MULTI-WINDOW COMPARISON OF SIR PERFORMANCE IN EXTRACTION OF MONO-AURAL VOCAL AND NON-VOCAL COMPONENTS IN REPET

Thesis submitted for fulfilment of the requirements for the degree of

Master of Technology

In

Electronics and Communication Engineering

By

VANSHA KHER

Enrol. No. 132006

Under the supervision

of

Prof. Dr. T.S Lamba (Dean, Academics and Research)

MAY, 2015

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT, SOLAN – 173234,

INDIA.

CERTIFICATE

This is to certify that the work reported in the M.Tech thesis entitled "**MULTI-WINDOW COMPARISON OF SIR PERFORMANCE IN EXTRACTION OF MONO-AURAL VOCAL AND NON-VOCAL COMPONENTS IN REPET**" which is being submitted by ***Miss. Vansha Kher*** in the **Department of Electronics and Communication, Jaypee University of Information Technology, Wakhnaghat, Solan, India** is a bonafide record of her original workcarried out under my supervision. This work has not been submitted elsewhere for any other degree or diploma.

Prof. T.S Lamba

Professor, Dean (Academics)

Department of Electronics and communication Engineering

Jaypee University of Information Technology (JUIT)

Wakhnaghat, Solan (H.P) – 173234,

India.

DECLARATION

I hereby declare that the work presented in this project has been carried out under the supervision of Prof. Dr. T.S Lamba, Dean (Academics), Department of Electronics and Communication Engineering, Jaypee University of Information Technology, Waknaghat, Solan-173234, Himachal Pradesh, and has not been submitted for any degree or diploma to any other university. All assistance and help received during the course of the investigation has been duly acknowledged.

Vansha Kher

M.Tech (ECE)

ACKNOWLEDGEMENT

The Indian tradition recognizes three kinds of non-repayable debt to the parents, to the teacher and to the spiritual guides. Therefore, words are not enough to express my sincere gratitude and indebtedness to my teachers who have inspired me to the subject. I would like to acknowledge them with great appreciation.

I would like to express my deepest gratitude to my advisor ***Prof.T.S Lamba,*** *Dean of Academics, J.U.I.T Wakhnaghat, Solan, for giving me an opportunity to work with him. I thank him for his excellent guidance and continual support during the course of my Master's at J.U.I.T. Working with him has been a very wonderful productive experience. His advice and his wide knowledge that he has shared during my association with him has been invaluable. I am very thankful for the support and encouragement he extended to me and the freedom to express my views.*

I am also thankful to ***Prof. Tapan Kumar Jain,*** *Assistant Professor, J.U.I.T, for his constant advice and all other staff members of the Department of Electronics and Communication, J.U.I.T, for their generous help in various ways for the completion of my thesis.*

I am thankful to my senior and mentor Amit Bhan and my school friends Nipun, Shalu, Akshay, Ninni, Manish, Vishal and Suraj and to my cousins Versha, Mayank, Aryan and Ananya, who are indeed an inseparable part of my life. They have given my courage, encouragement and support through various stages of my project and have been great sources of inspiration to me.

I am deeply indebted to my parents, grand-parents and most importantly, to my younger brother Shiva who have been a constant source of regard, inspiration, support, love, encouragement and affection to me.

To all acknowledged, I solemnly owe my work.

Place: *Vansha Kher*

Date: *M.Tech (ECE)*

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

GSM	Global System for Mobile Communications
POTS	Plain Old Telephone Service
ADSL	Asymmetrical Digital Subscriber Line
PCM	Pulse Code modulation
PSTN	Public Switched Telephone Network
EM	Electro-Magnetic
VF	Voice Frequency
CASA	Computational Auditory Scene
DTFT	Discrete Time Fourier Transform
STFT	Short-Time Fourier Transform
LFPC	Log Frequency Power Coefficients
PLPC	Perceptual Linear Predictive Coefficients
PLCA	Probability Latent Component Analysis
GMM	Gaussian Mixture models
HMM	Hidden Markov Model
TF Mask	Time-Frequency Mask
BPM	Beat-Per-Minute
FFT	Fast Fourier Transform
REPET	REpeating Pattern Extraction Technique
MIR	Music Information Retrieval
ICA	Independent Component Analysis
MFCCs	Mel-frequency cepstrum coefficients
NMF	Non-Matrix factorization
DB	Decibels
SIR	Signal to Interference Ratio
ANOVA	ANalysis Of VAriance

TABLE OF CONTENTS

ABSTRACT

The vocalized form of human communication is speech. In linguistics (articulatory phonetics),normal human speech is said to be produced with pulmonary pressure that are created by the lungs , thereby creating phonation in the glottis and larynx , which is then modified by vocal-tract to generate different vowels and consonants. Speech is composed of following three parts: Articulation, Voice and Fluency. The message or information that gets communicated through speech is intrinsically of a discrete nature; i.e. it can be designated by a concatenation of elements from a finite set of symbols. Phonemes are the set of symbols from which every sound can be classified. Thus, phonemes are the basic units of language phonology, which are usually combined with other phonemes to form meaningful units called Morphemes. The Audio signals can be classified as the class of sounds that pursue same frequency as that of human auditory range. The separation of vocals and music has evolved as an extremely quintessential area to be resolved in Automatic Karaoke, vocalist identification and audio pre-processing. The distinction of the lead varying vocals from the background music in an audio recording is an extremely demanding and exigent task. The speech-separation research usually inculcates Time-frequency masking technique that ultimately appraises the hearing-aid design. The core principle in music which is capitalized to discriminate underlying non-vocals from vocals (speech) is Repetition. The rudimentary principle in the field of Music Information Retrieval (MIR) is 'REPETITION', as premise of music, as an art. The 'Repetition' feature is especially enacted for pop songs where the singer often overlays frequently changing vocals on a periodically repeating background in a mixture. The basic approach of dissertation is the recognisation of periodically repeating segments in audio excerpts, analogize them with a repeating model and finally discrimate the repeating musical patterns via Time-Frequency masking. A TF mask is grounded on the basis of TF representation of any signal. In this project, the quality of foreground vocals and accompanying background can be analyzed in terms of SIR (Signal to Interference Ratio) value utilizing 'ANOVA' (Analysis Of Variation) computational method on different genres of musical audios and formulated the complete comparison of SIR values using hamming , hanning and blackmann windows using the software tool 'MATLAB' and concluded that separation of mono-aural vocal and non-vocal components applying blackmann window shows better SIR values and separation quality as compared to hanning and hamming windows.

Chapter 1: INTRODUCTION

1.1 Introduction

The purpose of speech [1] is communication. Speech is basically the signal that consists of changes or variations in pressure coming out from the mouth of a speaker. These pressure variations then propagate as waves though the air medium and will enter the ears of the listener who will finally decipher the waves as a received message. Gestures that are included in human communication are not part of the speech. Therefore, speech can be characterised in terms of the signal that carries the information as changes in pressure signal called as "Acoustic Waveform". The transfer of messages from one person to another via speech is called speech communication. The chain / series of events that commence from the origination or concept of a message in speaker's brain to the arrival of message in listener's brain are denoted as: Speech Chain [2]. The information that gets communicated through speech is generally of a discrete nature, called as Phonemes. Every language has its own distinctive set of phonemes, typically numbering between 30 and 50. English can be represented by a set of around 42 phonemes. It is based on the syntactic combination of lexical and names drawn from very large vocabularies. The two major concern of any speech communication system are the preservation of message content in the speech signal and the representation of speech signal in the form that is flexible so that modifications can be made in the speech signal without degrading the content of message.

Audio and Speech Processing [1-2] have steadily gained importance in the everyday life of enormous people in developed countries. From 'Hi-fi' music systems ,through radio to portable music players, audio processing is firmly entrenched in bestowing entertainment to consumers .Now-a-days , with the advent of CD players , Internet Radio , MP3 players and i-pods; digital audio techniques in particular have gradually attained a domination in audio delivery. Digital audio processing has continually subjugated even in the field of television and film studios; and in mixing desks for 'live' events. Myriad Sound effects and music are even becoming are even becoming more predominant within computer games. Speech Processing has equally experienced an upward world-wide trend, with the sudden rise and expansion of cellular communications especially the European GSM (Global System for Mobile Communications) standard. GSM has now been the most ubiquitous technology world-wide and also envisaged tremendous adoption even in the world's poorest nations.

Certainly, speech has been transmitted digitally over long distances, especially in satellite communication links; but even the legacy telephone network (named POTS

for Plain Old Telephone Service) is now succumbing to digitization in various countries. Several hundred metres of twisted pair copper wires running to a customer's home, was never designed or deployed with digital technology in mind. However, with DSL (digital subscriber Line), even this analogue twisted pair cable will convey high speed digital signals .The recently developed technology of ADSL (Asymmetrical Digital Subscriber Line) has allowed the rapid growth of telephony services such as Skype , that can even broadcast digitized speech.

- **SPEECH PROCESSING [3] :** In the case of speech signals, the information source is the human speaker. The measurement or observation is generally the acoustic waveform. Signal processing then involves representing a signal based on some given model and then the application of some higher level transformation is done to put the signal in a more convenient form. The concluding step is the extraction and utilization of the message information.

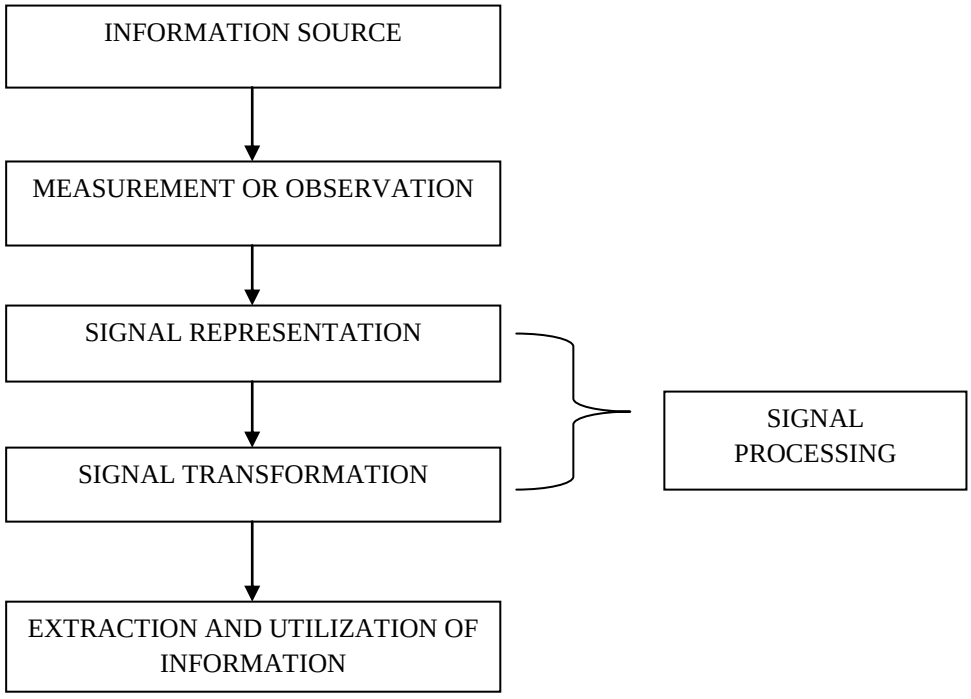

Fig. 1.1 (a) General View of information manipulation and processing (Ref: Rabiner)

- **DIGITAL SPEECH PROCESSING AND ITS APPLICATIONS**

The representation of signals in digital form is of fundamental concern. In this regard, a band-limited signal can be represented by samples that are taken periodically in time. The discrete representations of speech are classified into two broad groups, namely Waveform and Parametric representations [4]
Waveform representations of speech signals are concerned with primarily retaining the "wave shape" of the analog speech signal through a sampling and quantization process. On the other hand, parametric representations deal with representing the speech signal as the output of a model for speech generation.

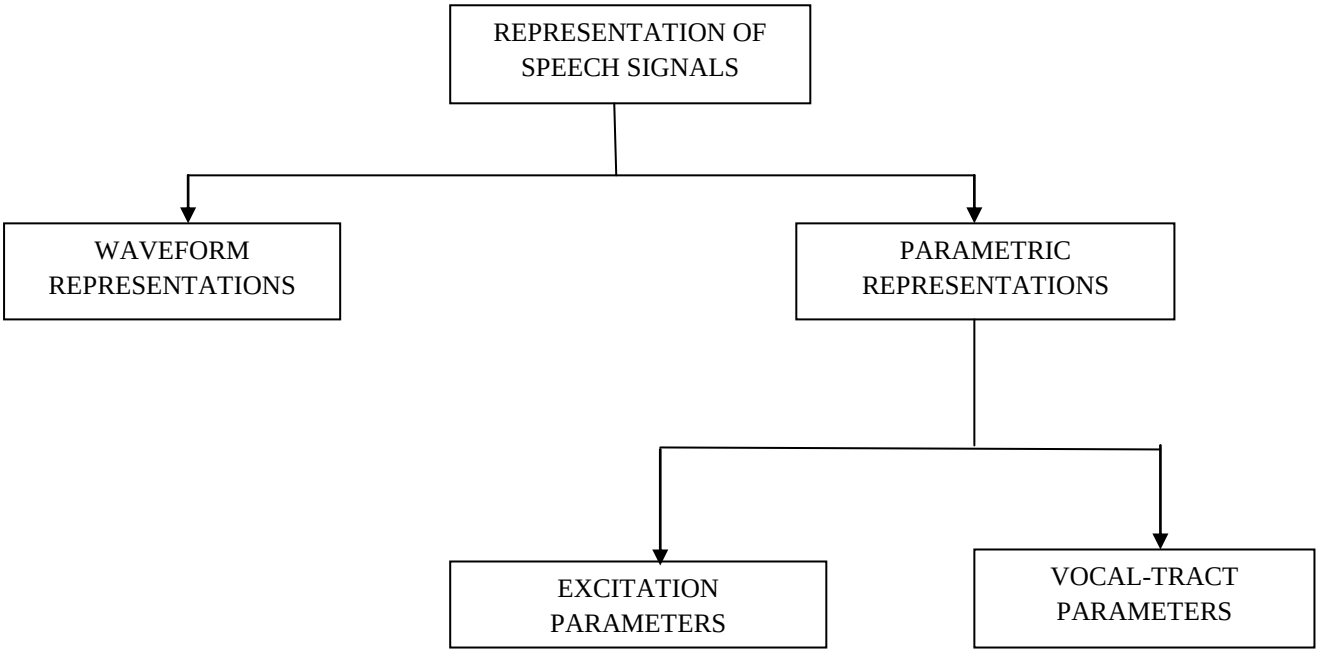

Fig. 1.1 (b) Representations of Speech signals. (Ref: Rabiner)

➢ APPLICATIONS OF SPEECH COMMUNICATIONS

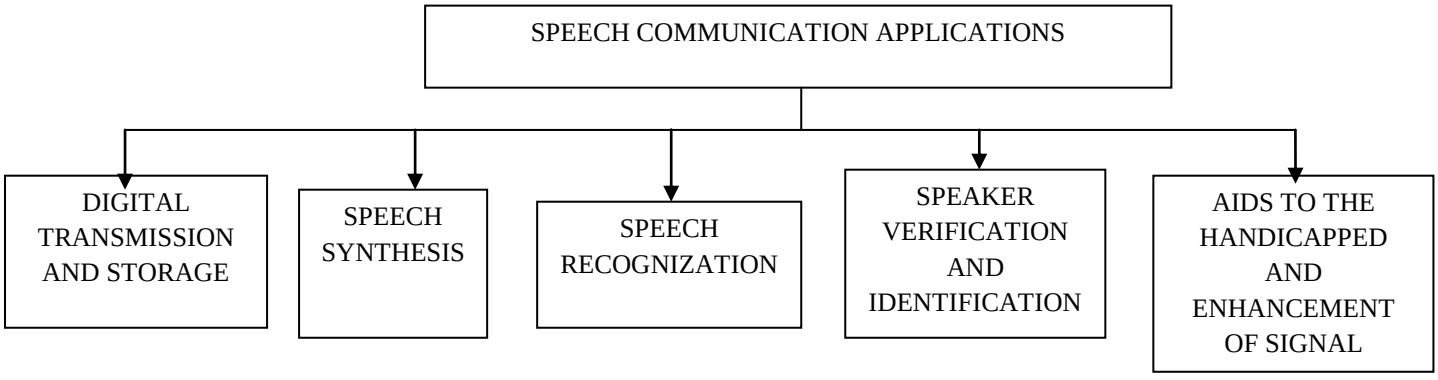

Fig.1.1(c) Some typical speech communications applications. (Ref: Rabiner)

1.1.1 THEORY OF AUDIO, SPEECH AND MUSIC SIGNALS

An audio signal can be enunciated as a sound, typically as an electrical voltage. Audio signals have frequencies in the audio-frequency range of typically 20 Hz to 20 KHz (Human range of hearing). Audio signals can be synthesized directly or can be originated at a transducer such as a microphone, musical instrument peer-up, tape head or phonograph cartridge. Headphones or Loud-speakers usually transform electrical audio signal into sound. In other words, the audio signals can be defined as sound waves – longitudinal waves through air, consisting of compressions and rarefactions .The unit for measuring these audio signals will be bels or in decibels.

An Audio track or an Audio channel [5] can be defined as an audio signal communication channel in any storage device.

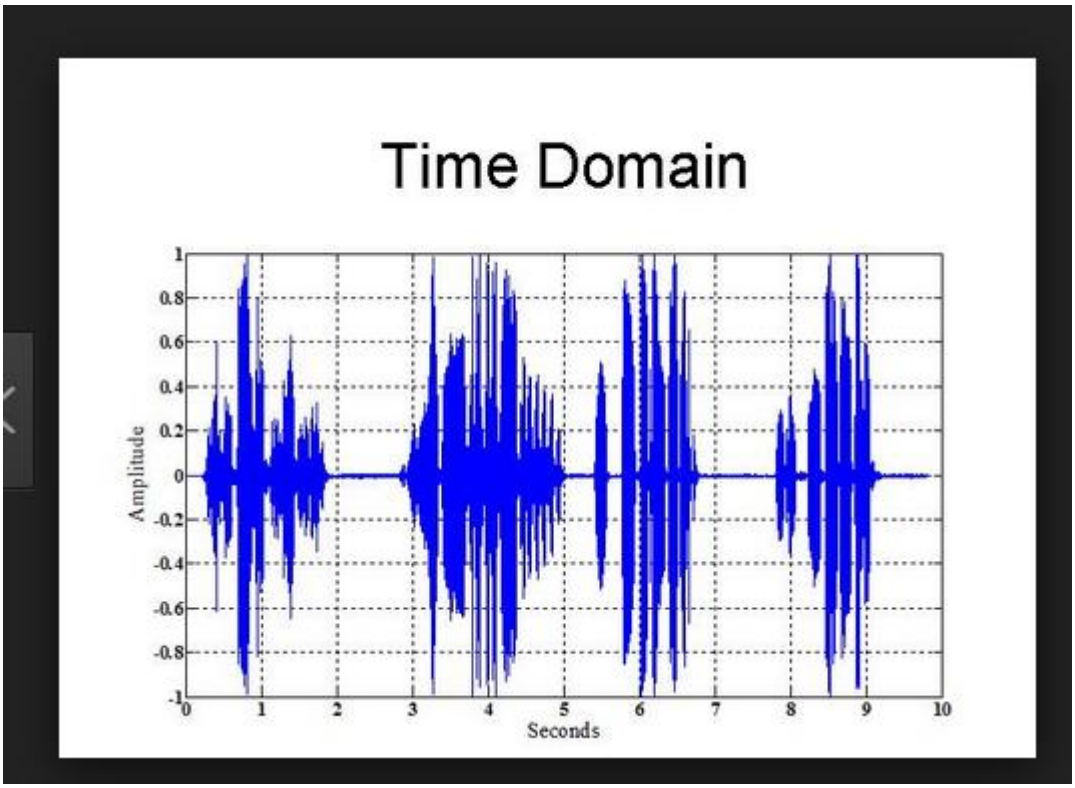

Fig.1.1.1 (a) Representation of Audio signal in Time-domain (Ref: Applied Audio and Speech Processing)

The basic incitement of speech is communication. That is, the vocalised form of human communication is “SPEECH”; relied on the syntactic combination of lexical and names that are usually drawn from a large set of vocabulary. The one-dimensional signal that comprises of changes or variations in pressure emerging out from the mouth of a speaker. These pressure variations propagate as waves through air and reach the ears of the listener, who will decrypt these waves into a received message.

According to the concepts of information theory, speech signal can thus be represented in terms of its message content or information. In general, speech can be signalized in order to carry the message information as pressure variations, named as 'acoustic waveform'.

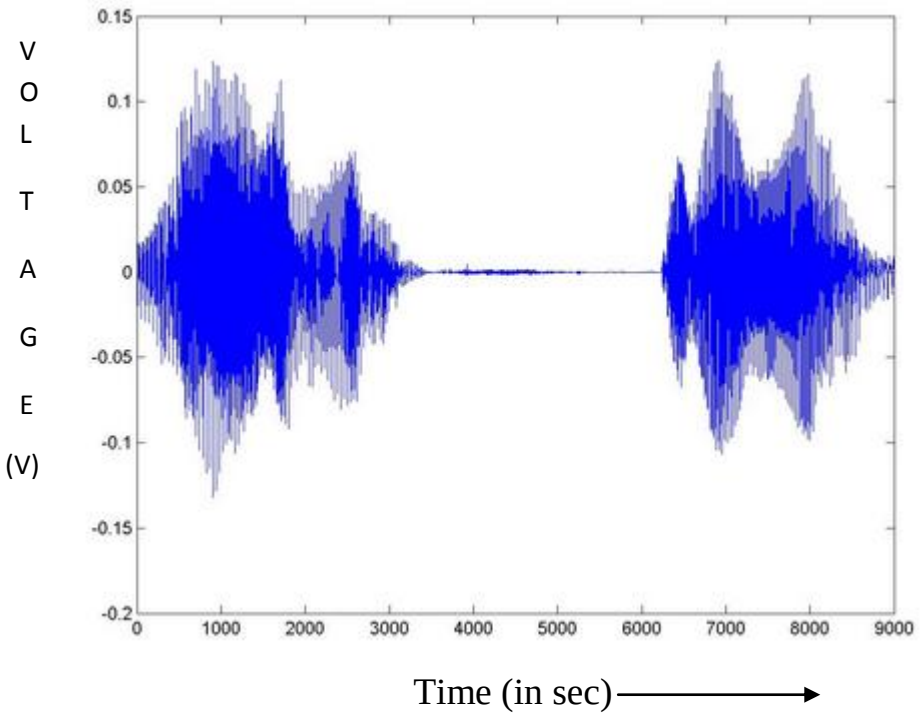

Fig. 1.1.1(b) Time domain representation of Speech Signal (Ref: Applied Audio and Speech Processing)

- **FREQUENCY RANGE OF AUDIO, SPEECH AND MUSIC SIGNALS**

In telephony or POTS (Plain Old Telephone Service); the usable voice frequency band ranges from 300 Hz to 3,400 Hz approximately. It is for the reason that the ultra low frequency band of the Electro-magnetic (EM) spectrum between 300 to 3000 Hz is also referred to as 'Voice Frequency' (VF) [4-6] ; being the EM energy that represents acoustic energy at base-band . The bandwidth that is assigned for a single voice frequency transmission is 4 KHz, with the inclusion of guard bands, with a sampling rate of 8 KHz to be applied at PCM system used for digital PSTN (Public Switched Telephone Network). Because a sampling rate of 8 KHz is used in order to avoid aliasing to be used at PCM (Pulse Code Modulation System) for digital PSTN. Voiced speech has a fundamental frequency "Fo" of 85-180 Hz.

Music is defined as an art form that nominates sound as its medium. The common elements for music consists of pitch (melody and harmony), sonic qualities of timbre and texture and rhythm associated concepts. The range of a musical instrument is the distance from its lowest to the highest pitch that it can play.

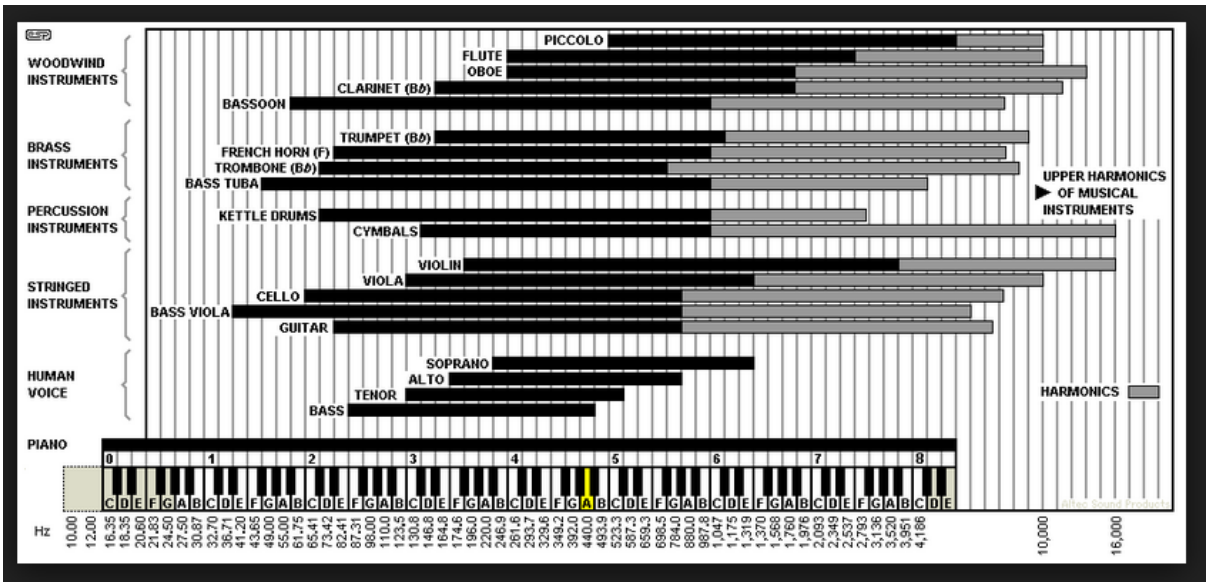

Fig. 1.1.1 (c): Comparative ranges of human voice and Musical instrument Frequencies (Reference: Rockford Fosgate)

1.2 **MOTIVATION**

The use of Time –Frequency masking is a new approach for the separation of speech from speech-in –noise mixtures. In our thesis, Time –frequency mask is computed using the computational auditory scene analysis techniques (CASA) [3-6].The separation of voice and music in an audio clip has several advantages in the field of digital audio processing like in the area of vocalist identification in chorus, Pitch estimation in songs, melody estimation and extraction, audio remixing, in automatic karaoke systems and in voice or music transcriptions. The main motivation of using separation and extraction techniques is for hearing aid design, thereby, showing its effectiveness in improving human speech recognisation in noise.

1.3 **THESIS OUTLINE**

The organisation of the thesis can be as follows:

Chapter 1 covers the Introduction part describing the digital signal processing, scope and research areas of Digital Audio, Speech and Music processing along with the frequency ranges of Audio, Voice Frequency and Musical Scale of different musical instruments.

Chapter 2 provides a brief idea on the human speech production and hearing process and discusses on various time-domain signal processing techniques inculcated in order to create an insight into the behaviour of speech and audio signals. The general outline about handling audio and speech in MATLAB has been demonstrated.

Chapter 3 creates an insight about the related works regarding the mono-aural separation of music and vocal structures in audio excerpts. The limitations of each proposed work has been illustrated and a general outline regarding the proposed work in dissertation has been discussed.

Chapter 4 emphasises on various techniques and algorithms that has lead to the formulation of a new approach and an extension to the REPET (Repeating Pattern Extraction Technique) on the basis of SIR (Signal - to - Interference Ratio) ,inculcating the ANOVA (ANalysis Of Variance) Technique using the audio processing steps carried out and simulated in MATLAB .

Chapter 5 formulates the Results, illustrating the better performance evaluation in terms of SIR values for hamming, hanning and blackmann windows. Subjective tests regarding the separation quality has been charted out.

Chapter 6 covers the Conclusion, future -work followed by the references.

Chapter 2: TIME – DOMAIN APPROACHES FOR SPEECH PROCESSING

Speech is the communication transfer from one speaker to one or more listeners. Vocal system is responsible for production of speech in humans. The speaker generally produces speech signal in the form of pressure waves that travel through air from speaker's mouth to listener's ears. Speech signals consists of variations in pressure as a function of time and is measured directly in front of speaker's mouth(primary sound source).These changes in pressure propagate through the air medium, the amplitude variations correspond to deviations in atmospheric pressure caused by these travelling waves. The nature of speech signal is non-stationary (time-varying) as the muscles of vocal-tract contract and relax.

Speech can be divided into 'Sound Segments' that share certain acoustic and articulatory properties; for a short duration of time. In order to produce each sound with message, there is a proper positioning of articulators such as: teeth, tongue, vocal-folds, cords, lips, jaws and velum.

2.1 SPEECH PRODUCTION IN HUMANS

Speech is produced in humans, as air is exhaled. Changes in articulatory positions generally influence the pulmonary egressive air-stream (air is expelled out from the lungs). Speech production in humans can be viewed as 'filtering' operation in which a sound source excites a vocal-tract filter. The voicing sound source occurs at larynx (at base of vocal-tract) where air-flow can be interrupted periodically by vibrating vocal-folds (cords). The sound source (larynx) can generate periodic vibrations leading to voiced speech and can generate noisy and a periodic speech, leading to unvoiced speech. The pulses of air that are generated by abduction and adduction of the vocal-folds create a periodic excitation for the vocal-tract .Here, a point should be noted that the air-pulse volume (Vs Time) resembles half a sine-wave, with the glottal closure more abrupt than its opening. The main aspect of naturalness of human-speech is small deviations from a periodic, smooth pulse waveform due to non-linearity in vocal-tract.

For both voiced and unvoiced speech/excitation, vocal-tract acts as a filter and amplifies certain sound frequencies while attenuating others. Harmonics are energy concentrations at multiples of fundamental frequency (F_o) [7]. Voiced Speech, being periodic has spectra consisting of harmonics of fundamental frequency (F_o). Unvoiced speech is noisy mainly due to the periodic and random nature of the signal, generated at a narrow constriction in vocal-tract. Generally, speech signals are quasi-periodic, due to random excitations in vocal-tract.

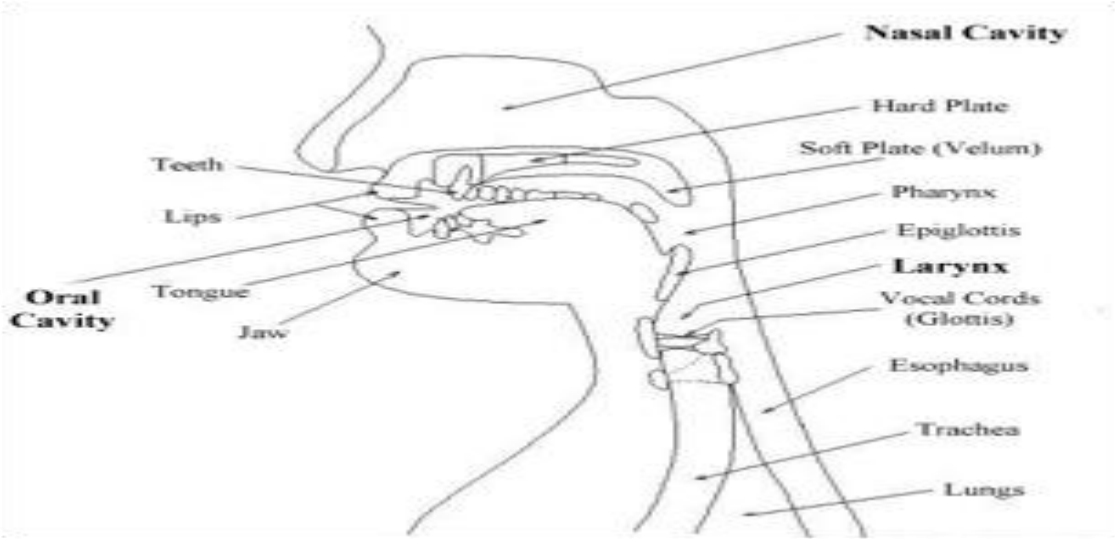

Fig.2.1: Sectional diagram of human –vocal apparatus. (Ref: Porticus Technology)

Source of most speech sounds is larynx where vocal-folds can obstruct air-flow from the lungs. The cartilages and membranes connect the lungs to the vocal-tract through a passage called as Trachea. The function of Epiglottis is to cover the larynx. Lungs provide air-flow and pressure source for speech. Vocal-folds modulate the air-flow for production of sound or speech.

a. Lung power mostly affects the volume of the volume of the sound, but rapid variations often distinguish a boundary between syllables.

b. If the glottis is closed temporarily during speech, a glottal stop results.

c. Vocal-chord muscle tension causes the chord to vibrate at different rates, forming the pitch frequencies. Voiceless sounds where the vocal-chords don't vibrate have little or no pitch structure.

2.2 HEARING PROCESS IN HUMANS

Listening is a process in which the signal (acoustic speech) entering the listener's ear is converted into a linguistic message. The Listening process consists of two parts: Audition or hearing that will register the speech sounds into the brain and the Sound perception that will decode the speech message from neural representation of sounds. Physcoacoustics is the study of auditory perception at psychological level, relating acoustic signals to what the human ear perceives.

STRUCTURE OF HUMAN EAR:

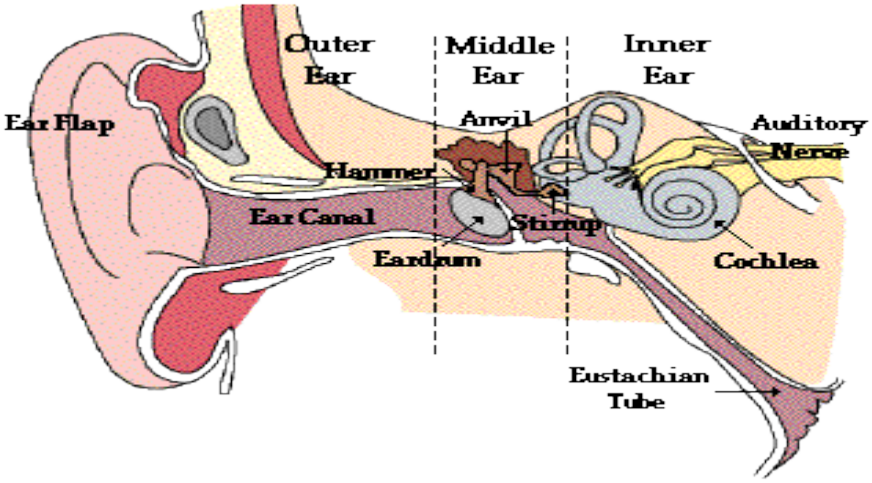

Fig.2.2: Sectional diagram of human –ear. (Ref: applied speech and audio processing forum)

The Ear is especially responsible to those frequencies in the speech signal that contain the most information relevant to communication. The ear consists of three main parts: Outer ear, Middle Ear and Inner Ear. The outer ear directs variations in sound pressure towards the ear-drum. The middle ear transforms these signals into mechanical motion. The inner ear converts the vibrations into electrical firing/ neural firing on to auditory neurons that lead to the brain.

- Outer ear consists of external/ visible part of ear called as Pinnae. The pinnae help in sound localization.
- The Ear-drum marks the beginning of the middle-ear, and accomplishes air impedance transformation between the air medium of outer ear and liquid medium of inner ear.
- Inner ear consists of Cochlea, a tube filled with gelatinous fluid called ‘end lymph’ that transforms mechanical at its oval window membrane into electrical/neural excitation over its neural fibre (auditory nerve) outputs.

2.3 TIME –DEPENDENT PROCESSING OF SPEECH

The underlying assumption in majority of the speech processing schemes is that the properties of the speech signal evolve with time. "Time-domain Processing" [7,8] of speech signals illustrate that the processing methods involve directly the waveform of the speech signal. Frequency domain processing of speech signals usually involves some kind of spectral representation.

The concept of 'Speech analysis' is to extract some properties from a speech signal usually called as "Features or Parameters". The advantage of time-domain analysis of speech lies in the fact that more relevant storage or manipulation of relevant speech parameters is allowed as compared to the original signal. Here, a point should be noted that speech gets sampled at 6000-10,000 samples/sec in order to preserve the bandwidth of 3-5 KHz and to maintain the nyquist rate for the process of sampling. Basically, the transformation of speech signals into a class of parameter signals can, therefore, decrease the sampling rates by two orders of magnitude. Speech analysis aids in necessitating the transformation of one signal s (n) into another signal or a set of signals or usually a set of parameters. The theory of accurate time resolution is rudimentary for segmentation of speech signals as well as ascertaining periods in voiced speech (e.g. locating phone boundaries). Good frequency resolution aids to identify different types of sounds.

2.3.1 SHORT TIME SPEECH ANALYSIS

Since the properties of the speech signal vary relatively slowly with time; we can consider the speech signal as dynamic, random and time-varying. Due to the small variations in vocal-cord vibration and vocal-tract shape; vowels are not even fully periodic. This assumption leads to a number of "short-time" processing methods in which short segments of the speech signal are isolated and processed individually as if they belong to short segments from a sustained sound source with constant properties. Investigation of a short-time window of speech in order to extract parameters is presumed to remain fixed for the entire duration of the window. Thus, in order to model the dynamic parameters; the signal must be divided into successive windows or analysis frames in a way that all the features or parameters can be computed often enough to follow relevant changes. Often these short segments called as analysis frames overlap each other. The outcome of processing on each frame must either be a single number or a set of numbers. Slow changing formants in long vowels may need as large as 100 ms but stops (rapid events) require short windows of 5-15 ms. Therefore, time-processing techniques introduce a new time-dependent sequence which serves as most appropriate representation of speech waveform.

Time-domain characteristics of speech signal usually constitutes of: Average Zero Crossing Rate (AZCR), Energy and auto-correlation function.

2.3.2 WINDOWING:

In digital signal processing, Windowing function [9] can be defined as a mathematical function that is considered to attain a zero value outside of some chosen interval, also called as Apodization or Windowing or Tapering function. The windowing function implies that when any signal or waveform or data function gets multiplied by a window function; the product is zero-valued outside the interval, the remaining part will be the signal that is overlapped by the window, that is weighted by the window's shape. The applications of Windowing involve Filter design, spectral-analysis and beam forming.

The process of multiplication of speech signals s (n) by a window sequence w(n) ;furnishing a set of speech samples x(n) ,usually weighted by the shape of the window. Some windows may have infinite duration but most practical windows have finite length in order to simplify the computations. The application of the windowing process yields the signal segment; represented mathematically in the form as:

$$\mathbf{Q_n} = \sum_{m=-\infty}^{\infty} \boldsymbol{T}[\boldsymbol{x(m)w(n-m)}] \qquad [2.1]$$

The output of the windowing process in order to isolate a particular desired frequency band is subjected to a transformation, T [ ], that can be either linear or non-linear, determined by some adjustable parameter or a set of parameters. The output Qn corresponds to the convolution of T [x(n)] with w(n).Therefore , Qn, being the output of a low-pass filter (the window) is a smoothed version of T[x(n)] and its bandwidth approximates to that of the w(n).The values Qn are therefore a sequence of local weighted average values of the sequence T[x (m)]. For the purpose of efficient manipulation and storage; Q_n should be decimated to a factor generally equal to the ratio of original sampled signal (Speech B.W) to that of the corresponding window.

Here, Q_n will possess a frequency response of that of a low pass filter that is basically a narrowband filter because of its slowly varying waveform. The advantage of smooth hamming window over rectangular window is that the former concentrates more energy at low frequencies where as the later has abrupt discontinuous edges, this energy concentration even at low frequencies will help to retain the integrity of spectral parameters obtained from windowed signals. Since the result of windowing process, i.e. x (n) = s(n).w(n) ; where , s(n) is the input non-stationary signal , corresponds to the convolution of spectra as :

$$X(e^{j\omega}) = \frac{1}{2\pi}\int_{\theta=0}^{2\pi} S(e^{j\theta})W(e^{j(\omega-\theta)} \quad [2.2]$$

In order to minimize distortion in output spectral representation of X ($e^{j\omega}$), the constraints on the window W($e^{j\omega}$) is that it should have a limited frequency range and a smooth shape.
For instance, the property of an ideal low pass filter is that it strictly limits the frequency range of the output spectrum and preserves a constant value. Here a point should be noted that the output spectrum is a smoothed version of S($e^{j\omega}$) where each of the frequency sample should be considered as the average of its neighbours over the entire range that is equal to the bandwidth of the low-pass filter.

The process of windowing applied to time-domain data is referred to as frequency weighing functions, meant to reduce the spectral leakage, associated with finite duration samples. Windows are smoothening functions that taper to zero at the edges and peak usually at the middle frequencies.
Majority of windows possess a finite duration impulse-response because of the certain reasons:

a) In order to strictly limit the analysis in time-range.
b) In order to maintain the phase component.
c) In order to allow and perform a discrete Fourier-transform (DFT) of windowed speech in a way to get the spectral coefficients.

- Frame Rate [9]: The number of times per second; the speech analysis can be performed in order to derive the parametric features .Normally, the frame rate is assumed to be twice the inverse of the window w (n) duration, so that the successive windows overlap by about 50%.
- For any type of window, the duration is always inversely proportional to the spectral bandwidth. That's the reason why traditional wide-band spectrograms [10] apply a window of about 3 ms, with a spectral bandwidth of 300 Hz. On the other hand, narrow-band spectrograms use a window of about 20 ms; having a bandwidth of 45 Hz. Narrow band spectrograms is used mostly for 'F_o' estimation. For viewing of vocal-tract parameters that vary rapidly and in a very short time; fine frequency resolution is not needed and thus wide band representations can be used. In order to perform the windowing of voiced speech, a rectangular window with usually one pitch period duration will generate an output spectrum that will closely resemble a vocal-tract impulse response. When majority of speech analyses have been performed; a fixed window size of longer duration can be used of about 25 ms so that to mitigate the problems of side effects. In case of narrow band windowing ; at least two pitch-periods can be considered for Fo estimation , therefore , pitch analysis will employ a long window of typical size of 30-50 ms.

TYPES OF WINDOWING USED IN SPEECH PROCESSING:

1. Rectangular Window :

A function that is constant inside the interval and attains zero value outside the interval is called as: 'Rectangular Window'. Rectangular window is also named as: Boxcar or Dirichlet Window.

$$W(n) = 1; \quad \text{for } n \leq 0 \leq N-1; \qquad [2.3]$$

Where, N= Length of the window.

W (n) = Sequence of the window.

Advantages : It provides high frequency resolution.

Disadvantages: It incurs sudden changes / discontinuities at the edges can cause spectral leakage during DTFT (Discrete Time Fourier Transform). Spectral Leakage causes the non-zeros values to happen at frequencies other than the interval; thereby causing difficulty in spectrally resolving two signals having same frequencies .Even if frequencies of sinusoids are different; then the leakage from higher components will obscure the weaker signal's presence.

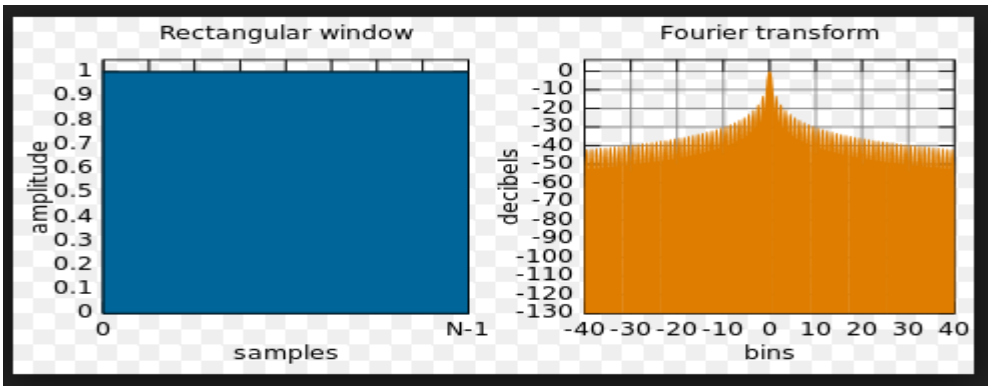

Fig 2.3.2(a): Hamming window design and its Fourier transform representation (Ref.: Signal Processing.com)

2. Hamming window

Hamming Window is also called as: Raised Cosine pulse. Hamming window acts as a low-pass filter h (n), that possess a bandwidth twice the bandwidth of that of a rectangular window because of the reason as more bandwidth; more spectral details can be obtained.

The representation of hamming window can be as:

$$W(n) = \alpha - \beta \cos\left(\frac{2\pi n}{N-1}\right) \qquad [2.4]$$

Here, $\alpha = 0.54$ and $\beta = (1-\alpha) = 1-0.54 = 0.46$

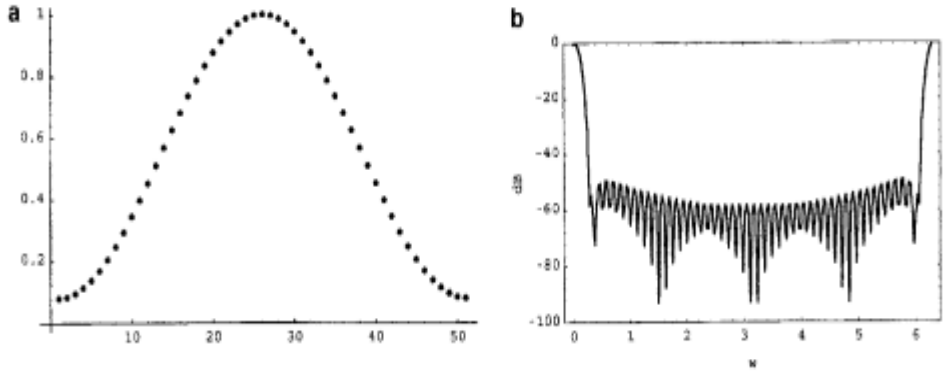

Fig 2.3.2(b): (a) Hamming Window (b) Amplitude spectrum of hamming window (Ref: Signal Processing.com)

2.3.3 FIGURES ILLUSTRATING THE DIFFERENCE IN HAMMING AND RECTANGULAR WINDOWS IN TERMS OF SPECTRAL REPRESENTATIONS

The following figures are demonstrating the magnitude-spectrum of a signal windowed via both hamming and rectangular windows using MATLAB; with window length of M=128 samples,

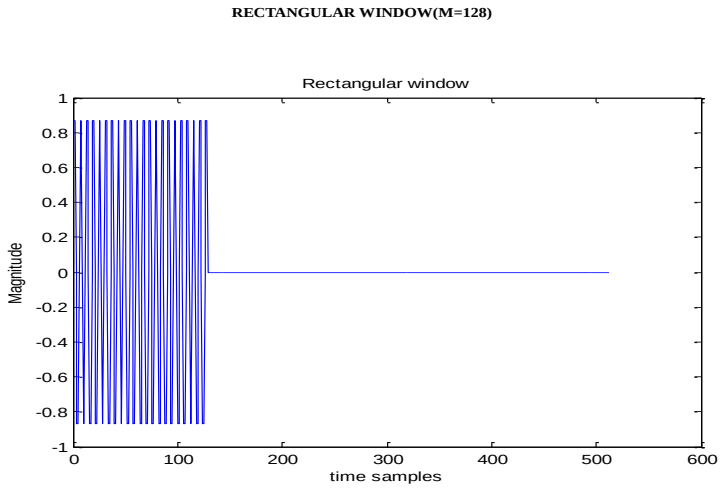

Fig.2.3.3 (a) Representation of a signal using rectangular window of M=128

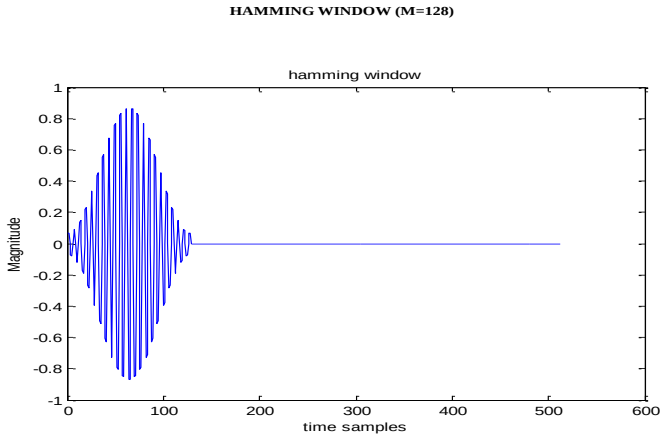

Fig.2.3.3 (b) Representation of a signal using Hamming window of M=128

From the figures of 2.3.3(a) and 2.3.3(b); it is evident that there are more sharp edges, discontinuities and spectral leakage in case of rectangular window as compared to hamming window.

REPRESENTATION OF MAGNITUDE SPECTRUMS USING HAMMING AND RECTANGULAR WINDOWS OF WINDOW LENGTH OF M=128

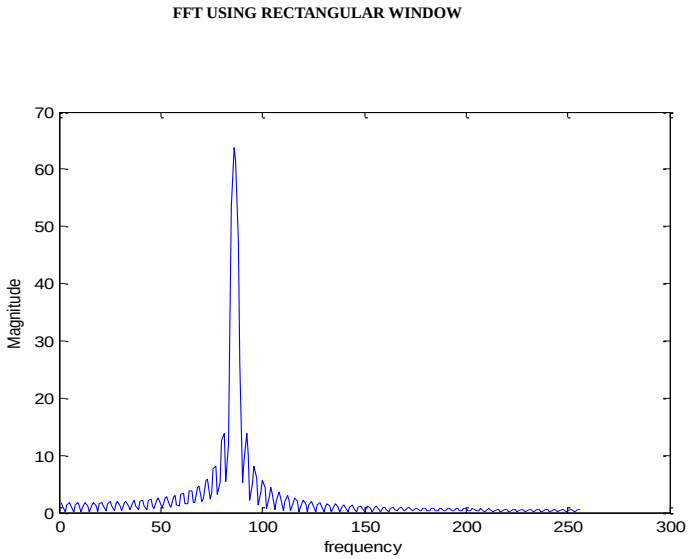

Fig.2.3.3(c) Representation of magnitude-spectrum using rectangular window of M=128

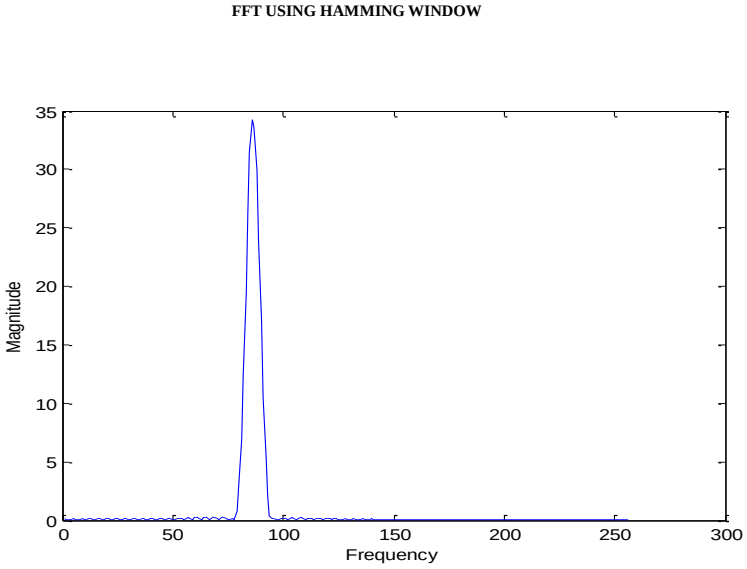

Fig.2.3.3 (d) Representation of magnitude-spectrum using hamming window of M=128

From the figures of 2.3.3(c) and 2.3.3(d) ; it is evident that there are more side-lobe levels and therefore more side –lobe level energy in case of rectangular window as compared to hamming window and the hamming window has its entire energy concentrated at the main side lobe level(more tapering at the ends).

2 Hanning Window:

The representation of hanning window is as under:

$$W(n) = \alpha - \beta \cos\left(\frac{2\pi n}{N-1}\right) \qquad [2.5]$$

When α=β=0.5; then the hamming window becomes the Hanning window.

$$W(n) = 0.5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right) \qquad [2.6]$$

For the Zero Phase version;.

$$W(n) = 0.5\left(1 - \cos\left(\frac{2\pi n}{N+1}\right)\right) \qquad [2.7]$$

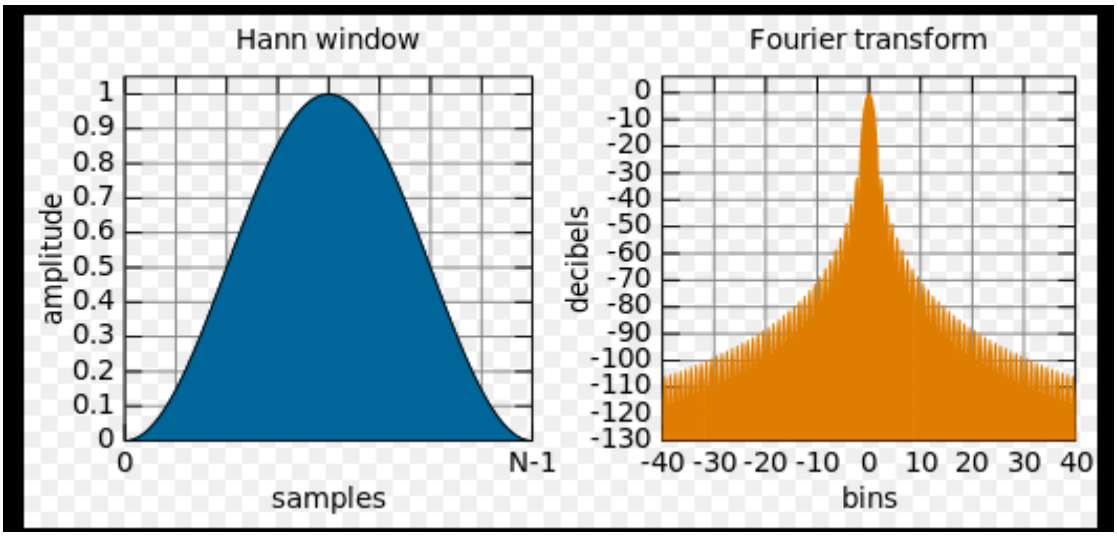

Fig 2.3.3(e): Hanning window design and its Fourier transform representation (Ref.: Signal Processing.com)

3 Blackmann windows :

The representation of the blackmann window is as under:

$$W(n) = a_0 - a_1 \cos\left(\frac{2\pi n}{N-1}\right) + a_2 \cos\left(\frac{4\pi n}{N-1}\right) \qquad [2.8]$$

Where, $a_o = \frac{(1-\alpha)}{2}$; here α = 0.16 and $a_1 = 0.5$ and $a_2 = \frac{\alpha}{2} = 0.08$

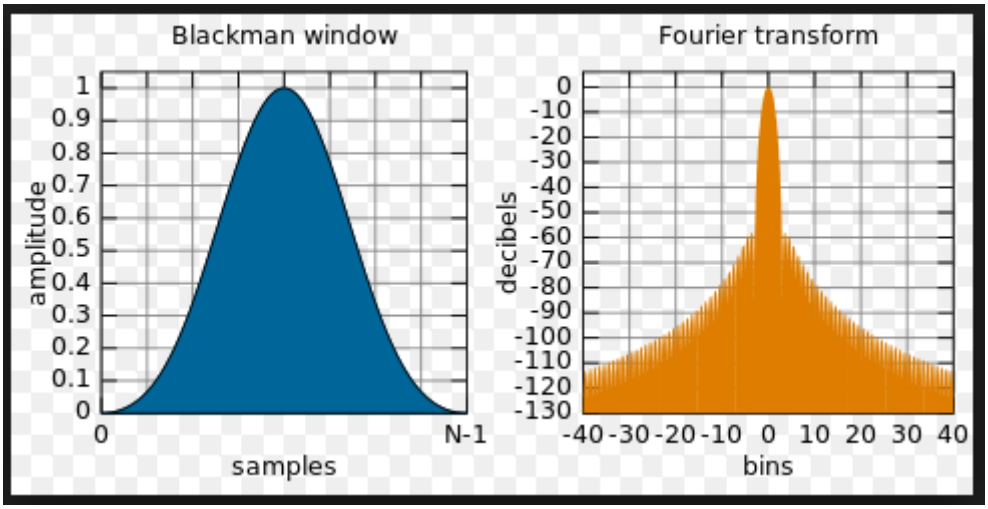

Fig 2.3.3(f): Hanning window design and its Fourier transform representation (Ref.: Signal Processing.com)

ADVANTAGES OF BLACKMANN OVER HAMMING AND HANNING WINDOWS:

i. The hamming and hanning window possess fewer terms as compared to blackmann windows in any window sequence function; as more number of terms will specify more accuracy in results.

ii. In Blackmann window function ; there is a presence of an extra cosine term ; that will reduce the side – lobe level ; which indicates more accuracy as less power gets dissipated through side-lobe levels.

iii. The value of equivalent noise-bandwidth is greater in case of blackmann windows as compared to hamming and hanning windows.

iv. The response of blackmann window in both time as well as in frequency domain is far better as compared to hamming as well as hanning windows.

2.4 FREQUENCY DOMAIN OR SPECTRAL REPRESENTATION OF SPEECH SIGNALS

Speech signals need to be more consistently and easily analysed spectrally (in the frequency-domain) as compared to that in the time-domain. With relative to phase and timing aspects; . The process of human hearing appears to pay much more attention to spectral aspects of speech. Therefore, in the process of extraction of parameters or features from speech; Frequency domain analysis [11] of speech or Speech spectral analysis is very mandatory.

- **FILTER-BANK ANALYSIS [12] :**

  One spectral analysis method involves usage of a filter-bank or a set of band-pass filters put in series (analog or digital); each filter analysing a different range or band of frequencies of the input speech.

  The usage of filter –banks in digital speech processing are more flexible as compared to DFT [13], since the band-widths need to be changed or manipulated in order to follow the frequency resolving power of the ear; rather than being fixed in DFTs. In addition to this; the amplitude that outputs from a bank of 8-12 band-pass filters will produce a more efficient spectral representation than a more detailed DFT.
  These filters often follow a Bark-scale, equally spaced, and basically fixed B.W filters up to 1 KHz.

2.4.1 SHORT-TIME FOURIER TRANSFORMS [14]

In order to obtain a speech representation in terms of amplitude and phase; as a function of frequency; the process of 'Fourier Transform' is being applied.

Considering the vocal-tract as a linear filter (Linear System); the Fourier-transform of speech is calculated as the product of the transforms of glottal excitation and vocal-tract impulse response. Speech is not stationary; and hence; the short-time analysis using windows is necessary.

- The short time Fourier transform of a speech signal s(n) can be given by :

$$S_n(e\hat{}j\omega) = \sum_{m=-\infty}^{+\infty} S(m)w(n-m)e\hat{}j\omega m \qquad [2.9]$$

Here, s (m) * w (n-m) is the "windowed" signal.

If we will consider 'ω' as a constant ; then the spectral transformation has an interpretation of Qn as in equation [2.1] ; where the transformation 'T' indicates the multiplication by a complex exponential of frequency 'ω'. Considering w (n) as a low-pass filter; {Sn (eˆjω)} is taken as a time signal that replicates its amplitude and phase of s(n) ; and is having the Band-width same as that of window ; being centred at ω radians. In order to simplify the computations; the Discrete Fourier transform is applied instead of standard Fourier transform so that the frequency variable takes on discrete values of N where, N= window duration /size of DFT.

$$S_n(k) = \sum_{m=0}^{m=N-1}[s(m)w(n-m)].e\hat{} - j2\pi km/N \qquad [2.10]$$

Where, N= DFT size / window size.

- **CHOICE OF 'N'**

The choice of N stands very crucial for Fourier analysis. Lower values of N will have worse frequency resolution but provide good time resolution since all speech parameters are averaged over short intervals. Large value of N will indicate poor time resolution and good frequency resolution. Short length of windows will serve better for formant analysis as well as segmentation where as long windows can serve for better formant estimation and detection in order to have better harmonic analysis. Since, speech s(n) is real ; S(eˆjω) is conjugate symmetric and hence the function needs to be preserved from ω=0 to ω=**π**.

2.4.2 SHORT-TIME AUTOCORRELATION FUNCTION:

The Spectral representation of speech in the form of Fourier –transform ascertains both spectral magnitude and phase. The auto-correlation function [15] of s(n) is obtained by performing the Inverse Fourier-transform of Energy / power Spectrum |S(eˆjω|ˆ².The purpose of auto-correlation function is to provide information about the formant amplitudes as well as in harmonicity in s(n) as well to know about the periodicity of the signal.

The special form of cross-correlation is Autocorrelation function; given by the expression:

$$\emptyset_{sy}(k) = \sum_{m=-\infty}^{m=\infty} s(m)y(n-k) \qquad [2.11]$$

Auto-correlation function basically measures the amount of similarity between the two signals s (n) and y (n) as a function of time-lag (k) between them.

- PROPERTIES AND ADVANTAGES OF SHORT-TIME AUTOCORRELATION FUNCTION :

1. Autocorrelation function is an even function; i.e. r (k) =r(-k).
2. If s (n) is periodic in 'P' samples; then autocorrelation function will also be periodic.
3. The maxima for the r(k) will occur predominantly at k=0 ,$\pm P$,$\pm 2P$.....
4. At k= 0; it will possess maximum energy and r(0) equals the energy in s(n) Average Power for random as well as periodic signals.

Short-time autocorrelation function is obtained by windowing s(n) and then applying equation [11] ; we get :

$$R_n(k) = \sum_{m=-\infty}^{m=\infty}[s(m).w(n-m)]\{y(m-k)w(n-m+k)\} \qquad [2.12]$$

The multiplicative product of speech signal s(n) with its delayed version s(n-k) is passed through a low pass filter having the impulse-response as :w(n) w(n+k). The windows with short length will reduce the complexity by reducing the calculations. That is, if w (n) has N samples; then (N-k) products are required for each value of R_n (k).

2.5 HANDLING AUDIO IN MATLAB:

I. Recording sound in MATLAB:

Recording sound directly in MATLAB requires the user to specify the number of samples to record, the sample-rate, the number of channels, and sample format. For example, to record a vector of double precision floating point samples on a computer with attached or integrated micro-phone, the following MATLAB command may be issued:

Speech = wavrecord (16000, 8000, 1, 'double');

This records 16,000 samples with a sample-rate of 8 KHz, and places them into a 16,000 element vector named *Speech*. The '1' argument specifies that the recording is mono rather than stereo.

If we need to use the audio recorder function, the procedure is to create an audio recorder object, specifying sample rate, sample precision in bits, and number of channels, then to begin recording:

Aro = audiorecorder (16000, 16, 1);

Record (Aro);

At this point, micro-phone is actively recording. When finished, stop the recording and try to play back the audio:

Stop (Aro);

Play (Aro);

To convert the stored recording into the more usual vector of audio, it is necessary to use the *getaudiodata ()* command:

*Speech = getaudiodata (*Aro, 'double'*);*

- ✓ **Storing and Replaying sound:**

  Replaying a vector of sound stored in floating point format is also easy:

  Sound (*Speech*, 8000);

  8000 Hz (8 KHz) is the sampling rate applied to the speech.

  The command that is used for scaling of speech in both directions, so that a vector is too quiet to be amplified, and one that is too large will be attenuated given as:

  Sound (*Speech*/max (abs (*Speech*)), 8000);

  A time-domain plot of a sound sample is given in MATLAB as:

  Plot (*Speech*);

  Although it is sometimes preferred for the x-axis to display time in seconds:

  Plot) [1:size (*Speech*)] / 8000, speech) ;

- ✓ **Normalization of Speech:**

  Normalization is important to perform in speech in order to prevent clipping:

  Sound (*Speech*/max (abs (*Speech*)), 8000);

Frequency domain operations, by contrast, require the audio to be first converted to the frequency domain, by use of a Fourier-transform, or Fast Fourier Transform (FFT):

```
a_spec = fft (a_vector);
```

in general, when the audio length is not a power of two, it is possible to zero-pad (truncate) the audio vector, to fill the size of FFT specified, as:

```
a_spec = fft (a_vector, 256);
```

The resultant frequency domain vector can be complex. In order to get a double-sided frequency representation using:

```
Plot (abs (a_spec));
```

✓ Continuous Filtering:

Create the filter, apply to the entire array of speech, then listen to the result, in vector y:

```
H= [1, -0.9375];
Y = filter (h, 1, s);
soundsc (y);
```

For a succession of 240 sample frames;

```
W = 240;
N = floor (length (s)/W);
For k = 1: n
Seg = s (1_ (k-1)*W: k*W);
Segf = filter (h, 1, Seg);
Outsp (1+ (k-1) *W:k*W ) = Segf;
End
Soundsc (outsp);
```

✓ Correlogram

A Correlogram is a plot of representation of the auto-correlation of a signal. Correlation is the process by which two signals are compared for similarities that may exist between them either at the present time or at the past.

```
[c, lags ] =    xcorr (x,y);
```

Chapter 3: RELATED WORKS ABOUT VOCAL – MUSIC SEPARATION

Identification of vocals and non-vocal segments is the foremost step in Music/Speech separation methods, thereby, followed by a variety of techniques to separate the lead vocals from the repetitive background including accompaniment model learning, spectrogram factorization and pitch-based inference techniques.

Certain audio features such as MFCCs, LFPC (Log Frequency Power Coefficients) [14], PLPC (Perceptual Linear Predictive Coefficients) [15] are applied to designate vocal and non-vocal regions. Apriori known non-vocal segments are used to train an accompaniment model; which is based on a Probability Latent Component Analysis (PLCA). Vocal and Non-Vocal Segmentation is performed using MFCCs and Gaussian Mixture models (GMM) [16]. Bayesian models are trained to adapt a background model learned from the non-vocal sections. Predominant pitch and melody estimators are accomplished in vocal segments to extract the pitch-contour, which was finally put to use to separate the speech components via Binary TF masking.

A song mixture is modeled as the sum of signal of interest (lead) and a residual (background) ,where the background is frame worked as an unconstrained NMF structure and lead vocals are parameters as a filter or source model. Broadly speaking, mono-aural T-F masking systems for song segregation into vocals and non-vocals can be distinguished into feature-based and model-based. Model-based systems imbibe trained and noise models to separate noisy speech. The mixture speech sources can be modeled using a Hidden Markov Model (HMM). During the separation algorithm, the song mixture is decomposed into the underlying HMMs concentrating a factorization method equivalent to binary T-F masking.

3.1 BEAT SPECTRUM – A NEW APPROACH TO RHYTHM ANALYSIS

The purpose of calculating Beat-spectrum [17] is to know about the repetition or symmetry in music. Beat Spectrum is generally calculated for characterizing the rhythm and tempo of audio and music.

- **RHYTHM:** Rhythm in music refers to the symmetry or regular occurring motions in music. It denotes the periodicity / frequency of music from microseconds to millions of years.

- **BEAT:** Beat refers to the regularly occurring pattern of rhythmic stresses in music. When we count tap / clap; along with music; we are experiencing the beat.

- **TEMPO:** Tempo is referred to as speed of the beat; usually measured in Beat per Minute (BPM).

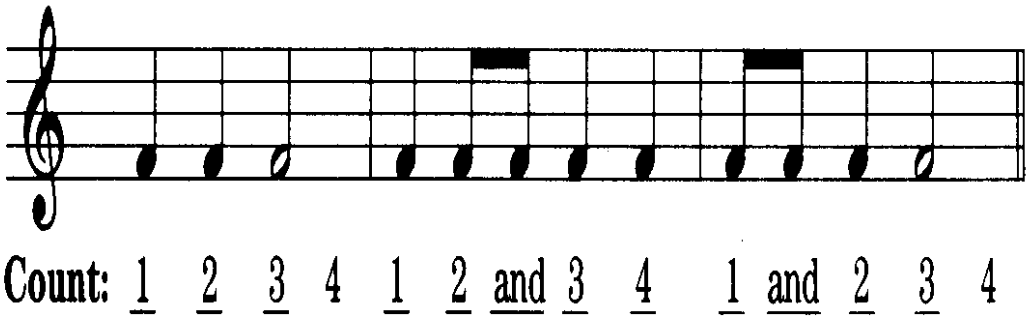

Fig 3.1: Reading a song melody: pitch and rhythm (Ref: Essential-Music-Theory.com)

- **BEAT SPECTRUM :**

The beat spectrum ‘b’ is defined as a measure of acoustic self-similarity, as a function of time-lag. Highly structured or repetitive music will possess strong beat-spectrum at the repetition peaks at the repetition times; describing tempo and relative strength of periodic peaks. These peaks in beat-spectrum will respond to the major rhythmic components of the source-audio, therefore, the audio components will periodically repeat at different levels.

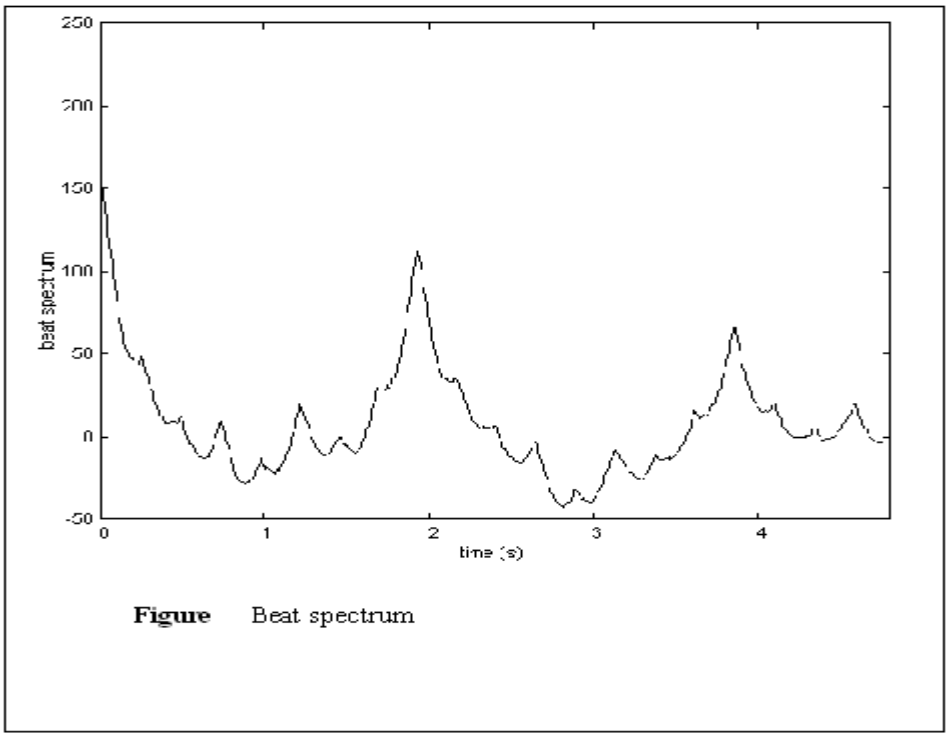

Figure Beat spectrum

Fig 3.1(a): Beat Spectrum of a 'Rock' genre music. (Ref: Beat Spectrum-New approach to Rhythm analysis)

The repetition time of each component can be determined by the "Lag-time" and the relative amplitudes of peaks corresponding to the strengths of the rhythmic components. In the paper; the Beat Spectrogram is also presented; that graphically illustrates rhythmic variation over time. The beat spectrogram is an image formed from the beat spectrum over successive windows. Strong rhythmic components are visible as bright bars in the beat spectrogram, showing changes in tempo or time signature. In addition, a measure of audio novelty can be computed that measures how novel the source audio is at any time .Therefore, periodic peaks correspond to the rhythmic periodicity in the musical structure.

- **BEAT SPECTROGRAM**

Beat Spectrogram [16-18] visualizes spectral evolution over successive windows. It evaluates the beat – spectrum over successive windows to show rhythmic variations over time. Beat spectrogram has units of time on X-axis and lag-time on the Y-axis. Therefore, Beat-spectrogram is an image formed by successive beat spectra.

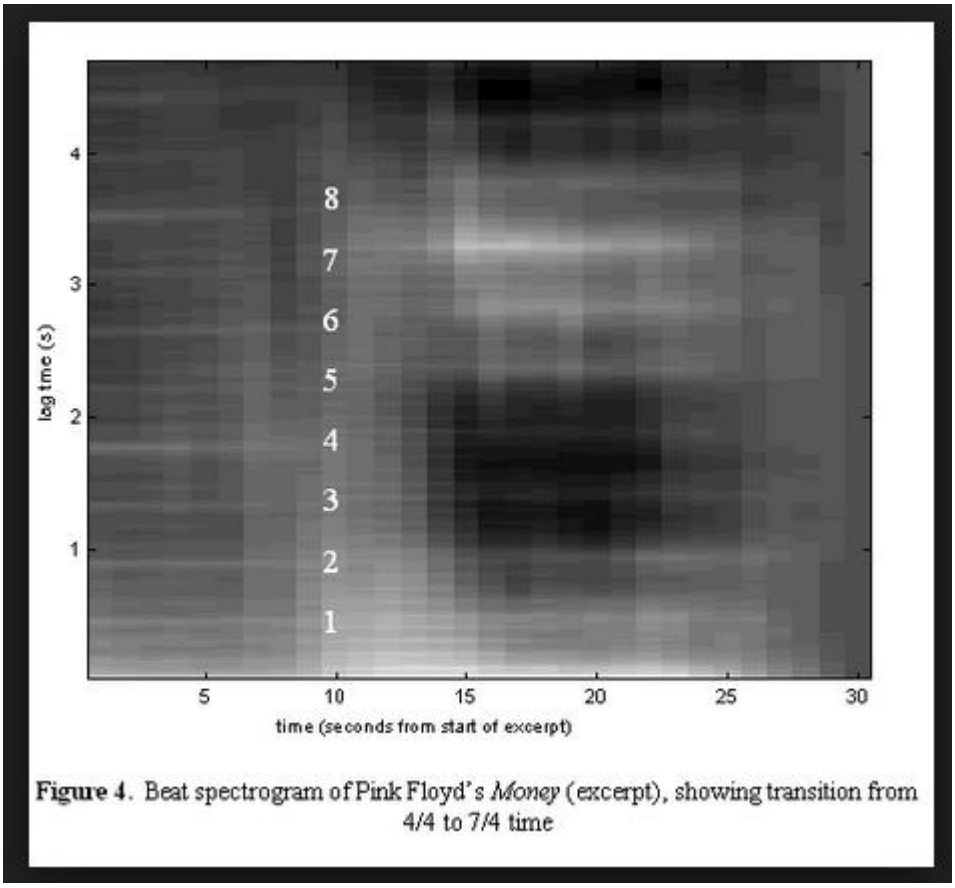

Figure 4. Beat spectrogram of Pink Floyd's *Money* (excerpt), showing transition from 4/4 to 7/4 time

Fig 3.1(b): Beat Spectrogram of a Pink Floyd's *Money* (excerpt) (Ref: Beat Spectrum-New approach to Rhythm analysis)

- **ALGORITHM STEPS**

The calculation of beat-spectrum from audio excerpt is calculated from three principal steps:

i. First, the audio is parameterized using spectral or some other kind of representation called as Spectrogram Formation. ; resulting in a sequence of feature vectors.
ii. "Distance Measure" is calculated to calculate the similarity between all pair-wise combinations of feature vectors.
iii. Also, the similarity between two instances of music is embedded in a 2-D representation called as a 'Similarity Matrix'.
iv. The Beat Spectrum results from finding periodicities in similarity matrix, using the paradigm of autocorrelation or Diagonal sum.

Ist STEP: PARAMETERIZATION OF AUDIO:

'Parameterisation of audio' [19] is performed by the windowing of the audio waveform. Then Fast Fourier Transform (FFT) is calculated on each frame. Log (FFT) or Log (Magnitude-spectrum) is estimated as the Power Spectrum of the signal

Parameterization of audio also refers to calculating the spectrogram of the signal.

SPECTROGRAM:

Spectrogram is represented as a graph of energy content of a signal, as a function of time. It is basically the visual representation of the spectrum of frequencies in any audio, expressed as a function of time or some other variable, also named as: Spectral water-falls, Voice Grams or Voice-prints. The common representation of spectrogram is a graph or image with two geometric dimensions: time is represented by the horizontal axis and frequency is represented by the vertical axis and the third axis will represent the amplitude of a particular frequency at a particular time and is indicated by the intensity of the colour; where black and dark colours will show maximum energy while white represents the least energy.

CREATION OF SPECTROGRAMS:

Spectrograms are usually approximated as a filter-bank; through a series of band-pass filters .They are also calculated from the time-signal using FFT. Analog processing is applied by band-pass filters in order to segment the input signal into different frequency – bands.

FFT is usually performed on digitally sampled signal in time-domain and further divided into chunks; the STFT of each chunk is calculated in order to get acquainted with the frequency-response. The spectrogram of a signal s(t) can be estimated by computing the squared magnitude of STFT of the signal as :

$$\text{Spectrogram}(t, f) = |\text{STFT}(t, f)|^2 \qquad [3.1]$$

The instrument used to capture the spectrogram for the purpose of measurement is the Spectrograph.

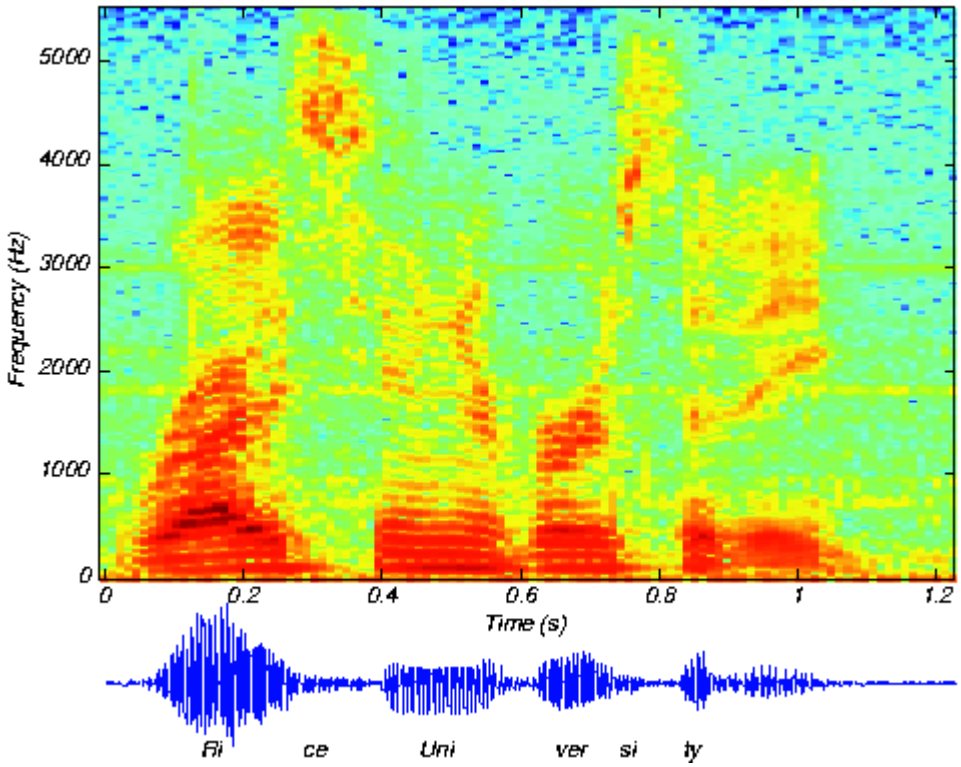

Fig 3.1(c): Spectrogram of a human speech (Ref: CNX.org)

2nd STEP: CALCULATION OF FRAME SIMILARITY MATRIX

The feature vectors are generated as a result of parameterization of audio. Once the audio has been parameterized, it is then embedded in a 2-dimensional representation. A (dis)similarity measure (D) is calculated between the vectors V_i and V_j of the frames 'i' and 'j'.

The distance measure or Euclidean distance is evaluated using the dot product in order to give the cosine of angle between the parameter vectors.

$$D_c\,(i\,,j) = \frac{Vi.Vj}{|Vi|.|Vj|} \qquad [3.2]$$

The cosine measure ensures that windows with low energy, such as those containing silence, can still yield a large similarity score, which is generally desirable.

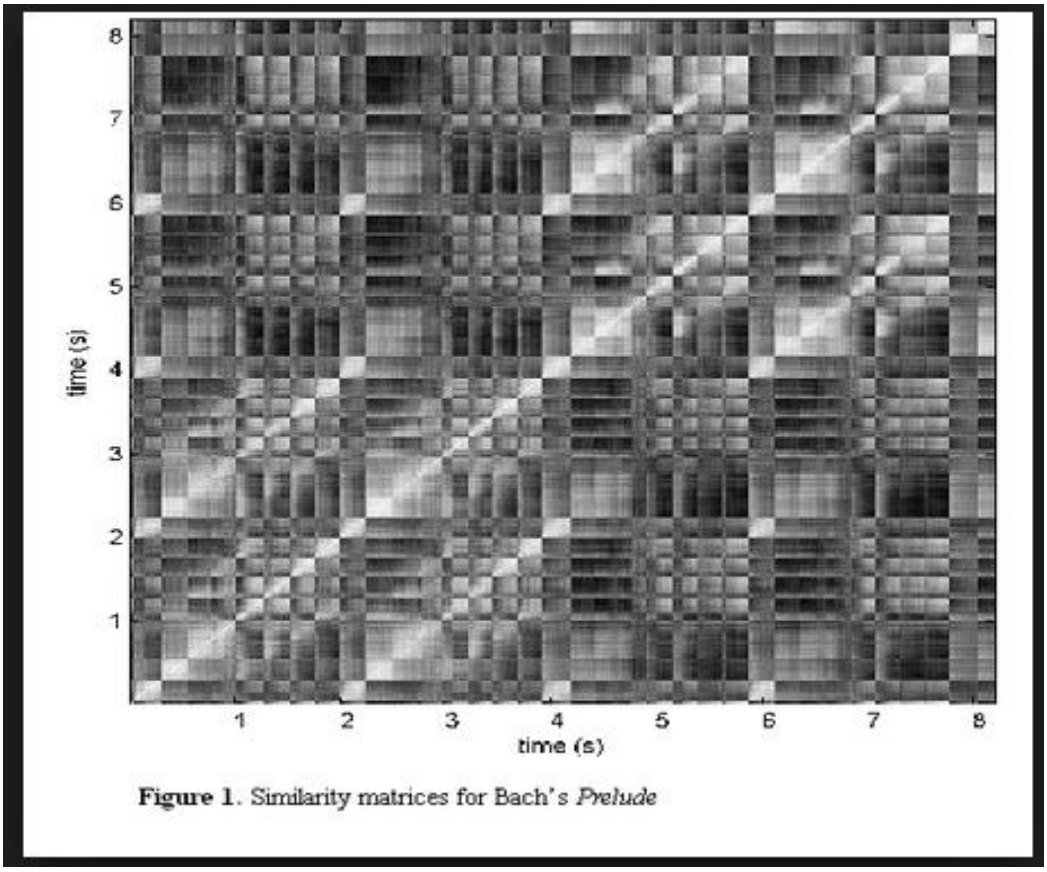

Figure 1. Similarity matrices for Bach's *Prelude*

Fig 3.1(d): Similarity Matrix (Ref: Beat Spectrum- A new approach to rhythm analysis).

3RD STEP: DISTANCE MATRIX EMBEDDING

It is convenient to consider the similarity between all possible instants in a signal. This is done by embedding the distance measure in a two-dimensional representation. The similarity matrix *S* contains the distance metric calculated for all frame combinations (hence time indexes *i* and *j)* such that the *i, j* th element of *S* is *D (i, j).* S can be visualized as a square image where each pixel *i, j* is given a gray scale value proportional to the similarity measure *D (i,j)* , and scaled such that the maximum value is given the maximum brightness. These visualizations let us clearly see the structure of an audio file. Regions of high audio similarity, such as silence or long sustained notes, appear as bright squares on the diagonal. Repeated figures will be visible as bright off-diagonal rectangles. If the music has a high degree of repetition, this will be visible as diagonal stripes or checkerboards, offset from the main diagonal by the repetition time.

4TH STEP: DERIVATION OF THE BEAT – SPECTRUM

Both the properties of periodicity and relative strength of rhythmic structure can be derived from the similarity matrix. Thus, the measure of self-similarity as a function of the lag is regarded as the beat spectrum B (l). Peaks in the beat spectrum correspond to repetitions in

the audio. A simple estimate of the beat spectrum can be found by summing *S* along the diagonal as follows:

$$B(l) = \sum S(k, k+l) \qquad [3.3]$$

Here, B (0) represents the sum along the main diagonal; and

B (1) represents the sum along the first super diagonal.

The Beat spectrum can also be computed from the auto-correlation of S; given by:

$$B(k, l) = \sum S(i,j)\, S(i+k, j+l) \qquad [3.4]$$

3.2 MUSIC / VOICE SEPARATION BASED ON SIMILARITY MATRIX

This proposed algorithm focuses on the condition when the repetitions in music occur intermittently; i.e. without a fixed or constant period, therefore, allowing the processing of music pieces having fast – varying repeating structures and isolated repeating elements. The proposed algorithm works on similarity matrix for the purpose of identification of repeating elements instead of the formation of beat-spectrogram.

In the next step; a 'repeating spectrogram model' is formulated with the help of calculation of median / segment model and thereby, extracts the repeating patterns via TF masking.

- **PURPOSE OF MEDIAN FILTERING**

  In order to form the repeating segment "S" from the back-ground, median filtering is performed on the repeating segments of music. In order to extract the repeating patterns from audio; a Time-Frequency Mask (TF Mask) is finally derived and thus, allowing the processing of musical pieces with fast-varying repetitive structures without the need of formation of beat- spectrum for identification of periods of repeating structure before-hand.
  Main function of Median Filtering is as below:

  a. In order to remove or filter noise that is basically considered as music in the audio excerpts. It is a sort of non-linear filter that is often applied in the system to filter the noisy random components.

ALGORITHM:

The usage of median filter lies in the fact that it runs or passes through the signal entry -by- entry or window-by-window ; replacing each frame with the 'median' of neighbouring frames or entries. The pattern of neighbours also called as Windows or frames slides entry by entry over the entire signal.

Now, consider that if any window has an odd number of entries; then the Median is taken as the 'middle' value of all entries in the numerically sorted windows / frames. For even number of windows, there is more than one possible median.

APLICATION OF MEDIAN FILTERING: PERCUSSSIVE BEAT-TRACKING USING MEDIAN FILTERING

For the transcription of 'drum-events' for learning the drum pattern of any song or audio; percussive signals are used. Median filtering is applied on the mixture spectrogram in order to separate the enhanced harmonic and percussive components or spectrograms of a signal. Percussive features consist of wide-band noise over all frequencies and appear as vertical lines in the spectrogram; whilst the harmonic components tend to consist of frequencies that persist and appear as horizontal lines. These are then compared to prepare a percussive mask, indicating the extent of each energy component belonging to the percussive component. By summing all the percussive elements along a given time frame over all frequencies will generate the 'Percussive detection function'. These median filters pass through the spectrogram and provide a substitution for the given value of a signal with the value of 'N' neighbouring windows. A two-dimensional matrix that is used to calculate the similarity or dissimilarity between two parameters or feature vectors 'a' and 'b' of a given sequence.

1st STEP: WINDOWING AND CALCULATION OF SIMILARITY MATRIX 'S'

i. A single (mono-aural) channel matrix 'x' is given. Calculate the STFT 'X' using half-overlapping Hamming windows of N samples length; leading to the generation of feature vectors or parameters.
ii. Calculation of Magnitude-spectrum 'V' by taking the absolute value of V=|X|; by discarding the symmetric part and considering only the DC component.
iii. Then the Normalization of columns of 'V' using the Euclidean norm $[\frac{V}{\sqrt{V}.\sqrt{V}}]$
iv. Similarity Matrix 'S' is defined as the matrix multiplication between transposed V and normalized V.
Each point or parameter (a, b) measures the cosine similarity between time frames j_a and j_b of the mixture spectrogram.

CALCULATION OF SIMILARITY MATRIX:

Similarity matrix is calculated by:

$$S(j_a , j_b) = \frac{V(i,ja).V(i,jb)}{|Vi,ja||Vi,jb|} \qquad [3.5]$$

$$= \frac{\sum_{i=1}^{n} V(i,ja).V(i,jb)}{\sqrt{\sum_{i=1}^{n} V(i,ja)^{\wedge}2}.\sqrt{\sum_{i=1}^{n} V(i,jb)^{\wedge}2}} \qquad [3.6]$$

Where, n = (N/2+1) ...frequency channels and,

$V\ j_a,j_b \in [1,m]$, where m= time frames

2ND STEP: CALCULATION OF REPEATING ELEMENTS

Similarity matrix 'S' is used to identify the repeating patterns in the mixture spectrogram 'V'. For all the total frames 'j' in V; we look for the frames that are most similar to frame 'j'and keep them in vector of indices 'J_j'. Using the similarity matrix therefore reveals the underlying repeating structures that don't necessarily happen in a periodic fashion.

3RD STEP: CALCULATION OF REPEATING MODEL

The repeating spectrogram model 'W' is derived for back-ground once all the repeating elements have been identified for all frames 'j' in mixture spectrogram 'V'. By taking the median of corresponding repeating frames, indicated by J_j ; repeating spectrogram model can be estimated.

$$W(i,j) = \text{median}\ [V_i , J_j(l)\ ] \qquad [3.7]$$

Where, i= [1,n] = number of frequency channels.

J=[1,m] =time frame index.

$Jj = [J_1\ J_2\ .....J_k]$ = indices of repeating frames.

R= maximum of repeating frames.

The TF bins that represent fewer deviations comprise of a repetitive pattern and can be captured by median. TF bins with larger deviations between repeating patterns can be removed by median.

4TH STEP: CALCULATION OF TIME-FREQUENCY MASK

The creation of refined repeating spectrogram model 'W'; by taking the minimum between V and W, for every TF bin. By normalization of W' by V; a time-frequency mask is created. The reason of normalization is that for TF bin; that possess repeating patterns in V will attain values near '1' in 'M' will get weighted towards the repeating background and the Tf bins that acquire values near '0' in M will be modelled as estimates of foreground vocals.

$$W'(i,j) = \min (W(i,j) , V(i,j)) \qquad [3.8]$$

And,

$$M(i,j) = \frac{W'(i,j)}{V(i,j)} \qquad [3.9]$$

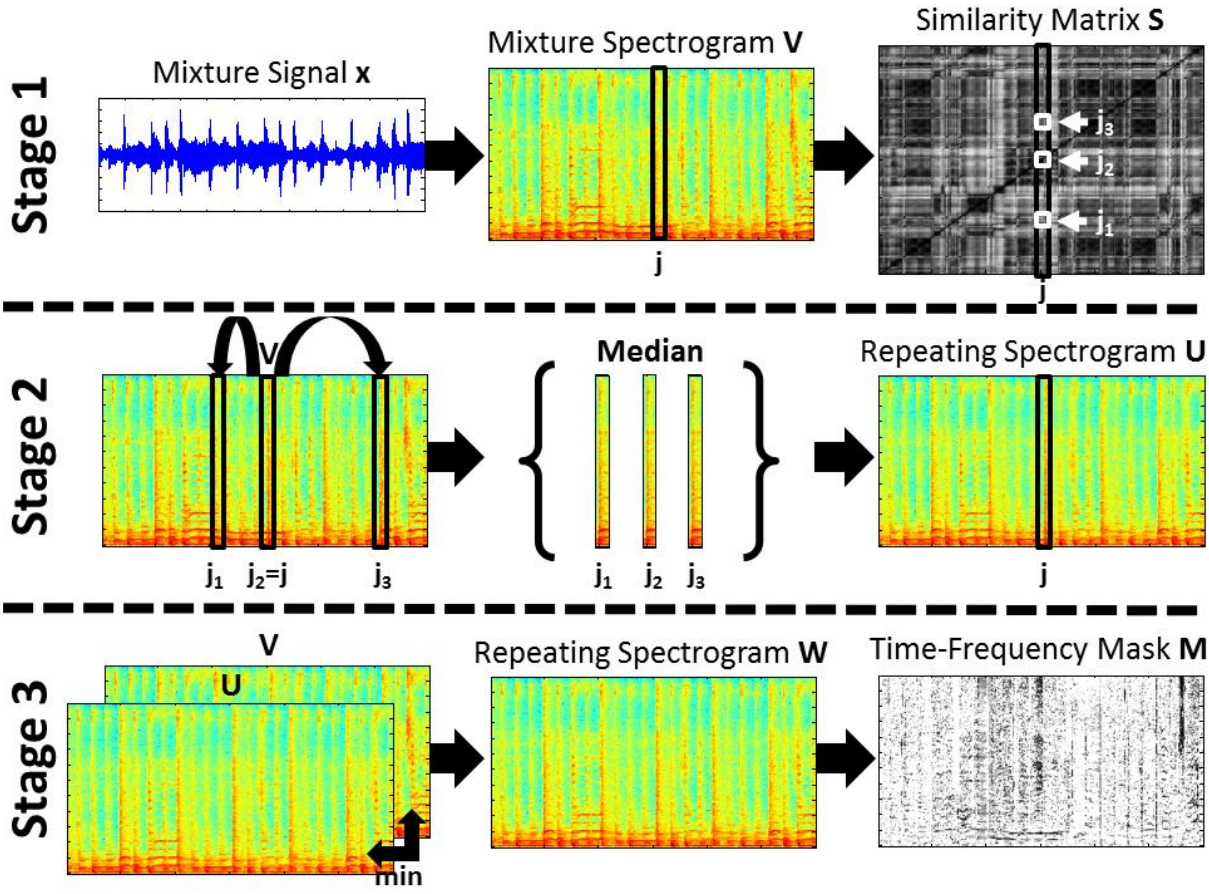

Fig 3.1(e): Similarity Matrix and the formation of Time-frequency Mask (Ref: Music/Voice separation based on similarity matrix)

Chapter 4: PROPOSED WORK

• EXTENSION OF REPET ALGORITHM

The Thesis work pivots on separation of musical background and singing voice signal in clipped audio excerpts on the basis of REPET [20] (REpeating Pattern Extraction Technique).REPET algorithm emphasizes on separating out human vocals from structured accompanying background noise (e.g. .music, hammering, engine noise). The rudimentary principle in REPET algorithm in the field of Music Information Retrieval (MIR) [21] is 'REPETITION' [21], as premise of music, as an art. The concept of repetition is, hence forth, quintessential for the analysis of musical structure, rhythm and pitch /melody estimation as well as summarization. The highly exacting task of efficiently separating a song into its musical and voice components has many important applications such as instrument /vocalist identification, melody transcription from musical song mixtures, automatic Karaoke gaming, removal of repetitive background noise, artifacts and interference for improved speech recognisation in hearing aid design, and more essentially defines the ability for any user to directly interact with musical context of audio tracks.

Recent trends in audio source extraction are relied on a filtering paradigm in which musical and speech sources get recovered through the direct processing of the song mixtures. For speech enhancement, the classical Wiener filter can be estimated as a TF (Time-Frequency) mask, where each TF element (unit) of mask represents the ratio of target energy to the mixture energy within the unit. The TF masking can be thus approximated as an element – wise weighing of the TF representations (Short-Time Fourier Transform) of the mixtures. Such representations deduced either by a Short-Time Fourier Transform (STFT) or a windowed auditory filter-bank in the form of a cochleagram. Usually TF masks consider the values of 0 (background) and 1(foreground) evolving in binary TF masking for audio source separation. Therefore, when individual TF bins are allocated weights of either 0 or 1, enunciated as Binary TF masking [23].

Considering the case for binary TF masking, energy from each specific TF bin is designated to be from just one source (foreground or background). Soft Weighing Strategy implies for assigning values between 0 and 1 in order to allocate energy proportionally to each source. The requisite of applying TF masks to the song mixture is to separate mono-aural sound sources. Furthermore, the algorithms meant for computing TF masks are basically of the kinds of Computational Auditory Scene Analysis (CASA) or Independent Component Analysis (ICA) [25]. CASA techniques aims at segregating sound sources on the basis of perceptual postulates of auditory scene analysis .On the other hand, ICA presumes that the source signals are statistically independent, thereby, constructs the separation problem as that of approximating a demixing matrix .

Several music / voice separation methods prioritizes on modeling either the music signal , by usually training an accompaniment model from non-vocal units or vocal segments, by plotting the pitch or melody contour. Audio features can also be predominantly derived using MFCCs (Mel-frequency cepstrum coefficients) [23]. REPET algorithm is an approach to explicitly analyze the repeating musical accompaniment for segregation of lead melody from

background accompaniment by finding the repeating patterns in the audio mixture and then applying spectral subtraction technique by extracting them from non-repeating elements. Thus, this algorithm can be designated as a procedure in which an audio mixture can be understood as a repeating musical accompaniment ,on which varying voice signal(Verse) is superimposed that is not repetitive. By correctly estimating the period of the immediate repetitive structure, background segment model can be computed and developed. This technique is considered to be extremely effective for audio clips with relatively stable repeating background.

In the proposed work, the performance evaluation of the REPET algorithm can be adjudged by computing the SIR (Signal to Interference Ratio) [26-29] values on the basis of foreground and background energies as well as variance values, rather than applying bss_eval tool-box.

- **ALGORITHM STEPS**

The entire algorithm is similar to REPET algorithm, as the following algorithm is organized as follows:

1. Identification of the repeating period through calculation of auto-correlation matrix.
2. Modelling of the repeating segment.
3. Extraction of the repeating patterns and subtraction of the background patterns from the audio mixture.

A. Repeating Period Identification

Since the principle of 'Repetition' of music has been exploited, the periodicities of the accompanying background can be determined by calculation of the auto-correlation matrix. The reason for calculation of auto-correlation is that if a signal is periodic, then its auto-correlation matrix is also periodic. Auto-correlation is considered as a mathematical tool for finding the repeating patterns, for example, the presence of a periodic signal corrupted by noise.

The phenomenon of auto-correlation 'R_{xx}' basically measures the similarity of a signal as reference and a lagged version of itself over successive time-intervals; given by equation (1) as:

$$Rxx(m) = \sum_n x(n).x(n+m) \qquad [4.1]$$

i. The audio mixture signal 'x' is considered; since the speech signal is varying in amplitude and time; and thus, it's non-stationary. The stationarity property is demonstrated by calculating the STFT (Short-Time Fourier Transform) 'X' ;using half-overlapping hamming windows of length 'N' samples . X, in frequency domain is also considered as "Cochleagram"[30] of the mixture x.

ii. The Magnitude spectrogram 'V' is obtained by taking the absolute value of the elements of X.

iii. The "Spectrogram" or "Power Spectrogram"[31] is procured by performing the squared magnitude of STFT (in Frequency domain) ;that is

$$\text{Power Spectrogram} = [V]^{\wedge 2} = |X|^{\wedge 2}.$$

iv. Then the auto-correlation matrix 'B' can be attained by computing the auto-correlation of each row of Power Spectrogram $[V]^{\wedge 2}$

v. The Beat Spectrum that is basically a measure of acoustic self similarity of 'x' is enumerated by calculating the mean over entire rows of 'B'. Beat Spectrum appraises the presence of beats, pitch and melody in the song corresponding to repetitive nature of back-ground ;given by:

Auto-Correlation Matrix 'B' is given by:

$$B(i, j) = \frac{1}{m-j+1} * \sum_{k=1}^{m-j+1} V(i,k)^{\wedge 2}.V(i,k+j-1)^{\wedge 2}$$

And,

$$b(j) = \frac{1}{n} \sum_{i=1}^{n} B(i,j) \qquad [4.2]$$

For i= 1, 2…….n (freq);
j= 1, 2……. . m (lag)

We conclude that if repeating patterns exist in audio mixture 'x', 'b' would form peaks that seem to be periodically repeating at different levels; thereby; divulging the hierarchical repetitive nature of the mixture.
The period of the beat spectrum can be concluded by estimating the period in which the beat spectrum has largest average accumulated energy over its harmonics.

(B) Modelling of the Repeating Segment Model

The beat spectrum is procured and estimated to roughly evaluate the period 'p' of the repetitive back-ground in audio mixture excerpts , the magnitude spectrum 'V' is then evenly fragmented into 't' segments of length 'p'. The repeating segment model is, thereby, illustrated as the element-wise median of 't' segments produced.

The computation of the repeating segment model 'S' [33-35] is given by:

$$S(i,l) = \text{median}\{V(i, l) + (k-1)p\} \qquad [4.3]$$

k=1, 2….t;

For i= 1, 2…..n (freq)
l= 1, 2…...p (time)
Where, p= repetitive period

t=No. of segments.

The ‘median’ is selected to lead to a better discrimination between repeating and non-repeating patterns. The formation of repetitive model is quintessential because the non-repeating and varied part i.e. voice has a dispersed and sporadic time-frequency representation compared with the periodic representation of accompanying music.

In the beam spectrum formation, time-frequency bins showing little deflection or irregularity at period ‘p’ would comprise the repeating music pattern and is apprehended by the median model [36-38].Accordingly ,the time –frequency bins of the magnitude spectrum having large deviations at period ‘p’ would imbibe a non-repeating vocal pattern and is separated by the median paragon.

(C)Extraction of the Repetitive Pattern and Calculation of Binary Mask

The repeating spectrogram ‘W’[36-38] is derived by taking the element-wise minimum between ‘S’ and each of the ‘t’ segments of the spectrogram ‘V’. The ‘minimum’ function is considered because it is based on the assumption that the non-negative spectrogram ‘V’ of the mixture x is same as that of a non-negative repeating spectrogram ‘W’ and non-negative non-repeating spectrogram (V-W).

$$W(i, l) + (k-1)\, p = \min \{S(i,l) , V(i,l +(k-1)p\} \qquad [23]$$

For i=1, 2…n;
l=1, 2….p and
k=1, 2….t

The cochleagram or soft time frequency mask ‘M’ is computed by normalizing W by V element-wise. the concept of time-frequency mask lies in the fact that are going to repeat at period ‘p’ in V will acquire values near ‘1’ on M during normalization will be weighted towards the repeating background and the time-frequency bins that are not likely to replicate at period ‘p’ in V will attain values near ‘0’ in ‘M’ and will be weighted towards verse fore-ground.

The determination of soft-time mask ‘M’ will be as in equation (15) as:

$$M(i,j) = \frac{W(i,j)}{V(i,j)} \qquad [4.4]$$

With M (i, j) € [0,1]

For i= 1, 2….n (freq)
j= 1, 2…m (time)

In the process of synthesis, the time-frequency mask is again applied to STFT 'X' of the mixture x. The estimated music signal is obtained by inversion of cochleagram (STFT). The estimated voice signal is obtained by subtracting the music signal in time domain from the mixture signal.

We have derived a binary TF mask by forcing time-frequency bins in M attaining values above a certain threshold € [0, 1] to 1 while the rest values are forced to 0. In other words, Apriori mask [41-43] will be 1 for a TF unit if the mixture energy is within 3 dB of premixed target speech energy, otherwise it is 0.

The indication of value 1 as the mask points on the fact that the acoustic energy in the corresponding unit during segregation contains mostly the target signal and should be retained and the mask value [44-45] of 0 indicates that energy in the corresponding unit should be eliminated.

The entire process of the REPET algorithm will be summarized as follows:

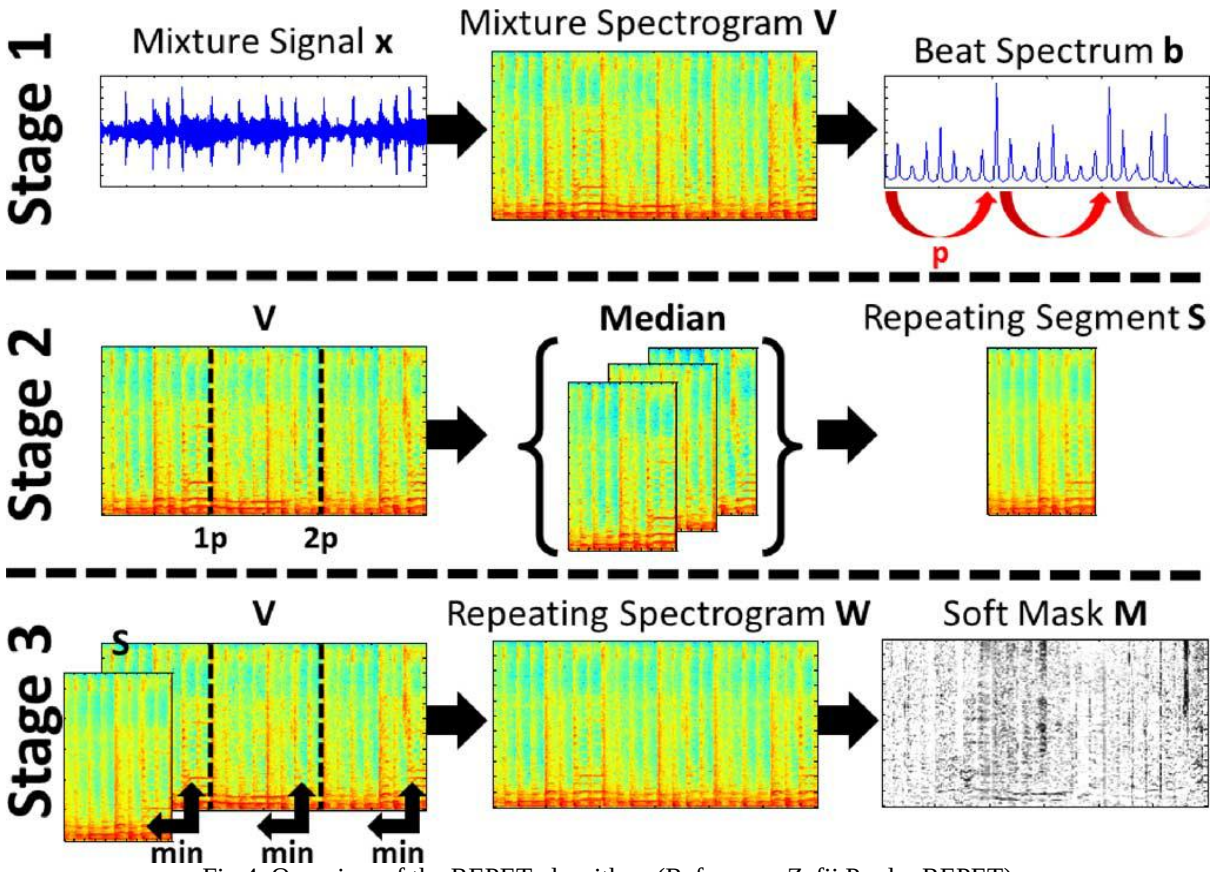

Fig.4. Overview of the REPET algorithm. (Reference: Zafii Pardo: REPET)

3.1 BEAT SPECTRUM FORMATION

The Beat Spectrogram visualizes evolution over successive windows. It demonstrates and picturizes rhythmic variations over successive windows in time. In other words,. Beat spectrogram is an image obtained by successive beat spectra. The beat spectrum can also be evolved through similarity matrix, which is basically a 2-Dimensional matrix that measures the similarity between two feature vectors ‘a’ and ‘b’ of a given sequence. In nut shell, the beat spectrum is considered as a measure of acoustic self similarity; as a function of time lag. Highly structured music will possess strong beat spectrum peaks at the repetitive periodic time, which describes the tempo, rhythm and strength of peaks.

The beat spectrum formation for jazz genre is as illustrated below:

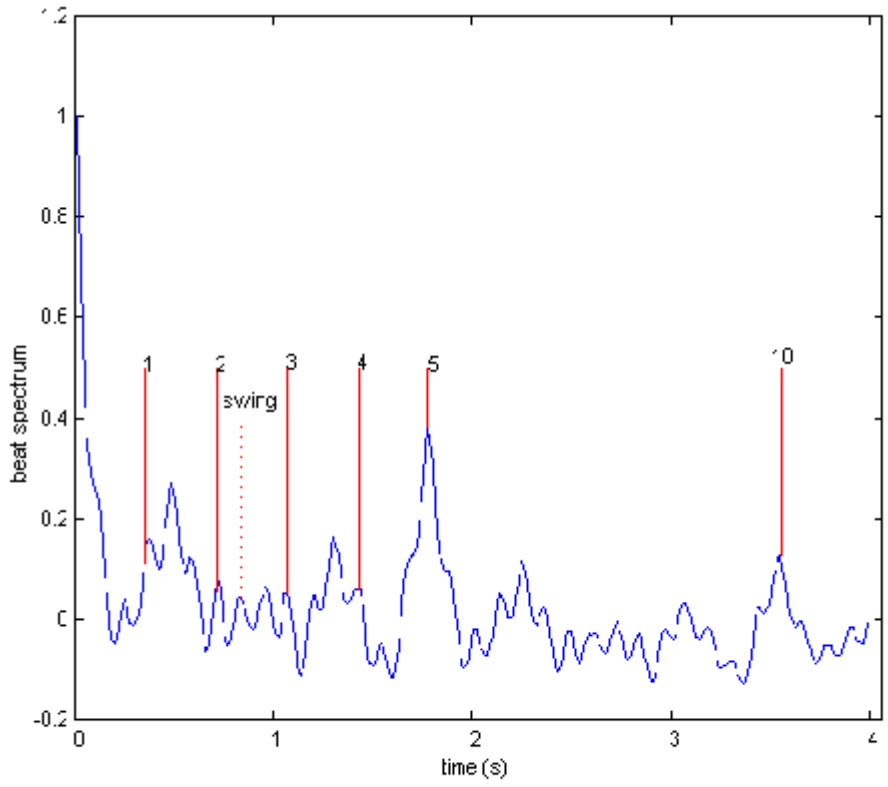

Figure Beat spectrum of jazz composition

Fig.4.1 Beat Spectrum of Jazz composition. (Ref: Beat Spectrum- A new approach to Rhythm analysis, IEEE 2001)

If some threshold value has been set, say 0.57; then we have considered a binary mask; if the value of soft weighing persists above 0.57 (threshold) value, it is set as 1, 0 otherwise.

Chapter 5: PERFORMANCE EVALUATION OF EXTENDED REPET ALGORITHM AND RESULTS.

5.1 PERFORMANCE EVALUATION:

The segregation of an audio mixture into vocal and non-vocal components can be adjudged via **ANOVA (Analysis Of Variation)** computation method as the music and voice local estimates are generated and we need to calculate and deduce their variation of points from the main audio excerpt mixture. The Performance evaluation can be evolved on the measure of **VARIANCE** and **ENERGY.** The first step in the separation process of vocals and non-vocals in audio include the Beat Spectrum generation.

The beat spectrum formation for different songs are as illustrated below:

- **Beat spectrum for Song1 (English Pop Song)**

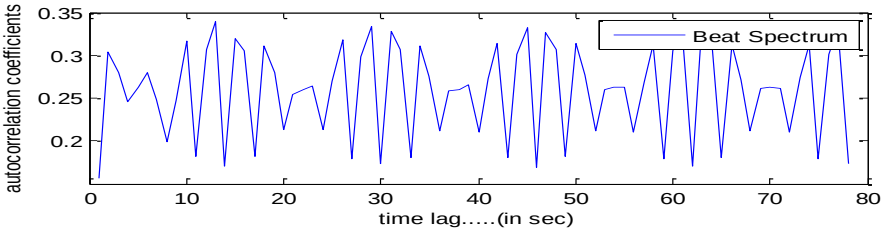

Figure 5.1(a): Figure showing beat spectrum for an English song

The beat spectrum showing the repetitive period as 14.The Y –axis of beat spectrum illustrating the auto-correlation coefficients and X- axis representing the time (in sec).

- **Beat spectrum for Song 2 (German Song):** The beat spectrum showing the repetitive period as 16

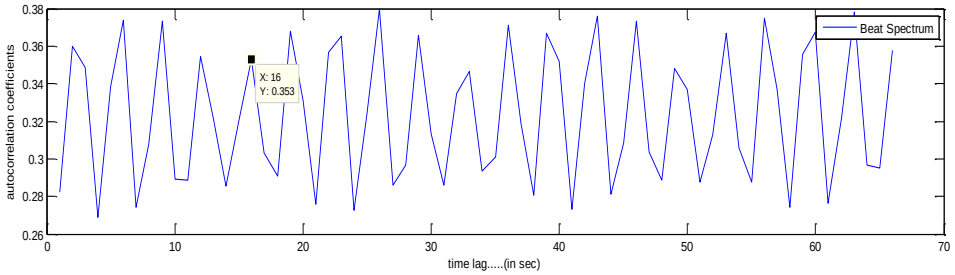

Figure 5.1(b): Figure showing beat spectrum for a German song

COCHLEAGRAM TF MASK FORMATION:

X-axis of TF mask represents Frequency bins, as it is derived from magnitude spectrum 'V' and Y axis represents the values of the normalized mask between 0 and 1. The different colour intensities show values of time frequency bins between 0 and 1.

- **TF Mask for Song 1(English Song)**

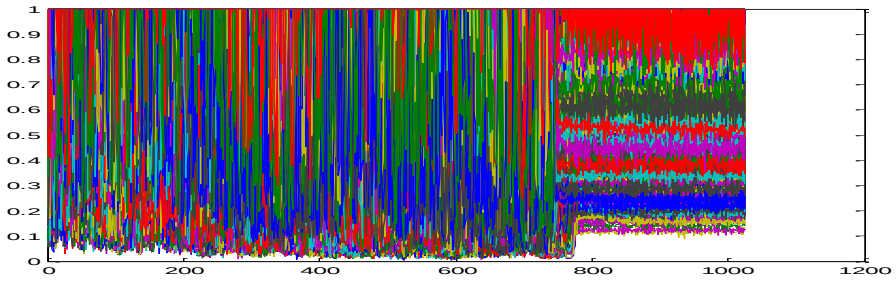

Figure 5.1(c): Figure showing cochleagram for song 1

- **TF Mask for Song 2 (German Song)**

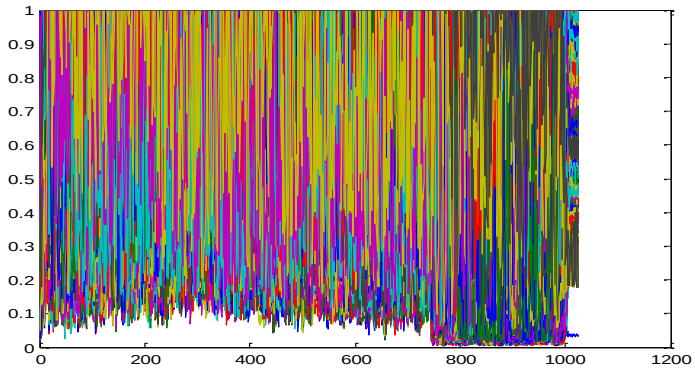

Figure 5.1(d): Figure showing cochleagram for song 2

If some threshold value has been set, say 0.57; then We have considered a binary mask; if the value of soft weighing persists above 0.57 (threshold) value ,it is set as 1 ,0 otherwise.

5.2 ANOVA ANALYSIS METHOD IN TERMS OF ENERGY AND VARIANCE

IMPORTANCE OF CALCULATION OF VARIANCE

In probability theory and statistics; variance measures how far a set of numbers is spread out. A variance of 'zero' indicates that all values are constant or identical. The value of variance is always non-negative. If the variance is small, the data-points or observations are very close to the mean (expected value) and a high variance illustrates that the data –points are spread out from the mean. The square-root of variance is denoted as the 'Standard Deviation'.

VARIANCE IN THE CALCULATION OF SIR

Since, we are considering the vocal and back-ground estimates and their deviation from the actual speech and music part contained in the mixture. Variance of a random variable is second central moment and is denoted as the squared deviation from the mean.

$$\text{Var}(X) = E(X-\mu)]\ \hat{}^{2} \qquad [5.1]$$

$$\text{Var}(X) = \text{Cov}(X, X) \qquad [5.2]$$

Where, Variance is also defined as the covariance of a signal with itself.

$$Var(X) = E[(X-E(X))^2] \quad [5.3]$$

$$=E[X^2-2XE(X) + (E(X))^2] \quad [5.4]$$

$$=E[X^2] - 2E[X]E[X] + (E[X])^2$$

$$=E[X^2] - (E[X])^2 \quad [5.5]$$

For continuous random variables, then variance can be given by:

$$Var(X) = \int (x-\mu)^2 . f(x) dx \quad [5.6]$$

$$= \int x^2 . f(x) dx - \mu^2 \quad [5.7]$$

Where, μ = mean or expected value of f(x)

And $\mu = \int x . f(x) dx$

For random variables, the variance can be understood by:

$$Var(X) = \sum_{i=1}^{n} (xi - \mu)^2 . pi$$

$$=\sum[pi.x^2] - \mu^2 \quad [5.8]$$

Where, $\mu = \sum pi . xi$

The ANOVA (Analysis Of Variation) technique can be used to calculate the estimates of fore-ground and back-ground during the segregation in order to determine how much energy and variance is present in these estimates measured after TF mask or formation of cochleagram.

The ANOVA computation method can be evaluated in the REPET algorithm while performing the parameterization using half-overlapping Hamming window.

RESULTS USING ANOVA ANALYSIS IN HAMMING WINDOWING

a. ON THE BASIS OF VARIANCE

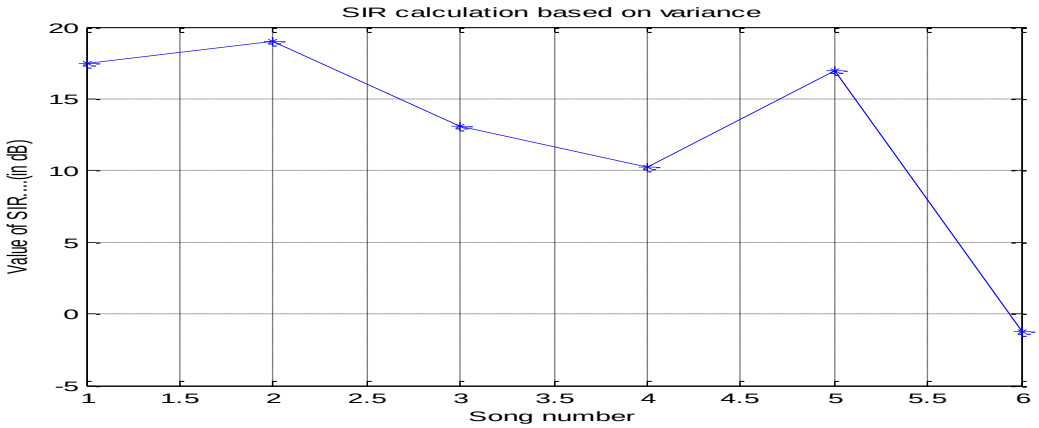

Figure 5.2(a): Figure showing performance evaluation of SIR value (in dB) versus song number in terms of variance.

Fig 5.2(a) illustrates the performance evaluation of SIR values versus different songs in terms of calculation of variance.

b. ON THE BASIS OF ENERGY :

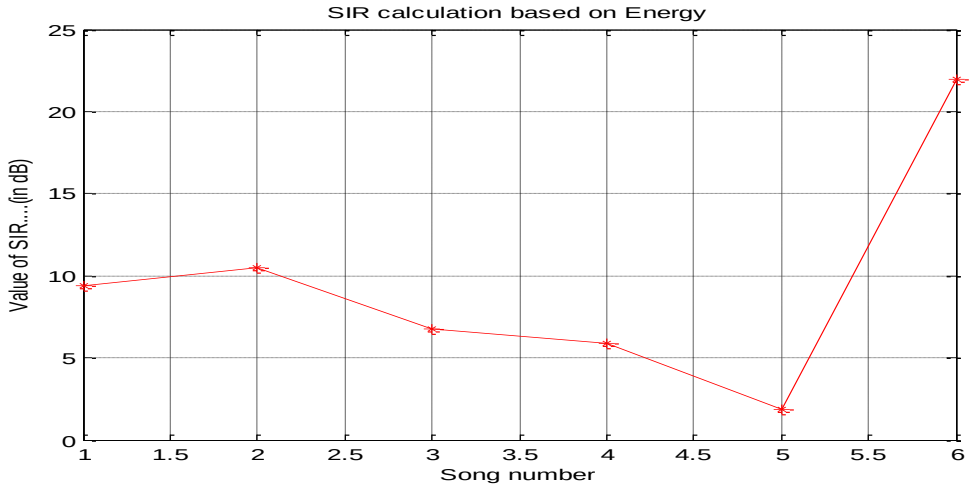

Figure 5.2(b): Figure showing performance evaluation of SIR value (in dB) versus song number in terms of energy calculation.

Fig 5.2(b) illustrates the performance evaluation of SIR values versus different songs in terms of calculation of energy.

5.3 MULTI-WINDOW COMPARISON OF SIR VALUES IN MONO-AURAL MUSIC SPEECH SEPARATION.

The segregation of an audio mixture into vocal and non-vocal components can be adjudged via ANOVA (Analysis Of Variation) computation method as the music and voice local estimates are generated and we need to calculate and deduce their variation of points from the main audio excerpt mixture. The Performance evaluation can be evolved on the measure of VARIANCE and ENERGY. The performance measure is SIR (Signal to Interference Ratio)[19,20] ,measured as ;given by equation [33]:

SIR Calculation in terms of Energy:

$$\text{SIR} = \frac{|Foreground|^2}{|Background|^2} \qquad [5.9]$$

Where, Foreground is the vocal energy (verse)

And Background is the music energy.

SIR Calculation in terms of Variance:

$$\text{SIR} = \frac{Var(var(Audio\ mixture))}{var(var(background))} \qquad [5.10]$$

SIR Calculation in terms of ANOVA analysis Method:

$$\text{SIR} = \frac{var(Foreground)}{var(background)} \qquad [5.11]$$

The ‘ANOVA’ paragon implies that the SIR value of any segregation method can be estimated by computation of variances of both foreground and background; as shown in equation (35).The variance is basically a parameter that represents the spreading of points. It denotes the deviation of certain points from the mean.

The Variance equation can be demonstrated by the equation (5.1); given by:

$$Var(X) = Cov(X, X);$$

Where, Cov (X) represents the covariance of X.

$$Var(X) = E(X)^2 - [E(X)]^2$$

Where E(X) is the expectation of the random variable ‘X’

5.4 RESULTS:

1. SIR CALCULATION OF MULTI-WINDOW BASED ON VARIANCE.

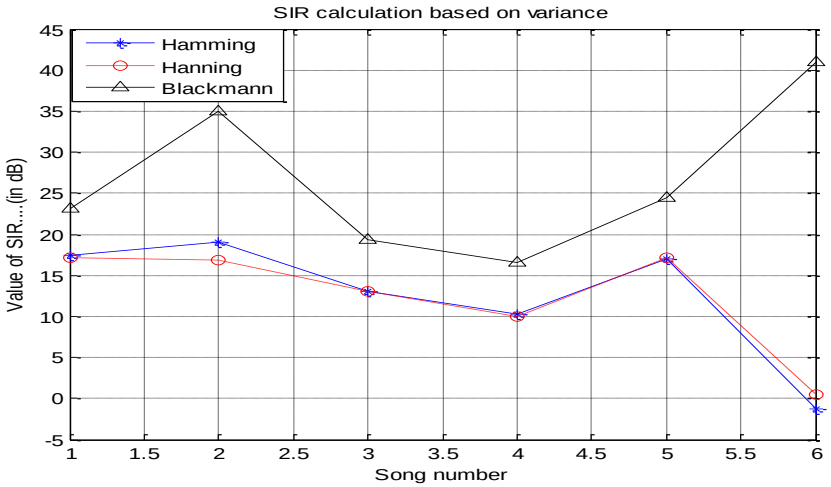

Figure 5.3(a): Figure showing performance evaluation of SIR value (in dB) versus song number in terms of variance for three different windows.

2. SIR CALCULATION OF MULTI-WINDOW BASED ON ENERGY CALCULATION

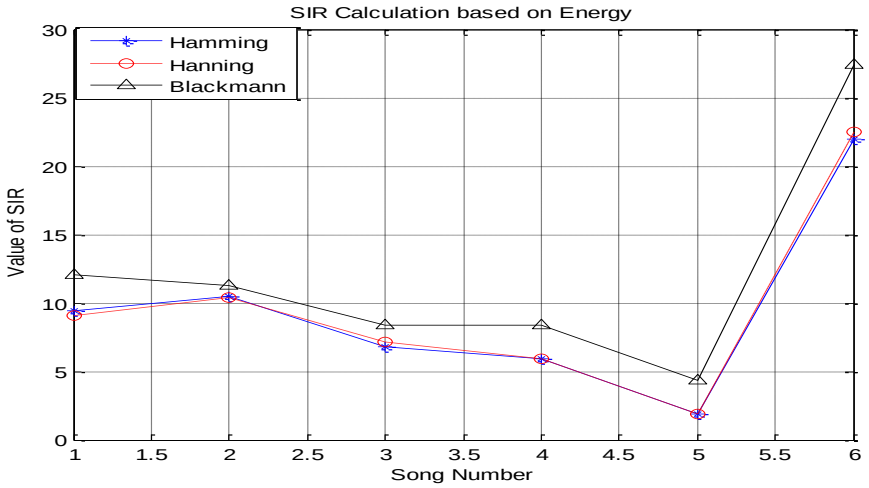

Figure 5.3(b): Figure showing performance evaluation of SIR value (in dB) versus song number in terms of energy calculation for three different windows

3. SIR CALCULATION OF MULTI-WINDOW BASED ON ANOVA ANALYSIS

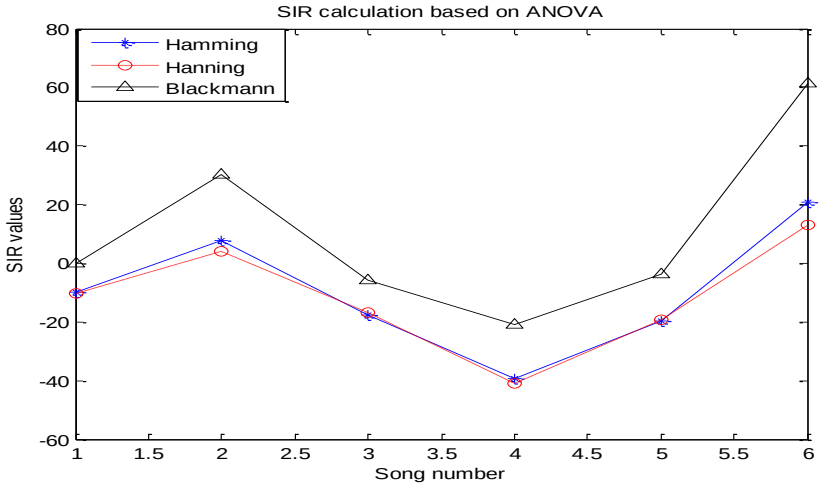

Figure 5.3(c): Figure showing performance evaluation of SIR value (in dB) versus song number by ANOVA analysis method for three different windows.

The performance evaluation in terms of Signal-to-Interference values for three different values (Hamming, Hanning and Blackmann Windows) show that Blackmann windows best quality segregation as compared to the other two in terms of all the three parameters, that is: Variance, Energy and ANOVA technique.

5.5 SUBJECTIVE TESTS: QUALITY TESTING OF SEPARATION OF VOCAL AND NON-VOCAL COMPONENTS USING MULTIPLE WINDOWS

ANALYSIS OF SEPARATION USING HAMMING WINDOWING: SONG 1

TABLE I: HAMMING WINDOW ANALYSIS FOR SONG 1.

No. of Persons	Very Good	Good	Average	Poor	Bad
1			✓		
2		✓			
3		✓			
4				✓	
5		✓			
6		✓			
7			✓		
8		✓			
9		✓			
10		✓			

Note: QUALITY PARAMETER'S VALUES – Very Good: 5, Good: 4, Average: 3, Poor: 2, Bad: 1

AVERAGE VALUE OF QUALITY USING HAMMING WINDOW FOR SONG 1: 3.6

song 1 is: **3.6**

ANALYSIS OF SEPARATION USING HANNING WINDOWING: SONG 1

TABLE II: HANNING WINDOW ANALYSIS FOR SONG 1.

No. of Persons	Very Good	Good	Average	Poor	Bad
1		✓			
2		✓			
3				✓	
4					✓
5			✓		
6			✓		
7			✓		
8		✓			
9			✓		
10		✓			

Note: QUALITY PARAMETER'S VALUES – Very Good: 5, Good: 4, Average: 3, Poor: 2, Bad: 1

AVERAGE VALUE OF QUALITY USING HANNING WINDOW FOR SONG 1: 3.0

The **average value** of quality while using the hanning window for segregation of audio for song 1 is: **3.0**

ANALYSIS OF SEPARATION USING BLACKMANN WINDOWING: SONG 1

TABLE III: BLACKMANN WINDOW ANALYSIS FOR SONG 1.

No. of Persons	Very Good	Good	Average	Poor	Bad
1	✓				
2		✓			
3		✓			
4		✓			
5	✓				
6			✓		
7	✓				
8			✓		
9		✓			
10			✓		

Note: QUALITY PARAMETER'S VALUES – Very Good: 5, Good: 4, Average: 3, Poor: 2, Bad: 1

AVERAGE VALUE OF QUALITY USING BLACKMANN WINDOW FOR SONG 1: 3.8

The **average value** of quality while using the Blackmann window for segregation of audio for song 1 is: **3.8**

INFERENCE: The blackmann windowing showing better quality of separation of audio song 1 as a result of hearing, as compared to hamming and hanning windows.

RESULT OF MULTI-WINDOW SEPARATION FOR SONG1

Blackmann Window > Hamming Window > Hanning window

ANALYSIS OF SEPARATION USING HAMMING WINDOWING: SONG 2

TABLE IV: HAMMING WINDOW ANALYSIS FOR SONG 2.

No. of Persons	Very Good	Good	Average	Poor	Bad
1		✓			
2		✓			
3		✓			
4		✓			
5		✓			
6			✓		
7			✓		
8			✓		
9			✓		
10			✓		

Note: QUALITY PARAMETER'S VALUES – Very Good: 5, Good: 4, Average: 3, Poor: 2, Bad: 1

AVERAGE VALUE OF QUALITY USING HAMMING WINDOW FOR SONG 2: 3.5

The **average value** of quality while using the Hamming window for segregation of audio for song 2 is: **3.5**

ANALYSIS OF SEPARATION USING HANNING WINDOWING: SONG 2

TABLE V: HANNING WINDOW ANALYSIS FOR SONG 2.

No. of Persons	Very Good	Good	Average	Poor	Bad
1					✓
2				✓	
3		✓			
4			✓		
5	✓				
6	✓				
7		✓			
8		✓			
9		✓			
10		✓			

Note: QUALITY PARAMETER'S VALUES – Very Good: 5, Good: 4, Average: 3, Poor: 2, Bad: 1

AVERAGE VALUE OF QUALITY USING HANNING WINDOW FOR SONG 2: 3.4

The **average value** of quality while using the Hanning window for segregation of audio for song 2 is: **3.4**

ANALYSIS OF SEPARATION USING BLACKMANN WINDOWING: SONG 2

TABLE VI: BLACKMANN WINDOW ANALYSIS FOR SONG 2.

No. of Persons	Very Good	Good	Average	Poor	Bad
1				✓	
2			✓		
3			✓		
4			✓		
5					✓
6	✓				
7	✓				
8			✓		
9			✓		
10			✓		

Note: QUALITY PARAMETER'S VALUES – Very Good: 5, Good: 4, Average: 3, Poor: 2, Bad: 1

AVERAGE VALUE OF QUALITY USING BLACKMANN WINDOW FOR SONG 2: 3.6

The **average value** of quality while using the Blackmann window for segregation of audio for song 2 is: **3.6**

INFERENCE: The blackmann windowing showing better quality of separation of audio song 2 as a result of hearing, as compared to hamming and hanning windows.

RESULT OF MULTI-WINDOW SEPARATION FOR SONG2

Blackmann Window > Hamming Window > Hanning window

ANALYSIS OF SEPARATION USING HAMMING WINDOWING: SONG 3

TABLE VII: HAMMING WINDOW ANALYSIS FOR SONG 3.

No. of Persons	Very Good	Good	Average	Poor	Bad
1	✓				
2				✓	
3					✓
4					✓
5			✓		
6				✓	
7	✓				
8				✓	
9				✓	
10				✓	

Note: QUALITY PARAMETER'S VALUES – Very Good: 5, Good: 4, Average: 3, Poor: 2, Bad: 1

AVERAGE VALUE OF QUALITY USING HAMMING WINDOW FOR SONG 3: 2

The **average value** of quality while using the Hamming window for segregation of audio for song 3 is: **2**

ANALYSIS OF SEPARATION USING HANNING WINDOWING: SONG 3

TABLE VIII: HANNING WINDOW ANALYSIS FOR SONG 3.

No. of Persons	Very Good	Good	Average	Poor	Bad
1		✓			
2		✓			
3			✓		
4			✓		
5				✓	
6				✓	
7				✓	
8				✓	
9		✓			
10		✓			

Note: QUALITY PARAMETER'S VALUES – Very Good: 5, Good: 4, Average: 3, Poor: 2, Bad: 1

AVERAGE VALUE OF QUALITY USING HANNING WINDOW FOR SONG 3: 3

The **average value** of quality while using the Hanning window for segregation of audio for song 3 is: **3**

ANALYSIS OF SEPARATION USING BLACKMANN WINDOWING: SONG 3

TABLE IX: BLACKMANN WINDOW ANALYSIS FOR SONG 3.

No. of Persons	Very Good	Good	Average	Poor	Bad
1				✓	
2				✓	
3					✓
4			✓		
5		✓			
6		✓			
7		✓			
8	✓				
9		✓			
10		✓			

Note: QUALITY PARAMETER'S VALUES – Very Good: 5, Good: 4, Average: 3, Poor: 2, Bad: 1

AVERAGE VALUE OF QUALITY USING BLACKMANN WINDOW FOR SONG 3: 3.4

The **average value** of quality while using the Blackmann window for segregation of audio for song 3 is: **3.4**

INFERENCE: The blackmann windowing showing better quality of separation of audio song 3 as a result of hearing, as compared to hamming and hanning windows.

RESULT OF MULTI-WINDOW SEPARATION FOR SONG 3

Blackmann Window > Hanning Window > Hamming window

ANALYSIS OF SEPARATION USING HAMMING WINDOWING: SONG 4

TABLE X: HAMMING WINDOW ANALYSIS FOR SONG 4.

No. of Persons	Very Good	Good	Average	Poor	Bad
1	✓				
2	✓				
3			✓		
4			✓		
5				✓	
6			✓		
7			✓		
8			✓		
9			✓		
10			✓		

Note: QUALITY PARAMETER'S VALUES – Very Good: 5, Good: 4, Average: 3, Poor: 2, Bad: 1

AVERAGE VALUE OF QUALITY USING HAMMING WINDOW FOR SONG 4: 3.3

The **average value** of quality while using the hamming window for segregation of audio for song 4 is: **3.3**

ANALYSIS OF SEPARATION USING HANNING WINDOWING: SONG 4

TABLE XI: HANNING WINDOW ANALYSIS FOR SONG 4.

No. of Persons	Very Good	Good	Average	Poor	Bad
1			✓		
2			✓		
3			✓		
4			✓		
5			✓		
6				✓	
7		✓			
8		✓			
9		✓			
10		✓			

Note: QUALITY PARAMETER'S VALUES – Very Good: 5, Good: 4, Average: 3, Poor: 2, Bad: 1

AVERAGE VALUE OF QUALITY USING HANNING WINDOW FOR SONG 4: 3.3

The **average value** of quality while using the hanning window for segregation of audio for song 4 is: **3.3**

ANALYSIS OF SEPARATION USING BLACKMANN WINDOWING: SONG 4

TABLE XII: BLACKMANN WINDOW ANALYSIS FOR SONG 4.

No. of Persons	Very Good	Good	Average	Poor	Bad
1			✓		
2			✓		
3		✓			
4		✓			
5		✓			
6		✓			
7		✓			
8		✓			
9		✓			
10		✓			

Note: QUALITY PARAMETER'S VALUES – Very Good: 5, Good: 4, Average: 3, Poor: 2, Bad: 1

AVERAGE VALUE OF QUALITY USING BLACKMANN WINDOW FOR SONG 4: 3.6

The **average value** of quality while using the Blackmann window for segregation of audio for song 4 is: **3.6**

INFERENCE: The blackmann windowing showing better quality of separation of audio song 4 as a result of hearing, as compared to hamming and hanning windows.

RESULT OF MULTI-WINDOW SEPARATION FOR SONG 4

Blackmann Window > Hanning Window = Hamming window

ANALYSIS OF SEPARATION USING HAMMING WINDOWING: SONG 5

TABLE XIII: HAMMING WINDOW ANALYSIS FOR SONG 5.

No. of Persons	Very Good	Good	Average	Poor	Bad
1	✓				
2	✓				
3	✓				
4	✓				
5		✓			
6		✓			
7		✓			
8			✓		
9			✓		
10				✓	

Note: QUALITY PARAMETER'S VALUES – Very Good: 5, Good: 4, Average: 3, Poor: 2, Bad: 1

AVERAGE VALUE OF QUALITY USING HAMMING WINDOW FOR SONG 5: 4.0

The **average value** of quality while using the Hamming window for segregation of audio for song 5 is: **4.0**

ANALYSIS OF SEPARATION USING HANNING WINDOWING: SONG 5

TABLE XIV: HANNING WINDOW ANALYSIS FOR SONG 5.

No. of Persons	Very Good	Good	Average	Poor	Bad
1		✓			
2		✓			
3		✓			
4		✓			
5		✓			
6			✓		
7			✓		
8				✓	
9				✓	
10				✓	

Note: QUALITY PARAMETER'S VALUES – Very Good: 5, Good: 4, Average: 3, Poor: 2, Bad: 1

AVERAGE VALUE OF QUALITY USING HANNING WINDOW FOR SONG 5: 3.2

The **average value** of quality while using the Hanning window for segregation of audio for song 5 is: **3.2**

ANALYSIS OF SEPARATION USING BLACKMANN WINDOWING: SONG 5

TABLE XV: BLACKMANN WINDOW ANALYSIS FOR SONG 5.

No. of Persons	Very Good	Good	Average	Poor	Bad
1		✓			
2		✓			
3		✓			
4		✓			
5		✓			
6		✓			
7	✓				
8		✓			
9		✓			
10		✓			
Note: QUALITY PARAMETER'S VALUES – Very Good: 5, Good: 4, Average: 3, Poor: 2, Bad: 1					

AVERAGE VALUE OF QUALITY USING BLACKMANN WINDOW FOR SONG 5: 4.1

The **average value** of quality while using the Blackmann window for segregation of audio for song 5 is: **4.1**

INFERENCE: The blackmann windowing showing better quality of separation of audio song 5 as a result of hearing, as compared to hamming and hanning windows.

RESULT OF MULTI-WINDOW SEPARATION FOR SONG 5

Blackmann Window > Hamming Window > Hanning window

ANALYSIS OF SEPARATION USING HAMMING WINDOWING: SONG 6

TABLE XVI: HAMMING WINDOW ANALYSIS FOR SONG 6.

No. of Persons	Very Good	Good	Average	Poor	Bad
1				✓	
2				✓	
3				✓	
4			✓		
5			✓		
6			✓		
7		✓			
8		✓			
9		✓			
10		✓			

Note: QUALITY PARAMETER'S VALUES – Very Good: 5, Good: 4, Average: 3, Poor: 2, Bad: 1

AVERAGE VALUE OF QUALITY USING HAMMING WINDOW FOR SONG 6: 3.1

The **average value** of quality while using the Hamming window for segregation of audio for song 6 is: **3.1**

ANALYSIS OF SEPARATION USING HANNING WINDOWING: SONG 6

TABLE XVII: HANNING WINDOW ANALYSIS FOR SONG 6.

No. of Persons	Very Good	Good	Average	Poor	Bad
1	✓				
2			✓		
3		✓			
4				✓	
5	✓				
6	✓				
7		✓			
8				✓	
9				✓	
10		✓			

Note: QUALITY PARAMETER'S VALUES – Very Good: 5, Good: 4, Average: 3, Poor: 2, Bad: 1

AVERAGE VALUE OF QUALITY USING HANNING WINDOW FOR SONG 6: 3.6

The **average value** of quality while using the Hanning window for segregation of audio for song 6 is: **3.6**

ANALYSIS OF SEPARATION USING BLACKMANN WINDOWING: SONG 6

TABLE XVIII: BLACKMANN WINDOW ANALYSIS FOR SONG 6.

No. of Persons	Very Good	Good	Average	Poor	Bad
1	✓				
2	✓				
3		✓			
4			✓		
5	✓				
6	✓				
7		✓			
8			✓		
9			✓		
10		✓			

Note: QUALITY PARAMETER'S VALUES – Very Good: 5, Good: 4, Average: 3, Poor: 2, Bad: 1

AVERAGE VALUE OF QUALITY USING BLACKMANN WINDOW FOR SONG 6: 4.1

The **average value** of quality while using the Blackmann window for segregation of audio for song 6 is: **4.1**

INFERENCE: The blackmann windowing showing better quality of separation of audio song 6 as a result of hearing, as compared to hamming and hanning windows.

RESULT OF MULTI-WINDOW SEPARATION FOR SONG 6

Blackmann Window > Hanning Window > Hamming window

Chapter 6: CONCLUSION AND FUTURE WORK

CONCLUSION

In the proposed work, we have evaluated the quality and performance measure of repeating music and superimposed varying vocals via the method of calculating variance named as 'ANOVA' computation procedure on REPET (REpeating Pattern Extraction Technique) separation algorithm. The main quintessence of REPET algorithm is to identify and extract the repeating musical patterns formed in beat spectrum by comparing them to a repeating segment model. ANOVA method is a statistical technique to scrutinize variation in response variable measured under conditions explicated by discrete factors. The performance evaluation results are generated in MATLAB 7.0 procuring SIR values based on energy values of foreground and background clips. Experiments on a data set of six different audio clip excerpts applying Hamming windowing, Hanning window and blackmann windowing has proved that the best separation performance in terms of SIR values for each and every one has been more in blackmann windowing as compared to hamming and hanning windows. That is, best quality music and vocals can be observed when blackmann window is applied to audio excerpts before performing short-time Fourier transform. Separation quality for both hamming and hanning window is the same when their SIR values are being compared.

FUTURE WORK:

In the future work, the **Formant structure analysis** of audio is performed to do the separation of vocal and non-vocal components. **Weiner filtering** technique can also be employed as it is most useful in estimation theory to generate the estimates of any random signal, to produce the final mask. REPET algorithm can be extended for the analysis where the music patterns **are not rhythmic** and thus, it's not repeating at regular intervals.

REFERENCES

1. Wold, E., Blum, T., Keislar, D., and Wheaton, J., "Classification, Search and Retrieval of Audio," in Handbook of Multimedia Computing, ed. B. Furht, pp. 207-225, CRC Press,1999.

2. Foote, J., "Automatic Audio Segmentation using a Measure of Audio Novelty," in *Proc. ICME* 2000.

3. Scheirer, E., "Tempo and Beat Analysis of Acoustic Musical Signals," in *J. Acoust. Soc. Am.* 103(1), Jan 1998, pp 588-601

4. Goto, M., and Muraoka, Y., "A Beat Tracking System for Acoustic Signals of Music," in *Proc. ACM Multimedia 1994*, San Francisco, ACM

5. Scheirer, E., "Using Musical Knowledge to Extract Expressive Performance Information from Audio Recordings," in *Readings in Computational Auditory Scene Analysis*, eds. Rosenthal and Okuno, Lawrence Erlbaum, 1998.

6. Foote, J., "Visualizing Music and Audio using Self-Similarity," in *Proc. ACM Multimedia 99*, Orlando, FL, pp. 70-80.

7. Rabiner, L., and Juang, B.-H., *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ, 1993

8. Foote, J., "Content-Based Retrieval of Music and Audio," in *Multimedia Storage and Archiving Systems II, Proc. SPIE,* Vol. 3229, Dallas, TX. 1997.

9. D. L.Wang and G. J. Brown, "Separation of speech from interfering sounds using oscillatory correlation," IEEE Trans. Neural Networks, vol. 10, pp. 684–697, Sep. 1999.

10. S. T. Roweis, "One microphone source separation," Advances in Neural Information Processing Systems, vol. 13, pp. 793–799, 2000.

11. E. A. Wan and A. T. Nelson, "Neural dual extended Kalman filtering: Applications in speech enhancement and monaural blind signal separation," in IEEE Workshop on Neural Networks and Signal Processing, Amelia Island, FL, USA, 1997, pp. 466–475

12. . G.-J. Jang and T.-W. Lee, "A probabilistic approach to single channel source separation," Advances in Neural Information Processing Systems, vol. 15, 2003.

13. B. A. Pearlmutter and A. M. Zador, "Monaural source separation using spectral cues," in Proc. 5th International Conference on Independent Component Analysis and Blind Source Separation(ICA'04), Granada, Spain, Sep. 2004.

14. T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in International Computer Music Conference (ICMC'03), Singapore, Sep. 2003

15. L. Benaroya, F. Bimbot, and R. Gribonval., "Audio source separation, S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook with a single sensor," IEEE Trans. Audio, Speech and Language Processing, vol. 14, no. 1, pp. 191–199, Jan. 2006.

16. Codebook-based Bayesian speech enhancement," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'05), Philadelphia, USA, March 2005, pp. 1077–1080.

17. L. Benaroya, R. Gribonval, and F. Bimbot, “Non negative sparse representation for Wiener based source separation with a single sensor,” in Proc. IEEE International Conference on Acoustics, Speech Signal Processing (ICASSP’03), Hong Kong, 2003, pp. 613–616.

18. A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” Journal of the Royal Statistical Society. Series B., vol. 39, pp. 1–38, 1977.
19. R. Gribonval, L. Benaroya, E. Vincent, and C F´evotte, “Proposals for performance measurement in source separation,” in Proc. 4th Symposium on Independent Component Analysis and Blind Source Separation (ICA’03), Nara, Japan, Apr. 2003.

20. Rafii, Z.; Pardo, B., "REpeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation," Audio, Speech, and Language Processing, IEEE Transactions on , vol.21, no.1, pp.73,84, Jan. 2013 H. [2]Schenker, Harmony. Chicago, IL: Univ. of Chicago Press, 1954.
21. A. Ockelford, Repetition in Music: Theoretical and Met theoretical Perspectives. Farnham, U.K.: Ash gate, 2005, vol. 13, Royal Musical Association Monographs.

22. Dubnov, S., Tabrikian, J., & Arnon-Targan, M. (2004, May) A method for directionally-disjoint source separation in convolutive environment in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (Vol. V, pp. 489-492), Montreal, Quebec, Canada.

23. M. Cooper and J. Foote, “Automatic music summarization via similarity analysis,” in Proc. 3rd Int. Conf. Music Inf. Retrieval, Paris, France, Oct. 13–17, 2002, pp. 81–85.

24. A. Pikrakis, I. Antonopoulos, and S. Theodoridis, “Music meter and tempo tracking from raw polyphonic audio,” in Proc. 9th Int. Conf. Music Inf. Retrieval, Barcelona, Spain, Oct. 10–14, 2008.

25. Kolossa, D., Klimas, A., & Orglmeister, R. (2005, October).Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques in *Proceedings of the IEEE* Workshop on Applications of Signal Processing to Audio and Acoustics (pp. 82-85), New Paltz, NY.

26. . Foote, “Automatic audio segmentation using a measure of audio novelty,” in *Proc*. IEEE Int. Conf. Multimedia and Expo, New York, Jul.-Aug. 30–02, 2000, vol. 1, pp. 452–455.

27. Kates, J. M., & Arehart, K. H. (2005), Multichannel dynamic-range compression using digital frequency warping in EURASIP Journal on Applied Signal Processing, 18, 3003-3014.

28. A. Liutkus and P. Leveau, “Separation of music+ effects sound track from several international versions of the same movie,” in Proc. 128th Audio Eng. Soc. Conv., London, U.K., May 22–25, 2010.

29. R. J. Weiss and J. P. Bello, “Identifying repeated patterns in music using sparse convolutive non-negative matrix factorization,” in Proc. 11th Int. Soc. Music Inf. Retrieval, Utrecht, The Netherlands, Aug. 9–13, 2010.

30. B. Fox, A. Sabin, B. Pardo, and A. Zopf, “Modelling perceptual similarity of audio signals for blind source separation evaluation,” in Proc. 7th Int. Conf. Ind. Compon. Anal., London, U.K., Sep. 09–12, 2007, pp. 454–461.

31. Madhu, N., Breithaupt, C., & Martin, R. (2008). Temporal smoothing of spectral masks in the cepstral domain for Speech separation. in Proceedings of the IEEE International Conference on

32. M. A. Bartsch, “To catch a chorus using chroma-based representations for audio thumb nailing,” in *Proc. IEEE Workshop Application. Signal Process. Audio Acoustics.*, New Paltz, NY, Oct. 21–24, 2001, pp. 15–18.

33. R. B. Dannenberg and N. Hu, “Pattern discovery techniques for music audio,” *J. New Music Res.*, vol. 32, no. 2, pp. 153–164, 2003.

34. K. Jensen, “Multiple scale music segmentation using rhythm, timbre, and harmony,” *EURASIP J. Adv. Signal Process*, vol. 2007, no. 1, pp.1–11, Jan. 2010.

35. R. B. Dannenberg, “Listening to “Naima”: An automated structural analysis of music from recorded audio,” in *Proc. Int. Computation Music Conf.*, Gothenburg, Sweden, Sep. 17–21, 2002, pp. 28–34.

36. R. B. Dannenberg and M. Goto, “Music structure analysis from acoustic signals,” in *Handbook of Signal Processing in Acoustics*, D.Havelock, S. Kuwano, and M. Vorländer, Eds. New York: Springer, 2009, pp. 305–331.

37. J. Paulus, M. Müller, and A. Klapuri, “Audio-based music structure analysis,” in *Proc. 11th Int. Soc. Music Inf. Retrieval*, Utrecht, The Netherlands, Aug. 9–13, 2010, pp. 625–636.

38. S. Vembu and S. Baumann, “Separation of vocals from polyphonicaudio recordings,” in *Proc. 6th Int. Conf. Music Inf. Retrieval*, London, U.K., Sep. 11–15, 2005, pp. 337–344.

39. . Raj, P. Smaragdis, M. Shashanka, and R. Singh, “Separating a foreground singer from background music,” in *Proc. Int. Symp. Frontiers of Res. Speech and Music*, Mysore, India, May 8–9, 2007.

40. Y. Li and D. Wang, “Separation of singing voice from music accompaniment for monaural recordings,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1475–1487, May 2007.

41. M. Ryynänen, T. Virtanen, J. Paulus, and A. Klapuri, “Accompaniment separation and karaoke application based on automatic melody transcription,” in *Proc. IEEE Int. Conf. Multimedia & Expo*, Hannover, Germany, Jun. 23–26, 2008, pp. 1417–1420.

42.] M. Ryynänen, T. Virtanen, J. Paulus, and A. Klapuri, “Accompaniment separation and karaoke application based on automatic melody transcription,” in *Proc. IEEE Int. Conf. Multimedia & Expo*, Hannover, Germany, Jun. 23–26, 2008, pp. 1417–1420.

43. . Virtanen, A. Mesaros, and M. Ryynänen, “Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music,” in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, Brisbane, Australia, Sep. 21, 2008, pp. 17–20.

44. K. Dressler, “An auditory streaming approach on melody extraction,” in *Proc. 7th Int. Conf. Music Inf. Retrieval (MIREX Eval.)*, Victoria, BC, Canada, Oct. 8–12, 2006.

45. C.-L. Hsu and J.-S. R. Jang, “On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 310–319, Feb

Acoustics, Speech and Signal Processing (pp. 45-48), Las Vegas, NV.