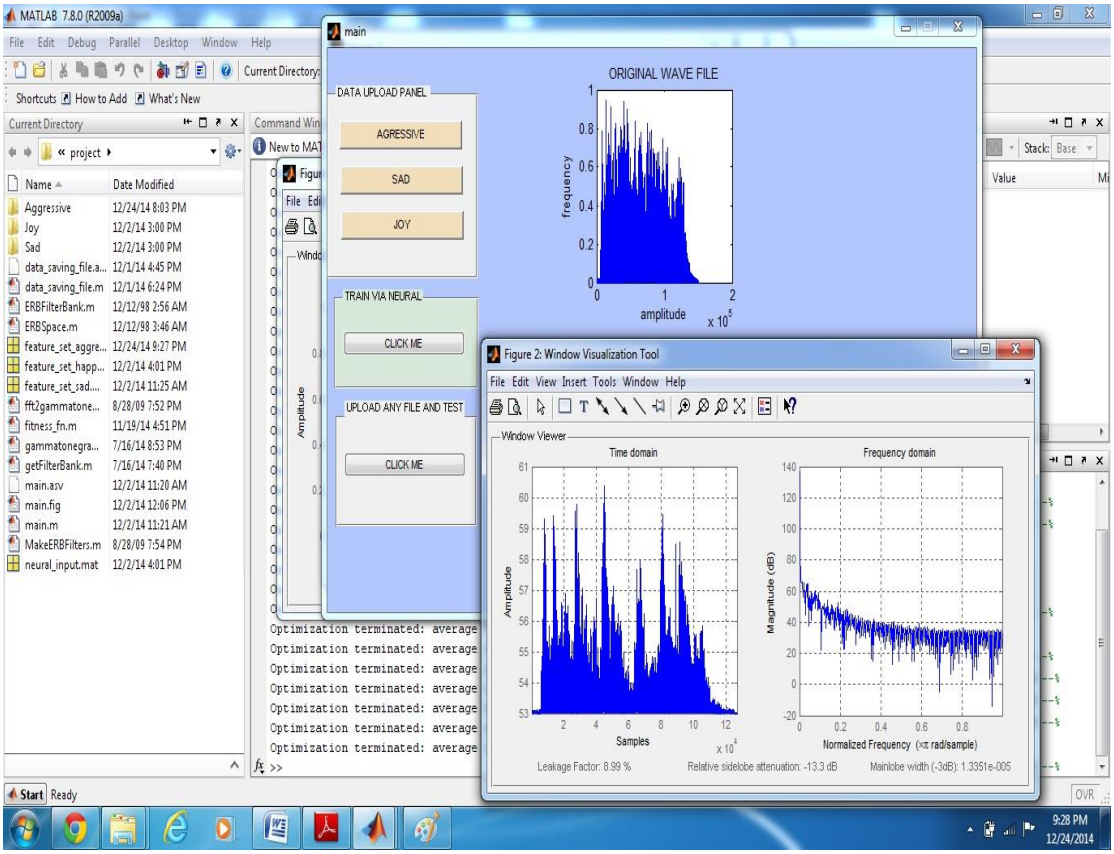

Figure 4.9: Showing time and frequency domain of signal

4.4.1 Gammatone Cepstral Coefficient (GTCC)

Figure 4.10: Block diagram of GTCC feature extraction technique.

Initially an audio signal is taken as input to the proposed system. This audio signal is then divided into short frames of size 10-50 ms. this process of dividing the signal into frames is known as windowing. Windowing is helpful in increasing the

efficiency of feature extraction process. Also short intervals contribute in assuming the non stationary audio signals as stationary ones.

After this each window of audio signals is converted into fast Fourier transform (FFT) form, which is further considered as input to frequency filters. Now the next step is GT filter bank which is the result of composition of various frequency responses. These responses are given by several GT filters. So, by passing the FFT form of signals through these GT filter bank, the required frequencies of sound signals can be emphasized. After getting the required frequencies, the log function is applied on these frequencies followed by the discrete cosine transform (DCT). These are applied to model the human loudness.

This snapshot shows the value of signal after applying FFT and GTCC algorithm o that audio signal. This snapshot also shows how the database is updated when an audio signal update the neural network.

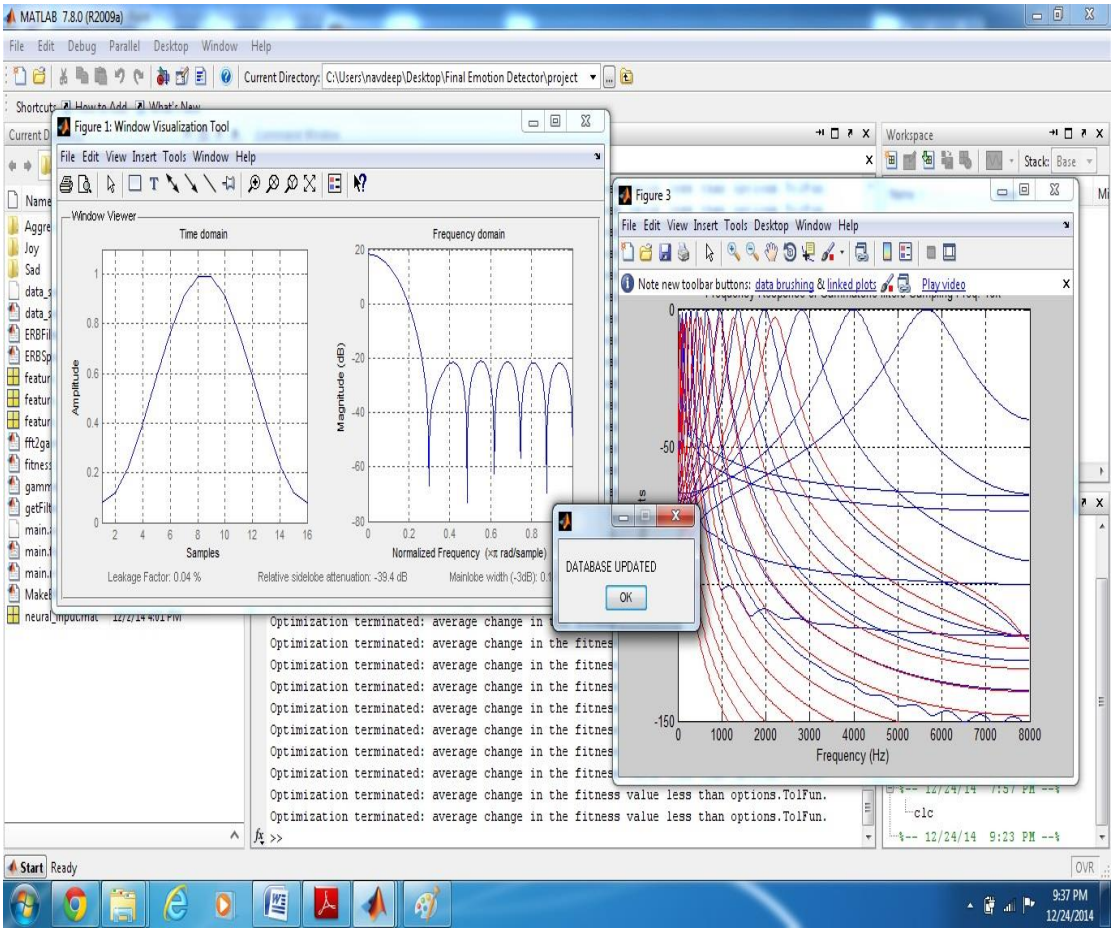

Figure 4.11: Signal after applying FFT and GTCC algorithm

4.4.2 Genetic Algorithm

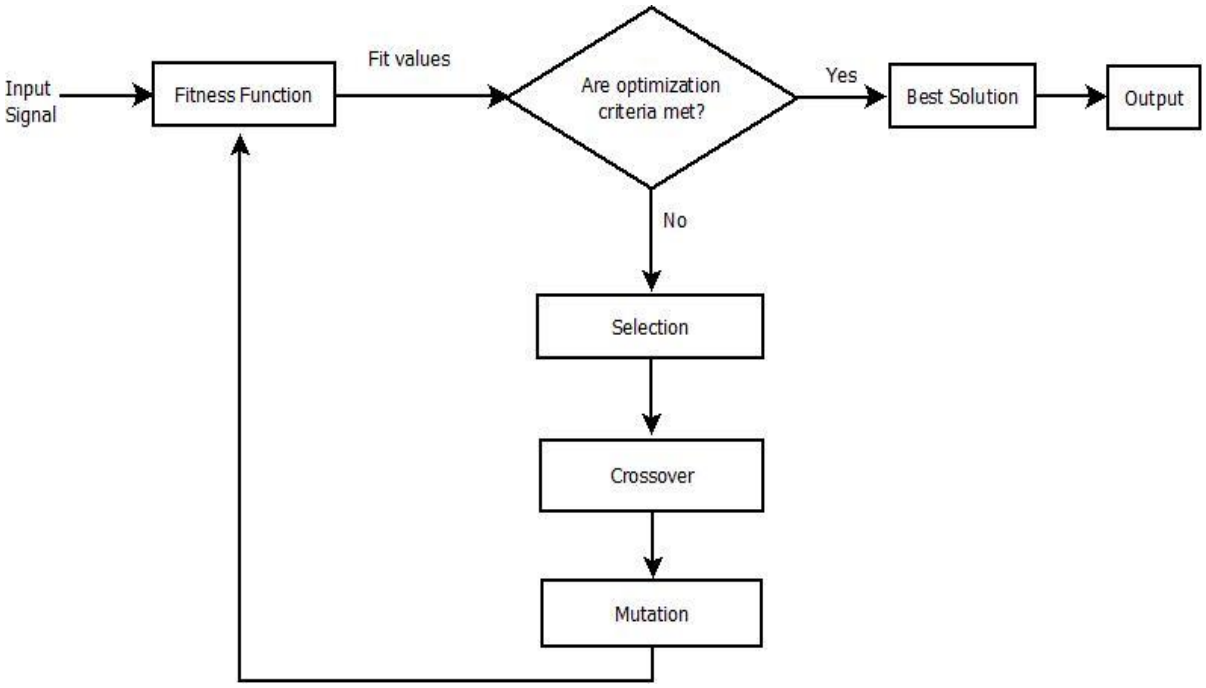

Figure 4.12: working of GA Algorithm

After applying the GTCC algorithm, now the GA algorithm is applied. GA is very much similar to natural selection. It is a type of search heuristic which is commonly used to find the solutions of search problems and optimization issues. GA algorithms are subpart of evolutionary algorithms which follows those methods which are inspired by inheritance, cross over, mutation and selection. Two things are mainly required by these genetic algorithms (GA): A solution domain and its genetic representation. Second one is a fitness function which is further used to evaluate the solution domain.

Input signals are considered for applying the fitness function. These input signals are actually the binary values of 01, 10 etc. Signals get these values as a result of feature extraction process. Now, fitness function is applied on these values in order to check whether they met optimized criteria or not. If yes, these are considered as best solution and if not then the genetic operators are applied on these input signals or binary values. As discussed above we have genetic operators such as selection, mutation and crossover. In these operators the binary values are shuffled and then

mutated again to get a different combination of values. Fitness function is again applied to these new values to check whether they satisfy the optimization criteria or not. This process is repeated again and again unless and until these input signals satisfy the optimization criteria and finally best solution is achieved.

Now these signals collaboratively made the database of proposed system, which is further used to apply back propagation algorithm and to provide training to the system.

4.4.3 Back Propagation Neural Networks Algorithm

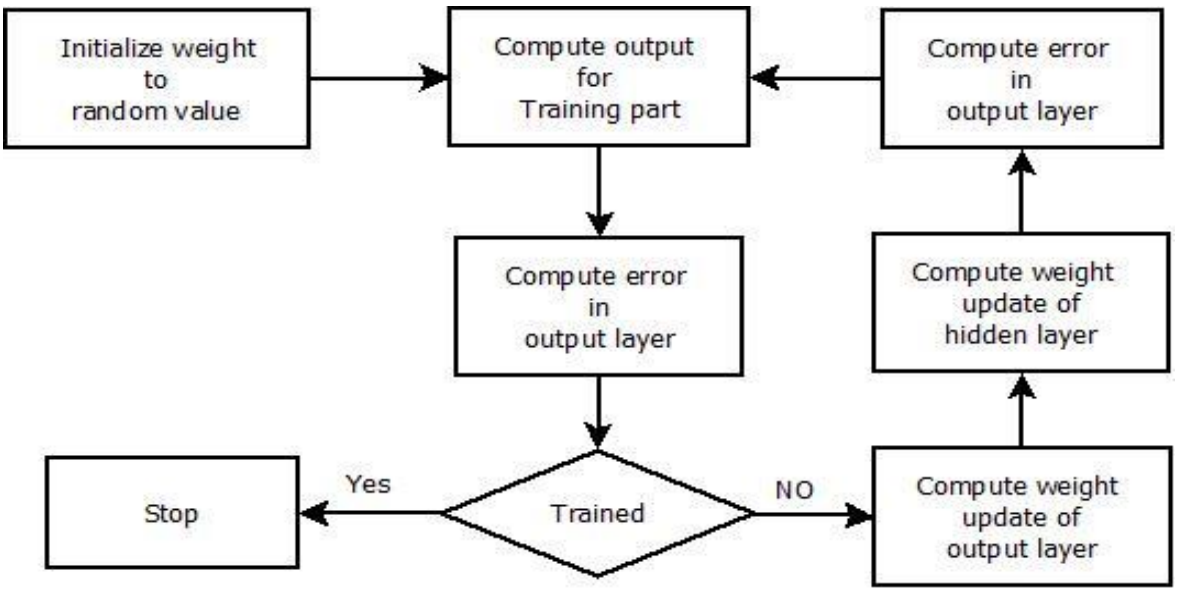

Figure 4.13: Back propagation algorithm.

This algorithm mainly takes multiple input signals and results in an output signal. Initially the database is used for creating a neural network using this algorithm. This back-propagation neural network contains various layers inside it. Main layers of this neural network are input layer, output layer and a hidden layer which lies in between the input and output layer. Each layer is attached to the next layer but there is no connection for moving in backward direction in between the PEs at the same level. Each PE has a predisposition input with non-zero weight The neural network is made up of P processing elements whose input/output function is defined as:

k = L (mW)

Where m = *{m,)* represents the input vector to the network,

k = {Kkj} represents the output vector from the network,

W is the weight matrix and is defined as:

$W = (w_i^T, wj^T, ..., Wn^T)^T$

Here the vectors w_i ,wj ,...,Wn are the individual PE weight vectors such as:

$$W_t = \begin{matrix} w_i \\ wj \\ wn \end{matrix}$$

So, basically the system is having a neural network generated at the training time. This neural network is generated from the database. Now when this algorithm is applied to a new signal then an output layer is generated. If there is any error, it is found only at the output layer. In order to remove that error, movement in backward direction is required. That is why this algorithm is named as back propagation algorithm.

Here the system is trained via neural network created using database.

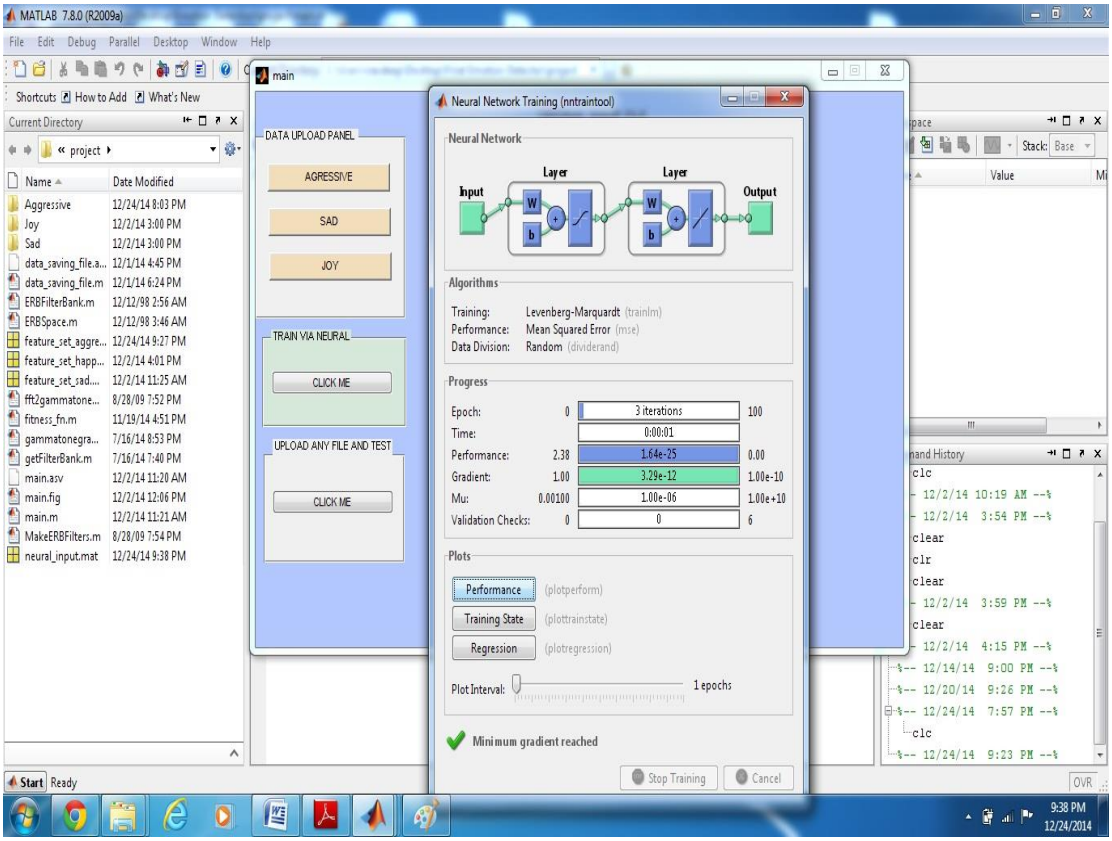

Figure 4.14: Training the system with neural networks

Here an audio signal is chosen for testing the proposed system. An audio file is selected and uploaded for testing. This snapshot shows the final output of system. Here the general features of input audio signal are shown along with the emotion detected by the system.

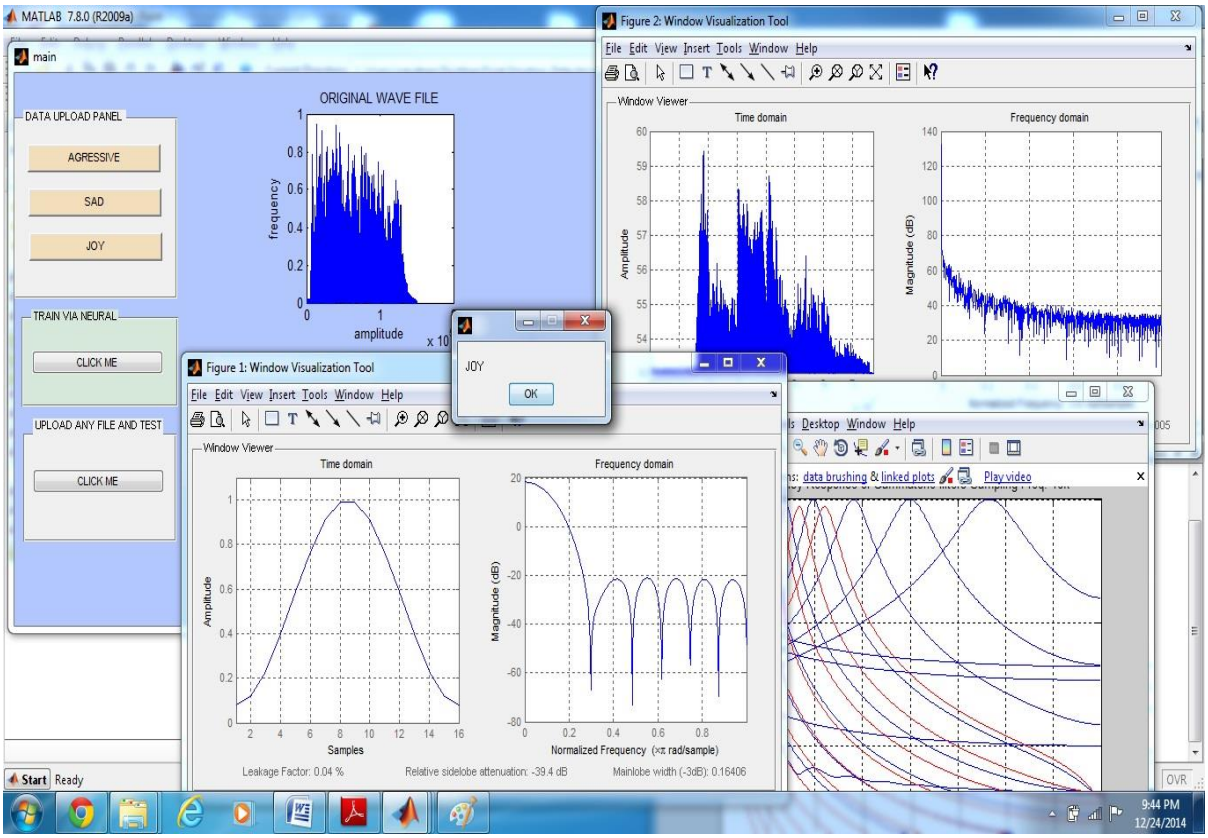

Figure 4.15: Final Output of emotion detection

Chapter 5

Experimental setup and results

Performance of proposed multi-f0 algorithm is evaluated on attest set of 135 mixed sounds. These 135 sounds consist of mixture of both male and female voice speech. Sampling of these speech signals is done at 8000 KHz originally. The FFT point is set to 16384 in order to achieve a high frequency resolution. Frame length is set to 25 ms.

Table 2 represents the performance evaluation of our test set. Groups contain multiple genders where male is represented by M and female is by F. It is observed from the results that as the number of person increase in the mixture audio signal its accuracy is going down for both persons as well as for gender detection. For one person or two person i.e. Group G0,G1,G2,G3 and G4 it show more than 90 % accuracy in person detection and more than 85% accuracy in gender detection. Gender detection accuracy for single person is near about 100%. When three people involves i.e. group G5, G6, G7 and G8 accuracy level drop down to 80% for person detection and 73 % to gender detection. It is also seen that it show good results when mixture is composed of different genders. Accuracy rate for similar gender is lower than that of different gender.

Table 5.1: Accuracy rates for mixture audio signals.

Group	Number of humans in mixed audio signal	Gender	Number of test cases	Numbers of times right person predicted	Numbers of times right Gender predicted	Accuracy Rate for Person detection	Accuracy Rate for Gender Detection
G0	1	M	15	14	15	93.33	100
G1	1	F	15	15	15	100	100

G2	2	M,M	15	14	13	93.33	86.66
G3	2	F,M	15	15	14	100	93.33
G4	2	F,F	15	14	13	93.33	86.66
G5	3	M,M,M	15	12	11	80	73.33
G6	3	M,M,F	15	13	12	86.66	80
G7	3	M,F,F	15	13	13	86.66	86.66
G8	3	F,F,F	15	13	11	86.66	73.33

M=male

F=female

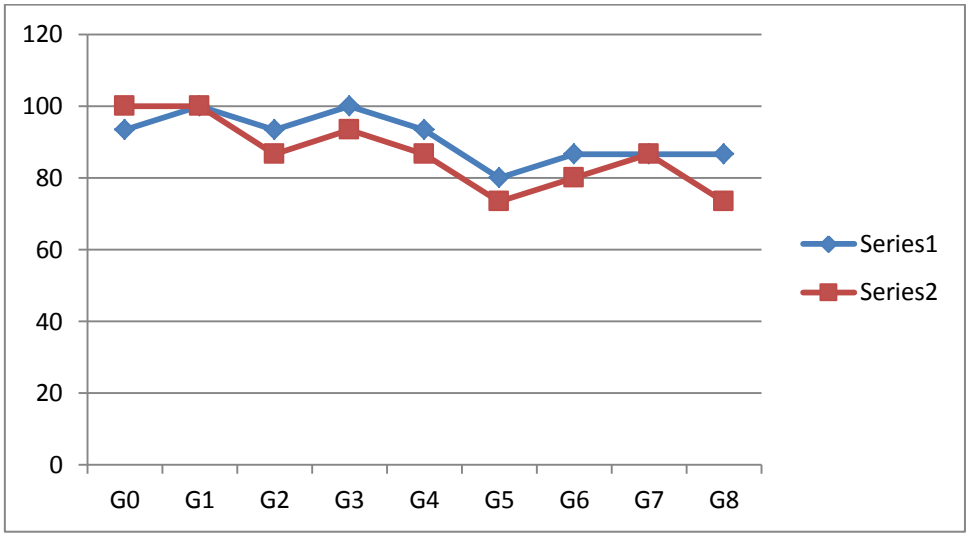

Figure 5.1: Accuracy graph

Series1 =Accuracy rate for person detection

Series2=Accuracy rate for gender detection

In order to check the accuracy of the proposed emotion detection system, the testing is done using in total 150 audio signals. 50 audio signals of each emotion (sad, joy, aggressiveness) are given as input to the proposed system. Results show that out of audio signals of sad emotion, the proposed system detects 43 accurately. Similarly in case of joy and aggressive emotions, 42 audio signals are correctively detected as joy signals and 44 are correctively detected as audio signals containing aggressive emotion.

Table 5.2: Accuracy rates for emotion detection.

CATEGORY	Number of test cases	Number of times right emotion detected	Number of times other emotion detected	Accuracy rate of emotion detection
SAD	50	44	6	88
JOY	50	45	5	90
AGGRESSIVE	50	46	4	92

Chapter 6
Conclusion and Future Scope

In thesis work, demonstration of a novel technique to determine multiple F0 in mixed signals and gender identification of those F0's. The basic idea of our approach is to identify the predominant pitch and then subtract it from the signal to get the residual signal for next iteration. Our approach results the number of persons in the signal along with their gender. The experiments results show that our approach performs well for human voiced signal and is able to identify their gender with a good accuracy rate. It shows higher accuracy for different gender mixture signal than similar gender mixture signal. But it only works for 3 persons mixture signal. For greater number its accuracy level drops down. In future we will try improving our person detection system so that it will be able to detect more than 3 people and try to improve our accuracy level for more than 3 people.

In second objective of our thesis the BPA algorithm is applied on input speech signals. Using this algorithm for detecting emotions in speech results in optimization and high accuracy. Two more algorithms play a vital role in achieving this optimization and accuracy rate named as GA and GTCC. Use of BPA algorithm with GTCC and GA results in 88 to 92 percent accuracy in classification.

The main limitation of this proposed system is that, this system needs to be trained with respect to every user whose speech is required to be identified in future. So, we will try to extend this research work to a level where no pre-training of system is required. Till now this proposed system only identify the emotions. There is no such system that identifies emotions along with the gender identification of the speech signal. We will try to recognize gender along with the spoken words in future.

References

[1] Human Voice. Available at: http://en.wikipedia.org/wiki/Human_voice.

[2] Fourier Transform. Available at: http://en.wikipedia.org/wiki/Fourier_transform

[3] The Fundamentals of FFT-Based Audio Measurements in Smart Live [Online]. Available at: https://www.rationalacoustics.com/files/FFT_Fundamentals.pdf.

[4] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Signal Processing*, vol. 88, no. 4, pp. 972–983, April 2008.

[5] M. Christensen and A. Jakobsson, "Multi-Pitch Estimation, *"Morgan & Claypool*, 2009.

[6] J. X. Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen, "A Robust and Computationally Efficient Subspace-Based Fundamental Frequency Estimator," *Transactions on Audio, Speech, and Language Processing, IEEE*, vol. 18, no. 3, pp. 487 –497, March 2010.

[7] Z. Zhou, H. C. So, and F. K. W. Chan, "Optimally Weighted Music Algorithm for Frequency Estimation of Real Harmonic Sinusoids," *International Conference on Acoustics, Speech and Signal Processing, IEEE*, Kyoto, Japan, March 25-30 2012.

[8] A. M. Noll, "Short-Time Spectrum and Cepstrum Techniques for Vocal-Pitch Detection," *J.A.S.A.*, vol. 36, no. 2, February 1964.

[9] A. M. Noll, "Cepstrum Pitch Determination,"*J.A.S.A.*, vol. 41, no. 2, 1967.

[10] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg and C. A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms", *ASSP*, vol. 24, no. 5, October 1976.

[11] M. Karjalainen and T. Tolonen, "Multi Pitch and periodicity analysis model for sound separation and auditory scene,"*IEEE*, 1999.

[12] R. Meddis and L. O'Mard, "A unitary model for pitch perception," *J. Acoust. Soc. Am.*, vol. 102, pp. 181 1-1820, Sept. 1997.

[13] R. Meddis and M. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: pitch identification," *J. Acoust. Soc. Am.*, vol. 89, pp. 2866-2882, June 1991.

[14] P. J. Walmsley, S. J. Godsill and P. J. W. Rayner, "Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters, "*Proc. 1999 IfiM Workshop on Applications & Signal Processing 10 Audio and Acoustics*, 1999.

[15] W. W. Zhao and T. Ogunfunmi, "Formant and Pitch Detection Using Time-Frequency Distribution," *International journal of speech technology*, pp. 35-49, 1999.

[16] J. S. Soler, R. JanC, J. A. Fiz2 and J. Morera2, "Towards automatic pitch detection in snoring signals," *Proceedings of the 22nd Annual EMBS International Conference*, Chicago IL, pp. 23-28, 200.

[17] A. P. Klapuri, "Multi pitch estimation and sound separation by the spectral smoothness principle", *IEEE*, 2001.

[18] M Wu, D Wang and G. J. Brown, "A Multi-pitch Tracking Algorithm for Noisy Speech", *IEEE Transaction on Speech and Audio Processing*, vol. 11, no. 3, November 2003.

[19] A. P. Klapuri, "Multiple Fundamental Frequency Estimation Based on Harmonicity and Spectral Smoothness", *IEEE Transaction on Speech and Audio Processing*, vol. 11, no. 3, November 2003.

[20] S. S. Abeysekera, "Multiple Pitch estimation of poly-phonic audio signals in a frequency-lag domain using the bispectrum", *IEEE*, 2004.

[21] Simon Godsill, "Bayesian Harmonic Models for the Musical Pitch estimation and Analysis", this work sponsored by the European research project MOUMIR. www.moumir.org.

[22] J. Wan, Y. Wu and H. Dai, "A Harmonic Enhancement Based Multi-pitch Estimation Algorithm", *Proceedings of ISCIT*, 2005.

[23] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "The multi pitch estimation problem: Some New Solution", *IEEE*, 2007.

[24] S. M. Kay, "Fundamentals of Statistical Signal Processing: Estimation Theory, "*Prentice-Hall*, 1993.

[25] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation using harmonic MUSIC,"*Rec. Asilomar Conf. Signals, Systems, and Computers*, 2006.

[26] X. Zhang, W. Liu, P. Li and B. Xu, "Multi-pitch Detection Based on Weighted Summary Correlogram", National Laboratory of Pattern Recognition, Beijing

[27] M.Y. Wu, D.L. Wang, and Guy J. Brown, "A Multi-pitch Tracking Algorithm for Noisy Speech*," IEEE Trans. Speech And Audio Processing,* Vol. 11, No. 3.

[28] R. Badeau, V. Emiya and B. David, "Expectation Maximization algorithm for multi pitch estimation and separation of overlapping harmonic spectra", *IEEE*, 2009.

[29] P. Mowlaee1, M. G. Christensen2, Z. -H. Tan1, and S. H. Jensen1, "A map criterion for detecting the number of speakers at frame level in model-based single-channel speech separation,"

[30] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modelling approach," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 45–66, Jan. 2010.

[31] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 77 – 93, Jan. 2010.

[32] E. Vincent, N. Bertin and R. Badeau, "Adaptive Harmonic Spectral Decomposition for Multiple Pitch Estimation", *IEEE Transaction on Speech and Audio Processing*, vol. 18, no. 3, March 2010.

[33] E. Benetos and S. Dixon, "Joint Multi-Pitch Detection Using Harmonic Envelope Estimation for Polyphonic Music Transcription", *IEEE Journal of selected topic in Signal Processing*, vol. 5, no. 6, 2011.

[34] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labelling sequence data," *Proc. 18th Int. Conf. Mach. Learn*, San Francisco, CA, Jun. 2001, pp. 282–289.

[35] A. Koretz and J Tabrikian, "Maximum A Posteriori Probability Multiple-Pitch Tracking Using the Harmonic Model", *IEEE Transaction on Speech and Audio Processing*, vol. 19, no. 7, September 2011.

[36] Q Huang and D Wang, "Multi-Pitch Estimation for Speech Mixture Based on Multi-Length Windows Harmonic Model", *Proc. Of IJCCSO*, 2011.

[37] S. I. Adalbjornsson, A. Jakobsson, and M. G. Christensen, "Estimating multiple pitches using block sparsity", *IEEE*, 2013.

[38] W. Gevaert, G. Tsenov and V. Mladenov, "Neural Networks used for speech Recognition," *Journal of Automatic Control*, Vol. 20, pp. 1-7, 2010.

[39] N. Pushpa, R. Revathi, C, Ramya and S. Shahul Hameed, "Speech processing of Tamil Language with Back Propagation Neural Network and Semi-Supervised Training", *International Journal of Innovative Research in Computer and Communication Engineering*, Vol.2(1), 2014.

[40] Y. Pan, P. Shen, and Liping Shen," Speech Recognition Using Support Vector Machine", *International Journal of Smart Home*, Vol. 6, No. 2, April, 2012.

[41] A. Utane, S.L Nalbalwar, "Emotion Recognition Through Speech Using Gaussian Mixture Model And Hidden Markov Model", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 3, Issue 4, April 2013.

[42] A. Sapra, N. Panwar, and S. Panwar , "Emotion Recognition from Speech", *International Journal of Emerging Technology and Advanced Engineering*, Volume 3, Issue 2, February 2013.

[43] Audacity Tool [Online]. Available at: http://manual.audacityteam.org/o/.

List of Publication

Navdeep Kumar, Ravinder Kumar, "A Novel Approaches for Multi-Pitch Detection with Gender Identification," *IEEE First International Conference on Networks and Soft Computing*, IEEE, (ICNSC-2014), 2014 (Status: Accepted)