High Resolution Property of Group Delay and its Application to Musical Onset Detection on Carnatic Percussion Instruments

A Project Report

submitted by

P A MANOJ KUMAR

in partial fulfilment of the requirements
for the award of the degree of

MASTER OF TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.

MAY 2015

THESIS CERTIFICATE

This is to certify that the thesis titled **High Resolution Property of Group Delay and its Application to Musical Onset Detection on Carnatic Percussion Instruments**, submitted by **P A Manoj Kumar**, to the Indian Institute of Technology, Madras, for the award of the degree of **Master of Technology**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. C S Ramalingam
Research Guide
Professor
Dept. of Electrical Engineering
IIT-Madras, 600 036

Prof. Hema A Murthy
Research Guide
Professor
Dept. of Computer Science
IIT-Madras, 600 036

Place: Chennai

Date:

ACKNOWLEDGEMENTS

The contents of this thesis would not have been possible without my guide Prof. Hema A Murthy. Her immense enthusiasm kept me motivated whenever research seemed a difficult endeavour. She has always been ready to hear out her students' ideas, something which enabled me a lot of freedom while pursuing my research. Amidst a tight schedule, she also has time to spare whenever I approached for my research. Whatever progress I have made from someone oblivious to the world of research to this moment, I attribute it directly and indirectly to her. I thank my committee members, especially Prof. C S Ramalingam, who introduced me to Prof. Murthy and gave me the liberty to chose a project of my choice.

I am grateful to my friends at DON Lab for the wonderful time and support, especially Raghava Krishnan who stayed with me in the lab through my first conference paper. Last but not least, I thank my parents for letting me pursue my interests including the decision for PhD, which unfortunately have always happened at the cost of time we could have spent together. I always received never ending support from home throughout my college and research life, a privilege that few people have.

ABSTRACT

KEYWORDS: Phase based processing; Group delay functions; High resolution property; Musical onset detection; Music Information Retrieval; Carnatic percussion instruments

Spectral analyses of speech and music processing have largely looked at the magnitude spectrum, while completely ignoring the phase component. The importance of phase on perception has been established through experiments on human listeners, but the wrapped form of phase has made it's processing difficult. More recently, the negative derivative of phase, or group delay function was successfully applied to various tasks in speech and music processing. The success of the group delay function and it's derived forms is attributed to the additive and high resolution property of group delay.

While the additive property is a well understood phenomenon, relatively little work has gone into understanding the high resolution property of group delay functions. In this thesis, a proof for the high resolution property is presented. It is shown that the group delay functions for various configurations of a single-resonator system have sharper peaks when compared to the magnitude spectra. Specifically, the strength of group delay function at the ndB bandwidth of magnitude spectrum is always lesser. This property is validated by numerical measures that quantify the peakedness of group delay functions - kurtosis and spectral flatness. A specific case is considered using the 3dB bandwidth and extended to multi-pole systems through numerical computations.

In the second part of the thesis, a novel method for musical onset detection is proposed. Onset detection is concerned with locating instants of significance in a musical recording, and is a key low-level description task in Music Information Retrieval (MIR). The proposed algorithm employs the high resolution property proven in the thesis. The music signal is treated as an amplitude-frequency modulated waveform, and the information about onsets are postulated to lie in the message signals. A demodulation technique using Hilbert transform is applied to extract the envelope. Minimum-phase group delay of the envelope is computed and a global peak picking threshold is applied to locate the onsets.

Experiments are performed on a large and varied database of Carnatic percussion instruments created through the course of this work. The proposed algorithm is shown to have an. F-measure of 0.93, and performs comparable to the state of the art machine learning technique. Later a method to reduce the parameter dependency of the algorithm without affecting the performance, is presented. Finally, the possibility of extending group delay processing to a more generic onset detection algorithm is explored.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

MIR	Music Information Retrieval
AM	Amplitude Modulation
FM	Frequency Modulation
STFT	Short Time Fourier Transform
ASR	Automatic Speech Recognition
TTS	Text-to-Speech

CHAPTER 1

Overview of Thesis

1.1 Introduction

Due to time varying properties of vocal characteristics associated with speech signals, it becomes necessary to divide a speech utterance into overlapping *frames*. Commonly known as short-term processing, it forms the beginning step for almost all speech and music processing applications. The speech signal is assumed to possess the same spectro-temporal characteristics within a frame - this is known as the *quasi-stationarity* property. Most applications in speech technology begin with spectral analyses that have considered the magnitude spectrum (or equivalently, the power spectrum) of the frame after a suitable windowing operation. For instance, filter bank energies computed from the magnitude spectrum and cepstral coefficients derived from filter bank energies are two commonly used features for speech recognition and verification tasks.

While research in speech technology has been primarily directed towards the processing of amplitude of the different frequency components (magnitude spectrum), the phase spectrum is often ignored. The question of whether or not phase spectrum plays a dominant role is to be addressed, but atleast for a minimum-phase signal, the phase spectrum has been shown to contain the same information as the magnitude spectrum ([37]). Closer to speech, the importance of phase in automatic speech recognition has been established in [1] through experiments on human listeners. The intelligibility of speech synthesised purely using magnitude spectrum or phase spectrum was found to be the same for shorter utterances and more for phase spectrum for longer utterances. [46] showed under various signal-to-noise ratios that having random phases for each frequency significantly altered the recognition rate as compared to true (and reconstructed) phase. Later, [20], [41] and [6] extracted features from the frequency derivative of phase spectrum and employed it in speech recognition.

In spite of the importance of phase being known, the main reason for the properties of phase not been fully exploited lies in the difficulty of processing it's wrapped nature (restricted within the principal values $[-\pi, \pi]$) ([49]). The problem of recovering the true phase given the wrapped form is an ill-posed problem if phase continuity is not considered. [53]. This problem can be overcome when the frequency derivative of phase is considered (specifically, the negative frequency derivative), more commonly known as group delay functions. In the last 30 years, the group delay function in it's original as well as derived forms as the minimum-phase group delay function ([35]) and the modified group delay function ([32]) have been exploited in tasks like pitch and formant estimation, syllable segmentation, speech recognition, speaker verification, etc [33] and shown to perform comparably and sometimes better than traditional magnitude spectrum based approaches. Two reasons have been presented to support the superior performance - the additive property and the high spectral resolution of group delay functions.

For a cascade of resonators, the resultant phase spectrum is a sum of the individual phases, while the magnitude spectrum is multiplicative:

$$h(t) = h_1(t) * h_2(t) \tag{1.1}$$

where $h_1(t)$ and $h_2(t)$ denote the individual response functions and $h(t)$ is the combined response function. In the frequency domain

$$|H(\omega)|e^{\angle H(\omega)} = |H_1(\omega)|e^{\angle H_1(\omega)}|H_2(\omega)|e^{\angle H_2(\omega)} \tag{1.2}$$

Equivalently,

$$\angle H(\omega) = \angle H_1(\omega) + \angle H_2(\omega) \tag{1.3}$$

The above property is relevant in the context of speech, where the resonances in vocal tract are modelled as formants in the frequency spectrum. The additive nature is argued to assist in discriminating the peaks in the spectrum and is a relatively well understood phenomenon.

While the additive property of phase explains the discriminating power between multiple poles, the implicit peakedness of each peak (similarly, of an isolated peak) has not been proven in existing literature. This is explored in the first part of the work.

1.2 High Resolution Property of Group Delay

The peakedness [1] of group delay function is illustrated using the magnitude spectrum for comparison. Frequency analysis has always been carried out using the magnitude spectrum, hence it forms an ideal measure for comparison purposes. We consider peaks due to system poles, and argue that the discriminating power of magnitude spectrum/group delay is inversely related to the bandwidth of the peaks.

First, a single-resonator system is considered, and it is proved that the group delay function has lesser ndB bandwidth than the magnitude spectrum, irrespective of the pole location. The result assures that at any frequency starting from the resonance, the group delay function will be lesser in strength than the magnitude spectrum. The proof is validated using numerical measures for a single (kurtosis and spectral flatness) as well as multi-pole system (numerical 3dB and acceleration). Effect of windowing on the spectrum is also studied, and it is shown that the discriminating power of group delay functions is unaltered for all configurations of the poles.

1.3 Musical Onset Detection using Group Delay

In the second part of the work, a novel application of group delay functions is proposed, which directly employs the high spectral resolution and makes use of it's minimum-phase equivalent form. A new onset detection algorithm is presented, specifically for the case of Carnatic[2] accompaniments. Similar to syllable segmentation, events are required to be located with high temporal accuracy in onset detection. The proposed method is evaluated on a large dataset of Carnatic percussion instruments and shown to perform comparable to a state of the art algorithm involving machine learning technique.

In the following section, a brief introduction to Carnatic music is presented alongwith the major percussion accompaniments. The task of onset detection is formally defined and the need for an algorithm with high accuracy is motivated.

[1] The terms peakedness and high resolution property will be used interchangeably in the rest of this work

[2] South Indian classical

Percussion Instruments in Indian Classical Music

Percussion instruments play an important role in many genres of world music. In addition to keeping track of rhythm, they have been developed to produce individual performances rich in artistic quality. Carnatic music constitutes one of the sub-genres of Indian classical music, the other being Hindustani music. A Carnatic performance (*concert*) comprises of an ensemble of the vocalist, and percussion and non-percussion accompaniments. Although most compositions are meant to be sung and the emphasis is on the vocalist, accompaniments have been used in complementing the vocalist as well as in exhibiting improvisations during the *tani-avarthanams* (solo percussion).

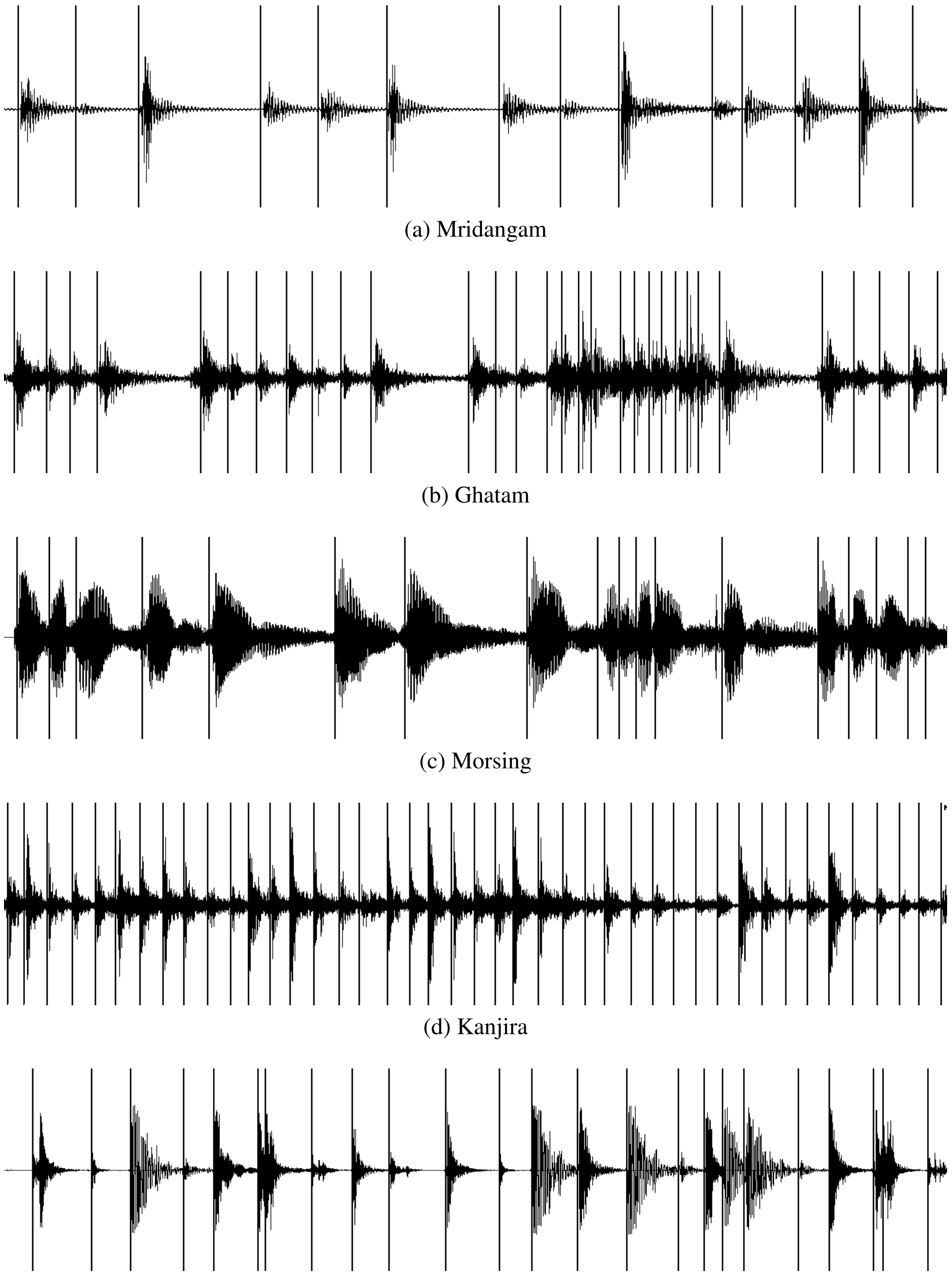
The major percussive accompaniments to Carnatic music include the mridangam, ghatam, kanjira, morsing and the thavil. The choice between these instruments is made based on the rapport between the lead and accompanying artists and nature of the performance itself. The thavil, for example, is widely found in musical performances associated with traditional festivals and ceremonies as well as in professional musical concerts.

Excerpts from the above instruments, each ≈ 3 seconds in duration are presented in Fig. 1.1. Mridangam is observed to have the slowest tempo [3]. A typical mridangam *tani avarthanam* contains a large number of silent strokes. Combination strokes exist, in which the left and right sides of mridangam are (expected to be) struck at the same instant. Ghatam strokes do not show significant amplitude variation but are played at a faster tempo. Morsing strokes lack a sharp onset. Being a wind percussion instrument, the vocal cavity is involved as the resonator in sound production. Kanjira is played at the fastest tempo among all the instruments - one possible reason being the smaller size and lighter weight of kanjira, thus enabling it to be struck the fastest. A brief description about each of these intruments can be found in Appendix A.

Task Motivation

A study of Carnatic percussion instruments on the lines of Music Information Retrieval (MIR) will be aimed at understanding the patterns involved in the performance - beat tracking, stroke classification, phrase/syllable classification etc. The information ob-

[3] On an average



(a) Mridangam

(b) Ghatam

(c) Morsing

(d) Kanjira

(e) Thavil

Figure 1.1: Waveforms of Carnatic percussion instruments with onsets marked

tained can be further extended to study higher level problems such as $t\bar{a}la$ classification, *sama* detection, artist and *song* recognition and so on.

As a first step towards automating the above tasks, it becomes necessary to accurately pinpoint the location of each of the strokes produced. For instance, the time instants thus obtained can enable the discovery of patterns/cycles that characterise the

$t\bar{a}la$ or artist and mark the end of a *tani-avarthanam* (*mohra*). In another application, waveforms can be segmented into strokes and used for training using Baum Welch re-estimation in isolated-style model for stroke recognition. Another idea would be to cluster strokes into patterns and perform pattern clustering on the output stroke transcription. A reliable onset detector forms the pre-processing step in such applications.

Task Definition

A percussion stroke in isolation can be divided into three parts - onset, attack and decay (Fig 1.2). The onset signifies the exact beginning of stroke activity and is a time instant. The attack and decay denote the rise and fall of activity respectively and are time periods. In the context of MIR, *Onset detection* is concerned with locating the onsets of an accompaniment - percussion, wind or string instruments. This thesis introduces onset detection to Indian classical percussion instruments. In the process, a novel algorithm employing signal demodulation is introduced, and improvisations that compete with and complement existing techniques are suggested.

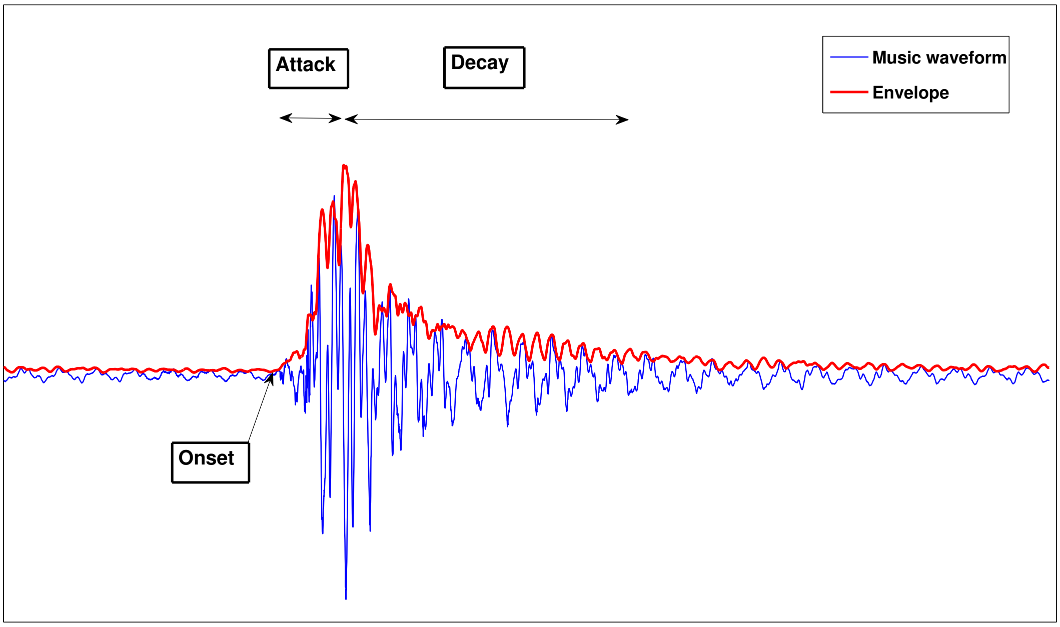


Figure 1.2: Components of an isolated stroke - onset, attack and decay. Illustration adapted from [3]

1.4 Overview of Thesis

In Chapter 2, two attempts to study the high resolution property of group delay functions are presented. It is shown that both works consider only the region closer to resonance, and conclusions are made directly from observations. The need for a formal proof is motivated. This is followed by a brief review of existing relevant approaches to onset detection task - magnitude based and phase based. In the backdrop of these approaches, the proposed algorithm is shown to use information from both magnitude and phase components.

The high resolution property for a single-resonator (pole/zero) minimum-phase system is proved using the concept of ndB bandwidth in Chapter 3. The group delay function is shown to have a lesser bandwidth than the magnitude spectrum for all pole configurations. The peakedness is validated using numerical measures like kurtosis and spectral flatness. Next, the peakedness property is argued to hold true for multi-pole systems too, and verified using numerical 3dB computation and acceleration measures.

Chapter 4 explains the algorithm, beginning with the characteristics of Carnatic percussion waveforms that motivated an amplitude-frequency modulation (AM-FM) approach. The database and evaluation measures are discussed, and results are shown where the proposed algorithm matches closely in performance with a state of the art algorithm. Two improvisations to the proposed algorithm are then proposed. Conclusion, criticisms and directions for future work are presented in Chapter 5.

1.5 Contribution

The main contributions of this thesis are as follows:

- A theoretical proof is presented, wherein the group delay response for a single-resonator minimum-phase system is shown to possess more peakedness in comparison to the magnitude spectrum.
- Three numerical measures are chosen to validate the proof. The analysis is extended to multi-pole systems, and shown to be true for all possible pole configurations.
- A novel onset detection algorithm is presented, that makes use of the high resolution of group delay functions. The algorithm is shown to perform comparably with state of the art algorithms.

- The dependence of the proposed algorithm on a windowing parameter that controls it's smoothness is removed by using the entire range of parameter values and producing a multi-resolution group delay function.

- A fairly large annotated dataset consisting of five Carnatic percussion instruments was developed. The audio comes along with label information as well as onset times.

CHAPTER 2

Previous Work

In this chapter, a brief review of previous work pertaining to use and study of the high resolution property of group delay functions is presented. The motivation for current work is developed in this context. Later, the major constituents of an onset detection algorithm - extraction of detection function and peak picking are explained. A survey of existing onset detection methods is then presented.

2.1 On the resolution of group delay functions

Equations (1.1-1.3) can be extended to a cascade of N resonators. The overall phase response of the cascade is a summation of the individual phase responses. This property is commonly referred to as the *additive property of phase*. In the context of an all-pole model of speech production, the individual resonators can be thought of as the formants while the output from the cascade as the speech signal. Additive property is a relatively well understood phenomenon. It served as a basis for proposing a non parametric method for formant estimation in [12]. In [52], this property was employed to extract spectral information even in the presence of noise. In [21], the ability of group delay to resolve closely spaced peaks was studied in detail and proposed to assist in automatic segmentation of speech into 'syllable-like' units. It was shown that nearby peaks are well discriminated in the group delay domain even when all peaks are not present in the magnitude spectrum.

The high resolution property although stated in most of earlier applications involving group delay and derived functions, has not been well explained or proven.

In earlier efforts to justify this property, the group delay function was studied in the vicinity of resonances. In [51], where formant estimation was attempted from the linear prediction phase spectra, a cascade of resonators was considered. For a constrained location of the poles, it was shown that squared magnitude behaviour of the group delay function around the resonance leads to its high resolution property. The magnitude spectrum for a cascade of N poles $\alpha_i \pm \beta_i$ is given as :

$$|H(\omega)|^2 = \prod_{i=1}^{N} \frac{1}{(\alpha_i^2 + \beta_i^2 - \omega^2)^2 + 4\omega^2\alpha_i^2)} \tag{2.1}$$

The corresponding group delay spectrum is :

$$\theta'(\omega) = \sum_{i=1}^{N} \frac{2\alpha_i(\alpha_i^2 + \beta_i^2 - \omega^2)}{(\alpha_i^2 + \beta_i^2 - \omega^2)^2 + 4\omega^2\alpha_i^2)} \tag{2.2}$$

For $\beta_i^2 >> \alpha_i^2$, the group delay can be approximated as

$$\theta'(\omega) = \sum_{i=1}^{N} \frac{K_i}{(\alpha_i^2 + \beta_i^2 - \omega^2)^2 + 4\omega^2\alpha_i^2)}$$

or

$$\theta'(\omega) = \sum_{i=1}^{N} K_i |H(\omega)|^2 \tag{2.3}$$

Hence most of the group delay strength was argued to lie around the resonance, and closely spaced formants could be better differentiated. Clearly, the proof places a significant constraint on the pole, that it must be close to the imaginary axis and have a small bandwidth. Later [4] considered a parallel connection of resonators while working on pitch histograms that resembled a non-constant Q factor for each of the peaks. An example of two resonators in parallel was considered, and once again, the group delay response was approximated as squared version of the magnitude spectrum around the peaks.

The above observations attempt to explain the high resolution property in the vicinity of the peaks. The choice of analysis (cascade or parallel resonators) is adapted based on the application considered. In this work, we take the case of a generic, single-resonator system. We prove that the peakedness of the group delay response is always greater than that in magnitude spectrum, irrespective of the pole location in the z-plane. Specifically, the strength of the group delay function at the ndB bandwidth of the magnitude spectrum is always less than the magnitude spectrum. Since both group delay and magnitude spectrum are found to possess a common maxima value at resonance, the result implies that at any point in the spectrum, the peakedness of group delay function is greater. Detailed analysis is present in the following chapter.

2.2 Approach to an Onset Detection algorithm

A number of techniques have been proposed to detect onsets. They can be classified based on the class of detection functions used - temporal, spectral, probabilistic, machine learning based, etc. Alternatively, they may be classified into online (real-time) algorithms, or offline algorithms. However, all of them consist of two major steps :

- Extraction of **detection function** : The raw waveform is transformed/reduced to a function that emphasises the onsets. Computation of the detection function (or *novelty function*) may involve noise removal, pitch removal or any other processing depending on the nature of the data.
- **Thresholding** : Onsets are located on the detection function by means of a threshold. This threshold may involve peak/valley picking and zero crossing detection. Generally, a global threshold is observed to be difficult to optimise, hence most algorithms tend to employ adaptive thresholds.

The performance of the algorithm depends largely on the choice of the detection function, which facilitates the thresholding step. In this section, we discuss the major approaches in literature, focussing specifically on magnitude and phase based algorithms. Each of the approaches emphasise information from different components of the Fourier transform, and the proposed algorithm attempts to use both components in the computation of the detection function.

2.3 Major Approaches

Magnitude based

Energy based methods look at significant rise in signal energy in the vicinity of an onset. This is achieved by dividing the waveform into smaller units (short-term processing) and computing the frame-wise energy, or by extracting the amplitude envelope. Early methods included the derivative of energy, which tend to result in sharper peaks at the onsets when compared to energy. Later approaches looked at filtering specific bands in the spectrum ([18]), and separating the transients from steady states ([27]). Energy based methods perform well for sharp onsets, especially for percussion instruments, but tend to suffer in the presence of noise and for instruments with softer onsets (string and wind types).

Since spectral methods analyse the Fourier spectrum, and are based on the premise that regions of transience manifest as broadband changes in the spectrum, different weighing functions that act on the spectrum have been proposed. The most widely used function is the HFC (High Frequency Contact) that uses linear weights. A generic form for the detection function $D[n]$ would be :

$$D[n] = \sum_k W[k]|X[n,k]|^2 \tag{2.4}$$

where n and k are the temporal and spectral indices respectively, and $W[k]$ is the weighing function. Analogous to the energy *difference*, the spectral difference or *spectral flux* is defined to be the geometric distance between successive frames. A number of forms exist, based on the choice of the distance. In [15] for example, the following definition is used :

$$SF[n] = \sum_k H(|X[n+1,k]| - |X[n,k]|)^2 \tag{2.5}$$

where $H(x)$ is the unit step function. This is added to include only for the bins with positive energy change. Another common implementation by [31] uses the L_1 norm, as opposed to the L_2 norm.

Phase based

Information about the temporal structure of the signal is also encoded in the phase spectrum. Phase based methods look at approximating the instantaneous frequency (IF) at a bin by change in phase spectrum between successive frames.

$$f[k,n] \propto \phi[k,n] - \phi[k,n-1] \tag{2.6}$$

During steady state regions, the IF is expected to be fairly constant and a significant change is expected at the onsets. Hence a suitable choice for the detection function would be a derivative of the IF, or the second-order difference of phase.

$$\delta\phi[k,n] = \phi[k,n+1] - 2\phi[k,n] + \phi[k,n-1] \tag{2.7}$$

While phase based approaches can detect softer onsets when compared to the energy based detectors, they are susceptible to phase distortions from noisy components.

Complex Domain

In [13], a new method for onset detection using both the magnitude spectrum and phase spectrum was proposed. The detection function was formed from the error between target STFT (Short Time Fourier Transform) and predicted STFT bins which are complex valued.

Let $S_k(m)$ denote the target (observed) value for the kth STFT bin at frame m.

$$S_k(m) = R_k(m)e^{\phi_k(m)} \tag{2.8}$$

The predicted value is constructed as follows : The magnitude is taken as that of the previous frame $|S_k(m-1)|$ and the phase is the sum of previous phase and the phase difference with it's previous frame $(\phi_k(m-1))$ + $(\phi_k(m-1) - \phi_k(m-2))$. The reason being the prediction captures deviations in magnitude as well as the phase spectrum. Hence

$$\tilde{S_k}\ (m) = |S_k(m-1)|e^{princarg(2\phi_k(m-1)-\phi_k(m-2))} \tag{2.9}$$

The error for the kth bin at mth frame is computed as the geometrical distance between $S_k(m)$ and $\tilde{S}_k(m)$. The detection function is this error value summed over the bins in a particular frame.

$$D(m) = \sum_k |S_k(m) - \tilde{S}_k(m)| \tag{2.10}$$

The above detection function was observed to be sharper than those derived from magnitude only and phase only based approaches. Similar to complex domain, we propose an approach that also takes into account magnitude and phase deviations albeit in a slightly different manner.

Other approaches

All the above methods were common in the sense that they worked upon the time-frequency aspects of the audio signal. Probabilistic methods for onset detection consider the waveform as observations from a single/multiple model(s), and either look for instants where the model is switched between the two, or a surprise instant with respect to a single global model. Recently, linear prediction analysis has been used in onset detection, with the error function as a possible detection function.

$$E[n] = X[n] - \sum_{k=1}^{p} X[n-k] \tag{2.11}$$

where p is the order for the LP analysis. Onsets are expected to give a large error function since they denote a significant change in signal evolution with respect to the previous samples. The state of the art algorithms that have been recently proposed involve training a (deep) neural network - recurrent and convolutional. Typically, the magnitude spectrogram or filter bank energies of onset frames alongwith context are used for training. However, these algorithms not only require large amounts of data for training but are also computationally intensive. [3], [10] and [5] discuss the major approaches in onset detection over the years and evaluate some of them.

An illustration of some of the common algorithms on a relatively clean (bereft of noise) mridangam recording is presented in Fig. 2.1.

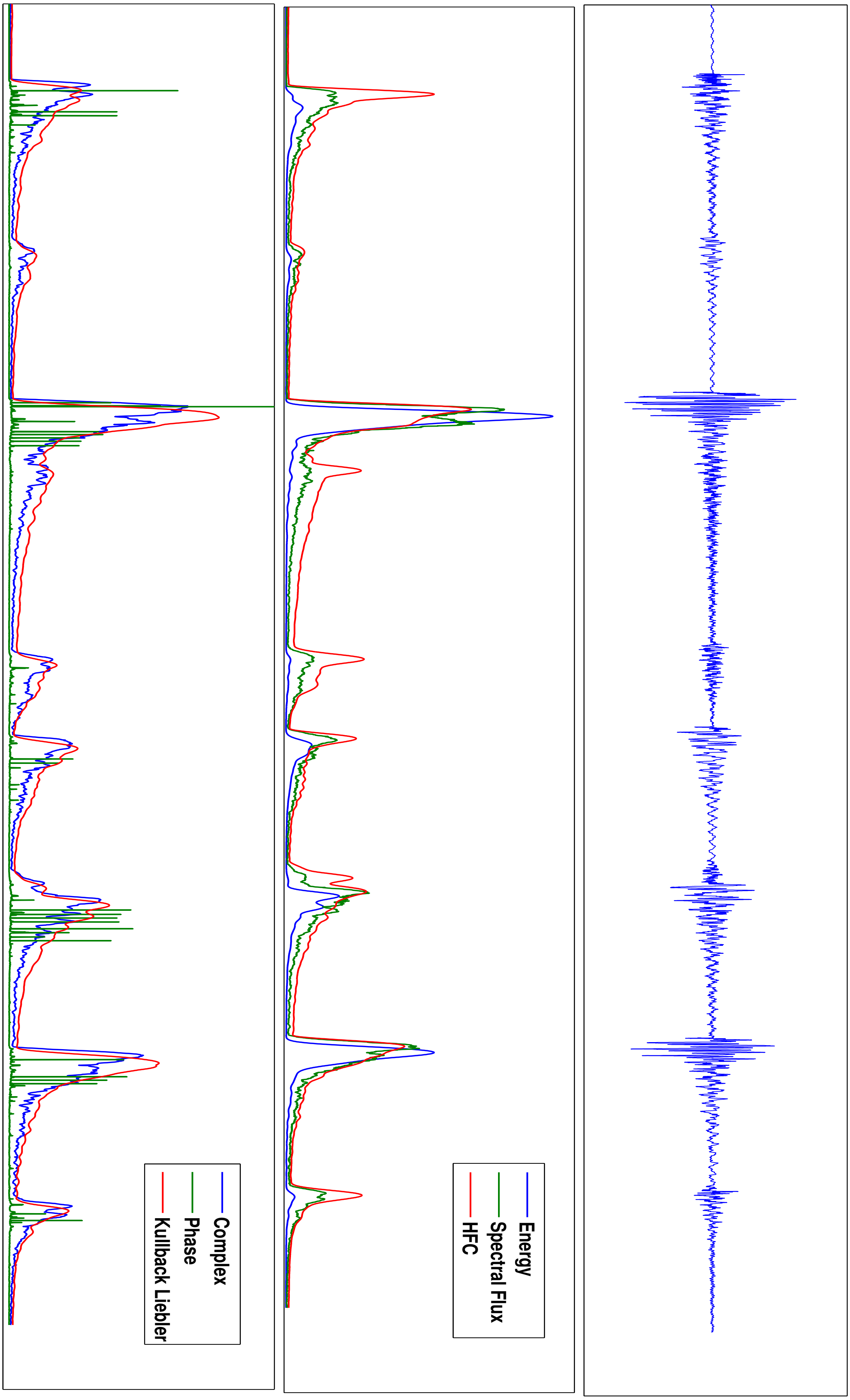
The mridangam waveform is chosen so as to contain onsets with significantly varying amplitudes. In the second subplot, energy based detection function is observed to have the largest dynamic range, followed by spectral flux. A large dynamic range necessitates fine tuning of the peak picking parameters (global and local). Linear weighing of the frequency scale results in a much more robust detection function in HFC. In the final subplot, phase based detection function is observed to be extremely noisy, although it possesses a very high precision[1]. Complex domain and Kullback Leibler approaches result in almost similar detection functions, although a faster decay is seen for the complex domain method.

2.4 Summary

This chapter discusses previous investigations of the high resolution property of group delay functions, and it was noted that both approaches were made only in the vicinity of poles, and did not prove this property. A survey of existing onset detection algorithms is then presented. Both theoretically and experimentally, the complex domain method for onset detection is found to be a superior signal processing algorithm, suggesting

[1] An evaluation measure defined in Chapter 4

Figure 2.1: Detection functions of some common onset detection algorithms evaluated on a mridangam *tani avarthanam* clip



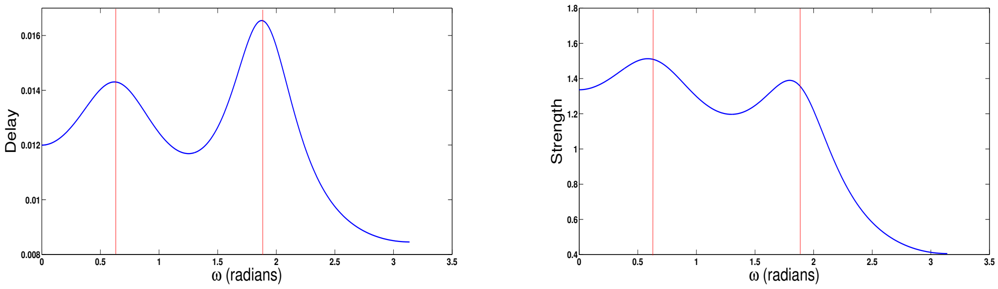
that a combination of magnitude and phase is a good choice for approaching an onset detection problem.

CHAPTER 3

Group Delay Analysis

In this chapter, we provide a theoretical basis to compare the peakedness of group delay functions and magnitude spectrum for a single-resonator minimum-phase system. This analysis of the group delay function gives an insight into the reason for the rapid decay around a resonance that leads to its high-resolution property. This is corroborated through the estimation of kurtosis, spectral flatness, and acceleration measures; and compared for various configurations of single and multi-pole minimum-phase systems for both the group delay spectrum and the magnitude spectrum.

First, an elementary single-resonator minimum-phase system is considered. The choice of this system is to allow the possibility of a closed form expression, comparing the strengths of the magnitude spectrum and group delay response at the ndB bandwidth of the magnitude spectrum. The ndB bandwidth for an isolated peak is an indication of the band of frequencies that are associated with a damped sinusoid. If the bandwidth is large, and the gain-bandwidth product constant, the location of the peak cannot be determined accurately from the magnitude spectrum. Specifically, it is observed that at the 3dB bandwidth corresponding to the magnitude spectrum, the strength of the group delay function significantly decreases away from the resonance, falling off to much less than $\frac{1}{\sqrt{2}}$ of the peak strength.



(a) (b)

Figure 3.1: **(left)** Group delay and **(right)** (Magnitude spectrum) spectra for a minimum-phase system with poles at $0.7e^{j0.2\pi}$ and $0.6e^{j0.6\pi}$. Large bandwidth in the case of magnitude spectrum results in the peaks moved from their true locations

3.1 A single-resonator minimum-phase system - Theoretical Approach

Consider a causal, discrete-time signal $x[n]$ with one pole whose location in the z-plane is given as $z_0 = re^{j\omega_0}$, or $z_0 = e^{-\sigma_0 + j\omega_0}$. σ_0 represents the bandwidth of the pole and ω_0 the angle with respect to the abscissa. The Z-transform of the above system is:

$$X(z) = \frac{1}{(z - z_0)(z - z_0^*)} \tag{3.1}$$

When evaluated at the unit circle:

$$X(\omega) = \frac{1}{(e^{j\omega} - e^{-\sigma_0 + j\omega_0})(e^{j\omega} - e^{-\sigma_0 - j\omega_0})} \tag{3.2}$$

The expression for the magnitude spectrum is given as:

$$|X(\omega)| = P \times Q \tag{3.3}$$

where

$$P = \frac{1}{\sqrt{1 + e^{-2\sigma_0} - 2e^{-\sigma_0} cos(\omega - \omega_0)}} \tag{3.4}$$

$$Q = \frac{1}{\sqrt{1 + e^{-2\sigma_0} - 2e^{-\sigma_0} cos(\omega + \omega_0)}} \tag{3.5}$$

Considering (3.4) alone, the maximum value of $\frac{1}{1-e^{-\sigma_0}}$ occurs at $\omega = \omega_0$. To compute the ndB bandwidth, we determine the angular frequency (ω_1) at which the magnitude spectrum falls to $\frac{1}{N}$ of its maximum value, i.e

$$\frac{1}{\sqrt{(1 + e^{-2\sigma_0} - 2e^{-\sigma_0} cos(\omega_1 - \omega_0))}} = \frac{1}{N(1 - e^{-\sigma_0})} \tag{3.6}$$

Here, $N = 10^{\frac{n}{20}}$. Solving for ω_1,

$$\omega_1 = \omega_0 \pm \cos^{-1}(N^2 + \frac{1 - N^2}{2}(e^{\sigma_0} + e^{-\sigma_0})) \tag{3.7}$$

The ndB bandwidth is the interval with ω_0 at the center, and is given by

$$\omega_{ndB} = 2\cos^{-1}(N^2 + \frac{1 - N^2}{2}(e^{\sigma_0} + e^{-\sigma_0})) \tag{3.8}$$

We repeat this analysis for the group delay spectrum. The phase spectrum for the system defined by (3.2) is given by

$$\theta(\omega) = -\tan^{-1}\left(\frac{\sin(\omega) - e^{-\sigma_0}\sin(\omega_0)}{\cos(\omega) - e^{-\sigma_0}\cos(\omega_0)}\right) - \tan^{-1}\left(\frac{\sin(\omega) + e^{-\sigma_0}\sin(\omega_0)}{\cos(\omega) - e^{-\sigma_0}\cos(\omega_0)}\right) \quad (3.9)$$

The group delay is defined as the negative derivative of the phase spectrum and is given by

$$GD(\omega) = \frac{1 - e^{-\sigma_0}cos(\omega - \omega_0)}{1 + e^{-2\sigma_0} - 2e^{-\sigma_0}cos(\omega - \omega_0)} + \frac{1 - e^{-\sigma_0}cos(\omega + \omega_0)}{1 + e^{-2\sigma_0} - 2e^{-\sigma_0}cos(\omega + \omega_0)} \quad (3.10)$$

Differentiating the first term in (3.10) and equating to zero, we find that it displays the same abscissa and ordinate for the maxima as the magnitude spectrum. Solving for the ndB frequency,

$$\frac{1 - e^{-\sigma_0}\cos(\omega_1 - \omega_0)}{1 + e^{-2\sigma_0} - 2e^{-\sigma_0}cos(\omega_1 - \omega_0)} = \frac{1}{N(1 - e^{-\sigma_0})} \quad (3.11)$$

$$\omega_1 = \omega_0 \pm \cos^{-1}\left(\frac{(1-N) + Ne^{-\sigma_0} + e^{-2\sigma_0}}{Ne^{-2\sigma_0} + e^{-\sigma_0}(2-N)}\text{‘}\right) \quad (3.12)$$

Hence, the ndB bandwidth is given as

$$\omega_{ndB} = 2\cos^{-1}\left(\frac{(1-N) + Ne^{-\sigma_0} + e^{-2\sigma_0}}{Ne^{-2\sigma_0} + e^{-\sigma_0}(2-N)}\text{‘}\right) \quad (3.13)$$

Since ndB bandwidth need not exist for all possible pole locations, we discuss its existence for the case of group delay and magnitude spectrum separately. The arguments of $\cos^{-1}$ function in (3.8) and (3.13) are constrained to lie within $[-1, 1]$:

Magnitude spectrum:

$$-1 \leq N^2 + \frac{1-N^2}{2}(e^{\sigma_0} + e^{-\sigma_0}) \leq 1 \quad (3.14)$$

e^{σ_0} being positive, the expression is always lesser than 1.

$$N^2 + \frac{1-N^2}{2}(e^{\sigma_0} + e^{-\sigma_0}) \geq -1 \quad (3.15)$$

Solving the quadratic equation in $e^{-\sigma_0}$,

$$e^{-\sigma_0} \in [\frac{N-1}{N+1}, \frac{N+1}{N-1}] \quad (3.16)$$

σ_0 being positive, the effective range of $e^{-\sigma}$ is reduced to $[\frac{N-1}{N+1}, 1]$

Group delay spectrum:

$$-1 \leq \left(\frac{(1-N) + Ne^{-\sigma_0} + e^{-2\sigma_0}}{Ne^{-2\sigma_0} + e^{-\sigma_0}(2-N)} ` \right) \leq 1 \tag{3.17}$$

The inequality gives rise to two quadratic equations:

$$(N+1)e^{-2\sigma_0} + 2e^{-\sigma_0} + (1-N) \geq 0 \tag{3.18}$$

$$(1-N)e^{-2\sigma_0} + (2N-2)e^{-\sigma_0} + (1-N) \leq 0 \tag{3.19}$$

The common range of $e^{-\sigma_0}$ in both of them being:

$$e^{-\sigma_0} \in [\frac{N-1}{N+1}, \infty] \tag{3.20}$$

Equations (3.16) and (3.20) result in $\sigma_0 \in [\frac{N-1}{N+1}, 1]$ as the interval for consideration of ndB bandwidth.

We now consider the value of the group delay function at the ndB bandwdith of the magnitude spectrum. Substituting for (3.7) in the first term of (3.10),

$$\tau(\omega) = \left(\frac{(1+N^2) + e^{-2\sigma_0}(N^2-1) - 2N^2e^{-\sigma}}{2[N^2(1+e^{-2\sigma_0}) - 2N^2e^{-\sigma_0}]} \right) \tag{3.21}$$

The magnitude spectrum at the same frequency was shown to have a value of $\frac{1}{N(1-e^{-\sigma_0})}$ in (3.8). The difference between this value and that of the group delay spectrum in (3.21) is given by:

$$\frac{1}{N(1-e^{-\sigma_0})} - \frac{(1+N^2) + e^{-2\sigma_0}(N^2-1) - 2N^2e^{-\sigma}}{2[N^2(1+e^{-2\sigma_0}) - 2N^2e^{-\sigma_0}]} \tag{3.22}$$

As the denominator is positive, the numerator is a quadratic expression positive in the interval $e^{-\sigma_0} \in [\frac{N-1}{N+1}, 1]$, which is the same for the existence of ndB bandwidth. Thus, it is proved that group delay is smaller in strength and greater in peakedness than the magnitude spectrum at all points in the frequency spectrum. Figure 3.2 shows the phenomenon explained in this section for a particular σ_0 in (3.22).

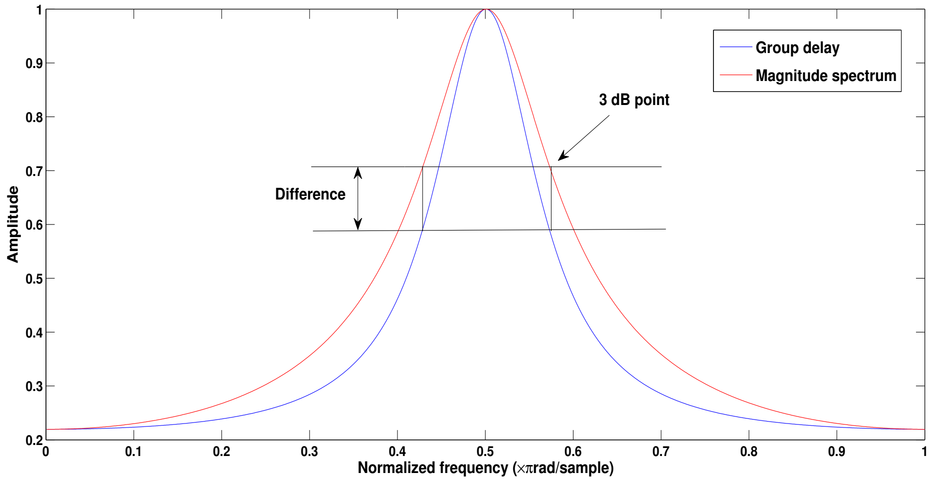


Figure 3.2: Illustrating the ndB analysis using 3dB as an example. Faster decay of group delay in comparison to magnitude spectrum is shown by the difference at half power of magnitude spectrum

Before validating the above statement in the following section, a few points must be noted:

- The above analysis has been carried out on the positive half of the spectrum. The effect due to the conjugate term in (3.2) has not been considered, although the additive property of group delay only increases it's spectral resolution.
- The same analysis can be extended to a single-zero system, in which case the magnitude spectrum is inverted, and the group delay function is negated in sign. Identical expressions for ndB bandwidths (3.8) and (3.13) can be derived. Except for group delay, the value will be less negative at the ndB point compared to that of the magnitude spectrum.
- The system is assumed to be minimum-phase. Although real systems are not necessarily minimum-phase, it is possible to derive a minimum-phase system from a mixed-phase system [28].
- Two different window sizes have been used for experiments - 512 samples, considering a typical speech frame $25ms$ long sampled at $16KHz$/$44.1KHz$ yields 400/1103 samples, respectively. Another window of 128 samples has also been used to illustrate the property for very short utterances too.

3.2 Numerical Analysis

For the above system, kurtosis and spectral flatness are chosen to extend the peakedness comparison. The analyses take the entire frequency spectrum into account, unlike the previous works. Kurtosis is defined as a scaled version of the fourth standardised moment about the mean of a distribution.

$$k = \frac{E[X-\mu]^4}{\sigma^4} \tag{3.23}$$

Since we do not model the group delay and magnitude spectrum using a pre-defined distribution, we replace the kurtosis measure with sample kurtosis measure defined as follows:

$$k \approx \frac{\Sigma(X_i - \bar{X})^4)/n}{(\Sigma(X_i - \bar{X})^2/n)^2} \tag{3.24}$$

A distribution with a higher kurtosis value has been argued to represent both higher peakedness and heavy tailedness ([9]). We employ this property to show the peakedness of group delay spectrum for a single pole system. For multi-resonator systems the responses resemble a multi-modal distribution, hence a single kurtosis value cannot quantify the peakedness of individual peaks.

Spectral flatness is defined as the ratio of the geometric mean to the arithmetic mean of a power spectrum ([23]). It has been used to characterize how noise-like (or tone-like) a waveform is ([24, 11]). An Additive White Gaussian Noise (AWGN) signal has a completely flat spectrum and has the maximum possible spectral flatness value of 1. The more the peakedness, the less the spectral flatness.

The results for kurtosis and spectral flatness by varying the bandwidth (σ_0) and the angular frequency (ω_0) are summarized in Figure 3.3. The behaviour of kurtosis and spectral flatness measures are observed to be similar irrespective of the pole positions. Both magnitude spectra and group delay converge asymptotically to a spectral flatness of 1 as the bandwidth increases (the pole moving closer to origin), as expected. The difference in peakedness is emphasized for pole locations with low bandwidths.

In the case of multi-pole system, the resonant frequency and bandwidth of individual poles determine the effect on each other in the spectrum and hence individual analysis is more relevant than a global measure. Further, in an ndB analysis, the system can be analysed as multiple single-resonator systems as long as the difference in resonant frequencies is atleast twice the ndB bandwidth. Owing to the group delay decaying exponentially away from the resonance, this statement is observed to be true in experiments. In this work, we compute numerically the 3dB bandwidth over all possible configurations ($\sigma \in (0.05, 0.8)$, $\omega \in (0.3\pi, 0.8\pi)$) of the two resonances. Further,

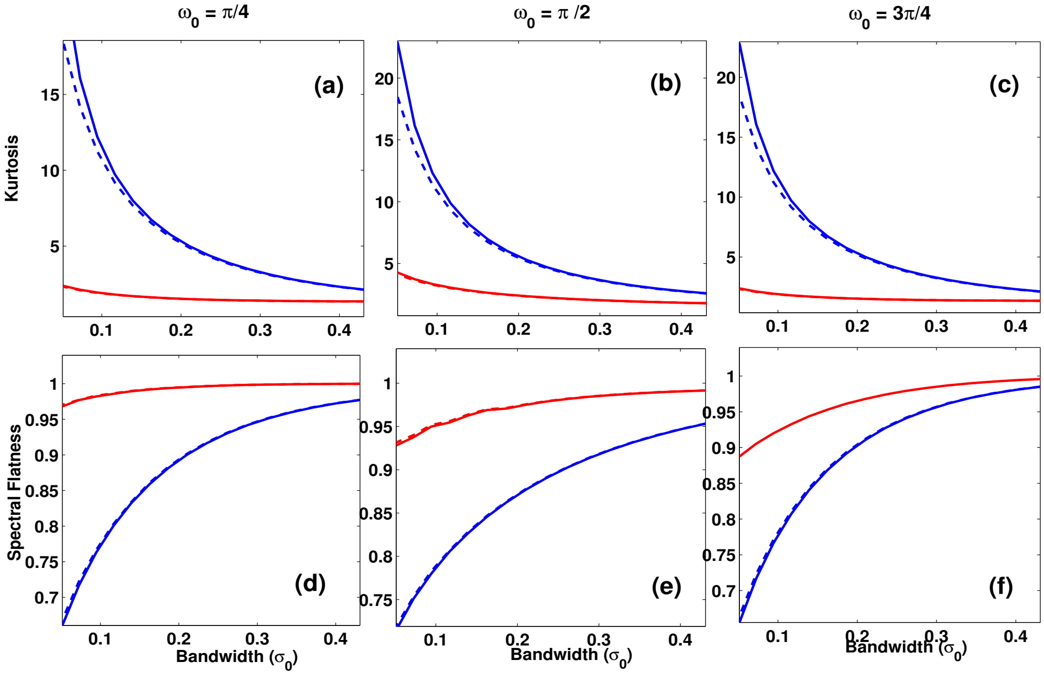


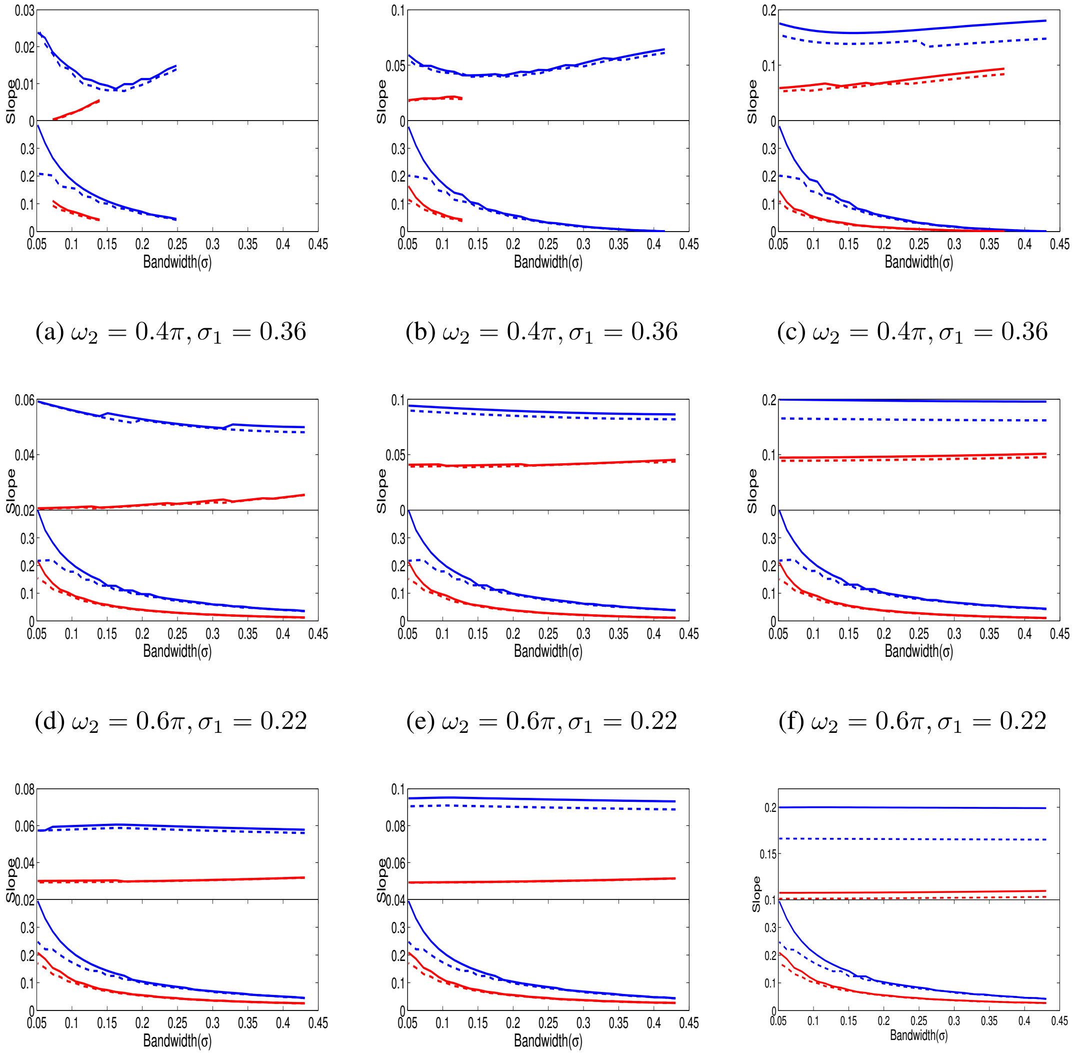
Figure 3.3: Demonstrating the peakedness of group delay functions (blue) over log-magnitude spectrum (red). Kurtosis measures over a bandwidth range of [0.1,0.4] are shown in (a),(b),(c). Spectral flatness measures for the same bandwidth range are shown in (d),(e),(f). Dotted lines correspond to the windowed response.

we use the average acceleration measures of group delay and magnitude spectrum (as a function of frequency) at the vicinity of the poles to directly compute the rate of rise and fall around the peaks. The pole locations are varied between the same ranges as those in the ndB bandwidth computations. Both numerical 3dB and acceleration measures are reported in Table 3.1. The peakedness property is clearly preserved for both measures.

Table 3.1: Evaluations illustrating lower 3dB bandwidths and higher acceleration measures of Group delay (GD) over Magnitude Spectrum (MS) for a double pole system

Measure	GD		MS	
	pole1	pole2	pole1	pole2
3DB (bins)	135.29	106.57	237.79	184.29
Acc.($\times 10^{-2}$)	2.28	2.26	1.13	1.25

The results are also reported by selecting a few representative configurations for the two poles and varying the bandwidth of the second pole in Figure 3.4. Pole 1 is fixed at an angular frequency of 0.3π (ω_1) in all cases. Bandwidth of Pole 1 (σ_1) and angular frequency of Pole 2 (ω_2) are varied across the various examples, and the bandwidth of Pole 2 (σ_2) is varied within each case. The magnitudes of normalized slopes at each of the two peaks for log-magnitude and group delay spectra are shown. For configurations where the two peaks are not distinguishable in the magnitude spectrum, the results could not be reported for the entire range of σ_2 (Fig 3.4a, 3.4b, 3.4c). It is observed that windowing operation reduces the acceleration measures for both magnitude spectrum and group delay. This is because the $sinc$ functions of a window increase the bandwidth of both functions upon convolution. Yet, group delay continues to possess higher acceleration around the peaks, implying faster decay than the magnitude spectrum.



(a) $\omega_2 = 0.4\pi, \sigma_1 = 0.36$ (b) $\omega_2 = 0.4\pi, \sigma_1 = 0.36$ (c) $\omega_2 = 0.4\pi, \sigma_1 = 0.36$

(d) $\omega_2 = 0.6\pi, \sigma_1 = 0.22$ (e) $\omega_2 = 0.6\pi, \sigma_1 = 0.22$ (f) $\omega_2 = 0.6\pi, \sigma_1 = 0.22$

(g) $\omega_2 = 0.8\pi, \sigma_1 = 0.11$ (h) $\omega_2 = 0.8\pi, \sigma_1 = 0.11$ (i) $\omega_2 = 0.8\pi, \sigma_1 = 0.11$

Figure 3.4: Comparison of acceleration magnitudes for group delay (blue) and log-magnitude spectrum (red). Top half of each figure corresponds to Pole 1 while bottom half corresponds to Pole 2. ω_1 of Pole 1 is kept constant at 0.3π for the entire experiment. σ_2 is varied from 0.44 to 0.05 in each plot. Dotted lines correspond to the windowed response.

3.3 Summary

In this chapter, examples of single-resonator and multi-resonator systems are considered and the resolution of group delay and magnitude spectrum are studied. The high resolution property that lends the superior performance to group delay functions is proved for the above systems and argued to be extended for any generic pole-zero system.

Expressions for the ndB bandwidth are derived in the case of single-resonator system, and it is proved that the group delay has lower strength at ndB bandwidth. Next, the peakedness is quantified using numerical measures. The results are consistent for various pole configurations and with the application of window of small duration.

CHAPTER 4

Proposed Algorithm

In this chapter, a novel algorithm for onset detection on percussion instruments that employs the high spectral resolution of group delay functions is presented. The approach is based purely on signal processing, and is agnostic to the data. The motivation for the algorithm is discussed using syllable segmentation, followed by a brief review of AM-FM demodulation and minimum-phase group delay processing. The significance of the algorithm is explained in detail using a test clip. The dataset used for evaluation is presented and the performance metrics are explained. It is observed that the performance almost matches that of the state of the art.

Syllable Segmentation

In this thesis, onset detection for MIR has been inspired by syllable segmentation for ASR and TTS applications. The constituents of an isolated stroke - onset, attack and decay were found to be similar to that of a syllable - onset, nucleus and cuda (Figure 4.1).

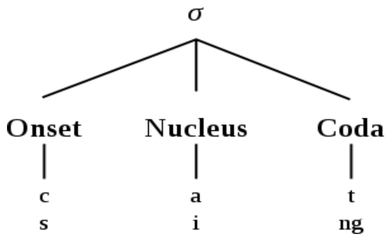


Figure 4.1: Break-up of a syllable

The syllable is considered as the fundamental unit of speech production and is found to be a better unit than the phoneme for ASR systems due to it's relatively better representational and durational stability [50]. Segmentation at the syllable level followed by isolated style recognition was found to improve performance when compared to a flat-start based recognition [26].

The syllable structure can be defined as C*VC* (C - Consonant, V - Vowel). The vowel at the centre is sometimes referred to as the *nucleus* between an *onset* and *coda*. The energy of a syllable is lower in the consonant and higher in the vowel parts. Hence, a syllable segmentation task would look for regions of low energy, while in the method proposed in this work, regions of high energy are chosen as onset (The attack region in percussion instruments can be considered smaller than tolerance limit or negligible).

The music signal is treated to be an amplitude-frequency modulated (AMFM) waveform. The amplitude of the difference signal is estimated in order to emphasise both amplitude and frequency components. The envelope of this signal is extracted which shows peaks at regions of high energy. Finally, minimum-phase group delay processing is applied on the estimated envelope to give onsets of high temporal resolution.

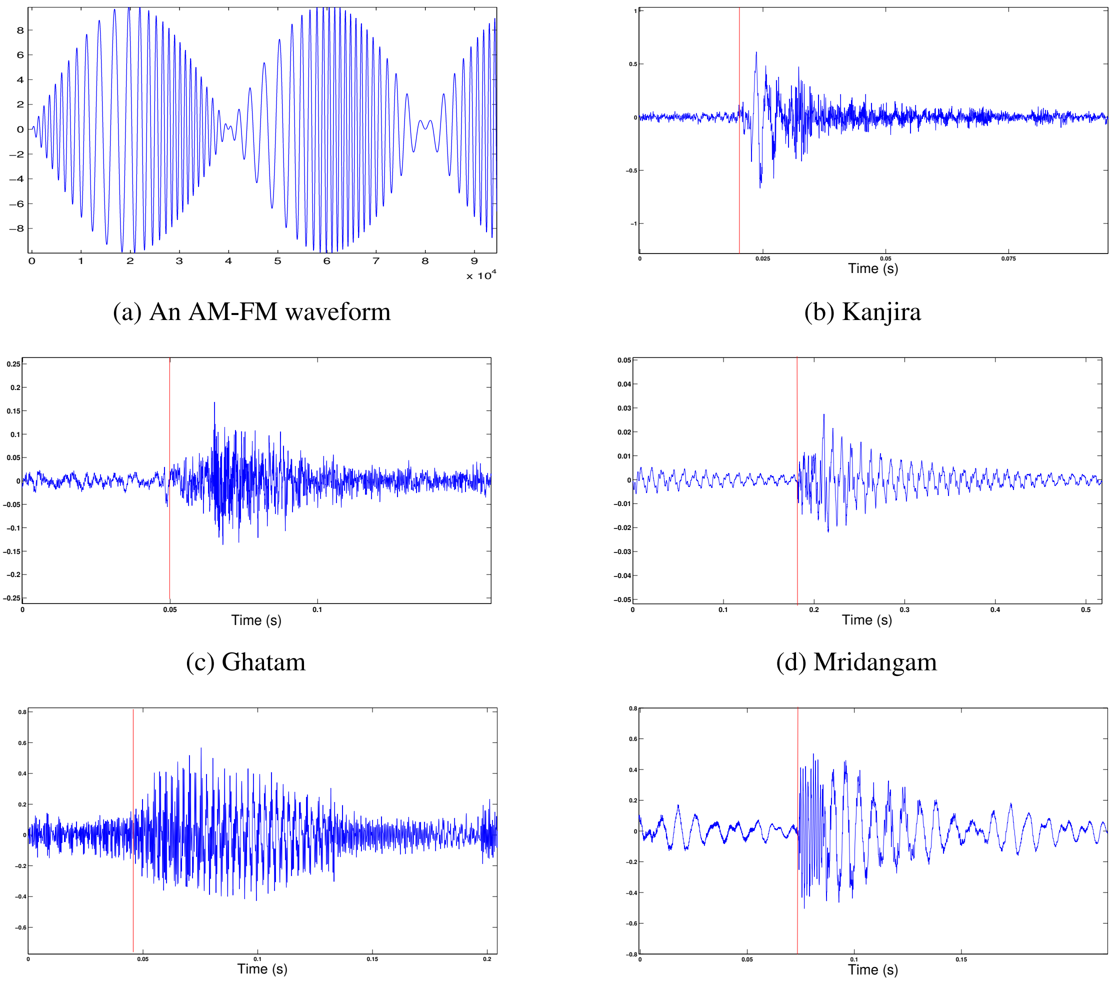
4.1 Resemblance to AM-FM

In the context of communications, a message signal $m(t)$ is encoded with a high frequency carrier signal $s(t)$ before transmission. Modulation is performed to reduce transmitter and receiver antennae sizes, and to multiplex different message signals. Various modulation schemes exist, and are characterized by the influence of $m(t)$ on $s(t)$. In the case of amplitude-frequency modulation (AM-FM), both the amplitude and frequency of the carrier signal are influenced by two message signals $m_1(t)$ and $m_2(t)$. The basic representation of an AM-FM signal is :

$$x(t) = m_1(t)cos(\omega_c t + k_f \int m_2(t)dt) \tag{4.1}$$

where $m_1(t)$, $m_2(t)$ represent the message signals, k_f is the frequency modulation constant and ω_c is the carrier frequency. Figure 4.2(a) presents an example using sinusoids in place of $m_1(t)$ and $m_2(t)$.

It is observed that most percussive strokes in Carnatic music can be modelled by an AM-FM signal, based on the variations in amplitude and frequency in the vicinity of an onset. Figure 4.2 illustrates this resemblance by comparing a carrier signal modulated using sinusoidal message signals with individual strokes of mridangam, ghatam, kanjira, morsing and thavil. In all the instruments, the onset is characterised by changes in both amplitude and frequency. We propose that the messages $m_1(t)$ and $m_2(t)$ contain information necessary to pinpoint the location of onsets. A demodulation technique is necessary to extract this information before proceeding to locate the onsets.



(a) An AM-FM waveform

(b) Kanjira

(c) Ghatam

(d) Mridangam

(e) Morsing

(f) Thavil

Figure 4.2: Resemblance of Carnatic percussion strokes (*with ground truth marked*) to an Amplitude-Frequency modulated waveform

4.2 Demodulation and Envelope Detection

Differentiating the AM-FM signal in (4.1) with respect to time :

$$x'(t) \approx -e(t)sin(\omega_c t + k_f \int m_2(t)dt) \tag{4.2}$$

where

$$e(t) = m_1(t)(\omega_c + k_f m_2(t)) \tag{4.3}$$

The term $m_1'(t)cos(\omega_c t + k_f \int m_2(t)dt)$ has been ignored in (4.2) since ω_c can be assumed large. Both the message signals are now part of the amplitude in (4.2). We postulate that all the information about an onset is now contained within the envelope function $e(t)$. $e(t)$ is extracted from $x'(t)$ using the Hilbert transform :

Any real-valued signal $S(t)$ with Fourier transform $S(w)$ can be represented by its analytic version (as introduced in [17]) and is given by

$$S_a(t) = 2\int_0^\infty S(w)e^{j2\pi wt}dw \tag{4.4}$$

From (4.4), it is the inverse Fourier transform of the positive frequency part alone. In terms of input signal $S(t)$,

$$S_a(t) = S(t) + iS_H(t) \tag{4.5}$$

where, $S_H(t)$ is the Hilbert Transform of $S(t)$. The real part of this analytic signal represents the actual signal and the imaginary part it's Hilbert Transform. The magnitude of the analytic signal gives an estimate of the envelope. An example of the envelope estimated by the above method is plotted in Fig 4.3.

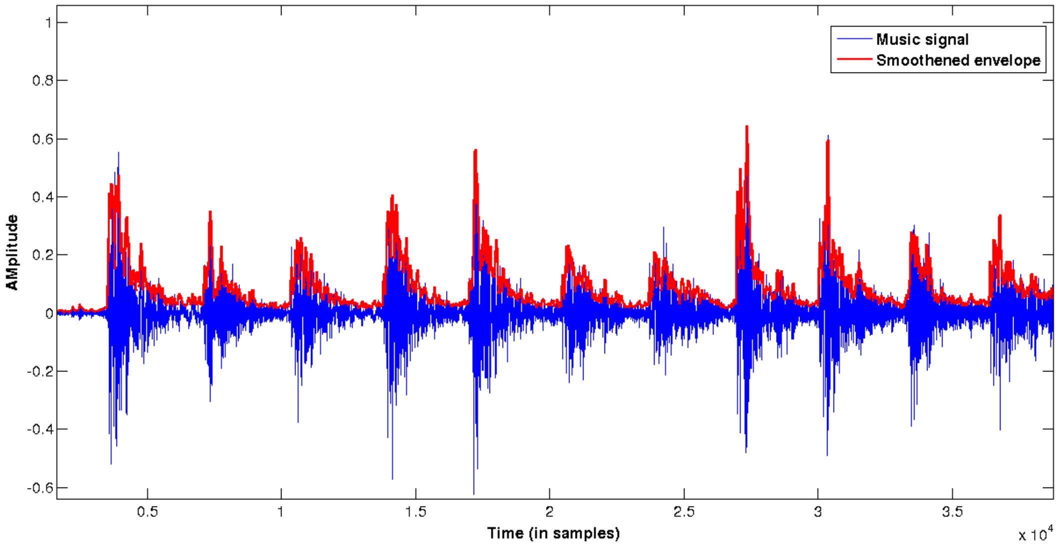


Figure 4.3: An example of a music waveform and Hilbert envelope

4.3 Minimum-phase group delay processing

Let $x[n]$ be a discrete-time signal, whose continuous phase spectrum is given by $\theta(\omega)$. The group delay function $\tau(\omega)$ is defined as

$$\tau(\omega) = -\frac{d(\theta(\omega))}{d\omega} \tag{4.6}$$

An alternate form to compute the same directly from magnitude spectrum exists:

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2} \tag{4.7}$$

where the subscripts R and I represent real and imaginary parts of the Fourier spectrum respectively. $X(\omega)$ and $Y(\omega)$ denote the discrete time Fourier transforms of $x[n]$ and $nx[n]$ respectively.

For a cascade of resonators, the group delay function exhibits high spectral resolution due to the additive property of phase - Figure 4.4 illustrates this phenomenon by considering two minimum-phase systems, complex conjugate poles at (i) ($0.8e^{j0.1\pi}$, $0.8e^{j0.3\pi}$) and (ii) ($0.8e^{j0.1\pi}$,$0.8e^{j0.5\pi}$). The peaks are not resolved in the case of the magnitude spectrum for system (i). Moreover, the peak locations do not coincide exactly with the poles in both the systems, while the ability of group delay function to clearly differentiate between the poles is visible.

It was also shown that for minimum-phase signals, the group delay function and the magnitude spectrum resemble each other [51]. This property, alongwith the high spectral resolution has since been used in group delay based feature extraction [21, 39, 38]

Zeroes outside the unit circle appear as peaks in the group delay domain. This makes it difficult to differentiate them from poles inside the unit circle, which exhibit the same phenomenon in the group delay domain. This results in a drawback of group delay functions for representing non minimum-phase signals. As practical signals are rarely minimum-phase, and zeroes in the vicinity of the unit circle are common, the group delay function cannot be applied for estimation or segmentation analysis in it's original form. [36] showed that it is possible to derive a minimum-phase equivalent from any non minimum-phase signal using the root cepstrum method. The causal por-

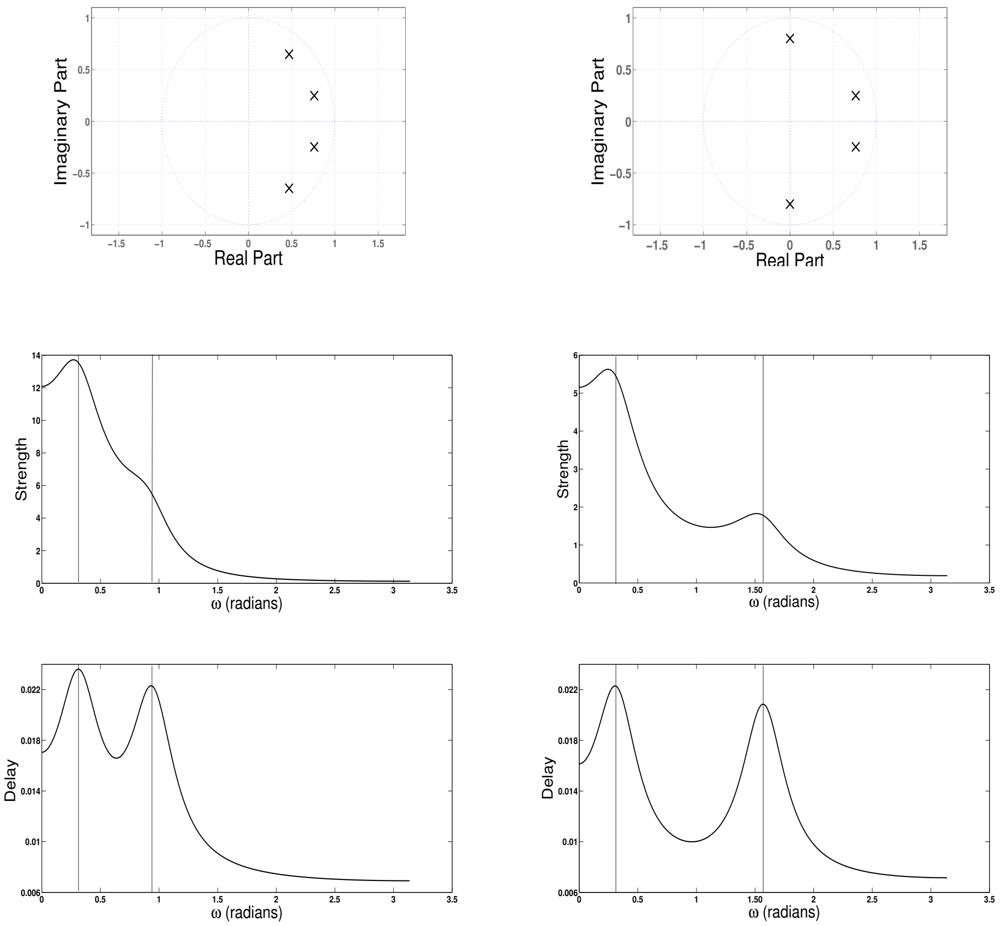


Figure 4.4: Resolving power of the group delay function *(Top)* Pole-Zero plots for two minimum-phase systems. *(Center)* Corresponding magnitude spectra with the resonant frequencies marked *(Bottom)* Group delay spectra.

tion of the inverse Fourier transform of the squared magnitude spectrum (in fact, to any power γ) was proved to be minimum-phase. This property has since been exploited for segmentation of speech into syllables [40, 35, 34].

4.4 Implementation and Results

The music signal $s(t)$ is considered as an AM-FM signal. Information about the onsets is postulated to lie in the message signals $m_1(t)$ and $m_2(t)$ which are extracted in the final detection function. First, the difference operation is applied, which brings both message signals in the envelope (4.1). Next, the analytic function is computed for the resulting signal using the Hilbert transform. The absolute value results in the envelope function. As it is, the envelope function is quite noisy, and a median smoothing of window $10ms$ is applied.

The smoothed envelope function $e(t)$ is considered as one half of the magnitude spectrum for a hypothetical signal $x(t)$. The assumption is valid as $e(t)$ is a positive function. $x(t)$ is evaluated using the inverse Fourier transform after making $e(t)$ symmetric. $x(t)$ is windowed using a one-sided Hanning/Hamming window to remove high frequency components. Group delay of $x(t)$ is computed using (4.7) since $x(t)$ is always a minimum-phase signal. The resultant group delay spectrum is termed as minimum-phase group delay function and is considered as the *detection function* in this thesis.

A brief overview of the algorithm is presented :

Onset Detection Algorithm

```
Differentiate the music signal $s(t)$ to obtain $s'(t)$.
Compute the analytic signal $s'_a(t) = s'(t) + s'_H(t)$.
Obtain the envelope $e(t) = |s'_a(t)|$.
Compute $x(t) = \mathcal{F}^{-1}(e(t) + e(-t))$. $x(t)$ is necessarily minimum-phase.
Compute detection function $d(t)$ = Group delay of $x(t)$ using (4.7).
```

The working of the algorithm is illustrated with an example in Figure 4.5. A phrase of mridangam tani is shown, with intermediate outputs from the algorithm along with the *detection function*. Notice a silent stroke at 1,20,000 samples in the original signal which gets amplified by the difference operator, and the subsequent envelope detection and group delay processing steps.

Datasets

No annotated datasets exist for Carnatic music in literature, and a new one has been created during the course of this work. *Tani-avarthanam* recordings have been collected for mridangam, ghatam, kanjira, morsing and thavil instruments. The recordings have been split into meaningful phrases in the case of mridangam by professional musicians. For all other instruments, segments measuring $20s$ in length have been used in this work. All recordings are sampled at $44.1KHz$. The details are provided in Table 4.1.

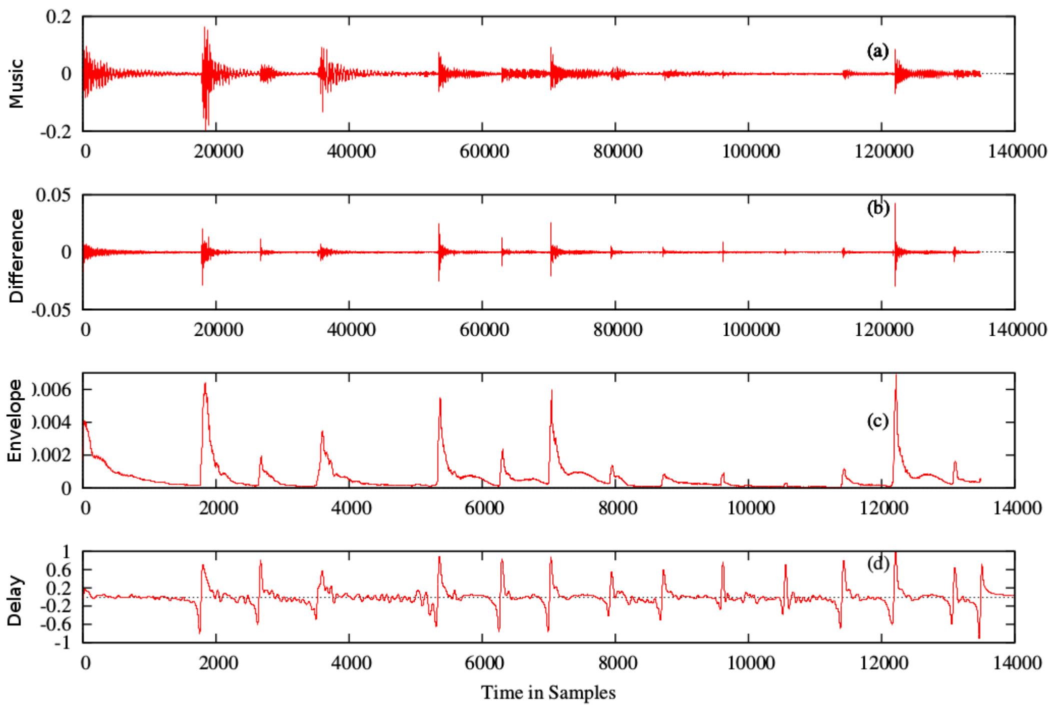


Figure 4.5: **a)** Music excerpt of Mridangam tani. **b)** Difference signal of (a). **c)** Estimated envelope of (b). **d)**Group delay function of the inverse Fourier transform of (c)

Table 4.1: Details of dataset used for evaluation

S No	**Instrument**	**Duration**	**Onsets**
1	Mridangam	18:41	5982
2	Ghatam	4:14	2616
3	Kanjira	3:11	1377
4	Morsing	6:35	2184
5	Thavil	4:39	2904
6	Ensemble	5:00	2529
	Total	42:20	17592

Performance Measures

An onset is treated as correct (*True Positive*) if it is reported within a threshold ($\pm 50ms$) of the ground truth. The tolerance is introduced to account for errors in manual annotation. A *False Positive* does not fall within the threshold of any of the time instants in the ground truth whereas a *False Negative* is a missed onset. The following metrics are used to evaluate the algorithm :

$$Precision(P) = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{4.8}$$

$$Recall(R) = \frac{N_{FN}}{N_{TP} + N_{FP}} \tag{4.9}$$

$$F\text{-}measure = \frac{2PR}{P + R} \tag{4.10}$$

where N_{TP}, N_{FP} and N_{FN} represent the number of *True Positives*, *False Positives* and *False Negatives* respectively. Previous evaluations of onset detection algorithms [16, 5] merged closely spaced onsets (most commonly, within $30ms$) and treated them as one, based on pyscho-acoustical studies of human perception of onsets [19]. In such cases, the arithmetic mean of consecutive onsets was taken to replace the multiple onsets. We do not perform the above step since it becomes impossible to differentiate between simple and composite [1] strokes, the latter being quite common in mridangam, kanjira and thavil. Further, we have not considered the case of multiple onset outputs within the threshold of a target and a single onset output within the threshold of multiple targets. They are treated as false positives and false negatives respectively.

Comparison is made with a state of the art algorithm [44] based on convolutional neural networks (CNNs). The network is trained with 80-band Mel filter banks scaled logarithmically in magnitude from spectrograms of multiple resolutions. 102 minutes of monophonic and polyphonic instrumental recordings are used for training the network. The output activation function is smoothed and a local threshold is used to detect onsets.

By varying the thresholds in the proposed algorithm as well as in [44], we report the optimum results in Fig 4.6.

It is observed that the proposed algorithm performs comparably with the state of the art approach. It outperforms in the case of mridangam, suggesting that silent strokes are better detected by the algorithm. The proposed algorithm, however has a significantly lower F-measure than the CNN based detector in the case of kanjira and thavil, suggesting that strokes in regions of very high tempo cannot be differentiated sufficiently by the proposed method.

Through the example in Figure 4.5 and results in Figure 4.6, it is seen that group delay processing performs comparably with the state of the art machine learning techniques, and can be successfully applied as an onset detection algorithm for music signals. Onset detection using group delay function also serves as another application for

[1] Composite refers to both left-right strokes occurring together in mridangam, and strokes which involve multiple fingers in case of kanjira and thavil

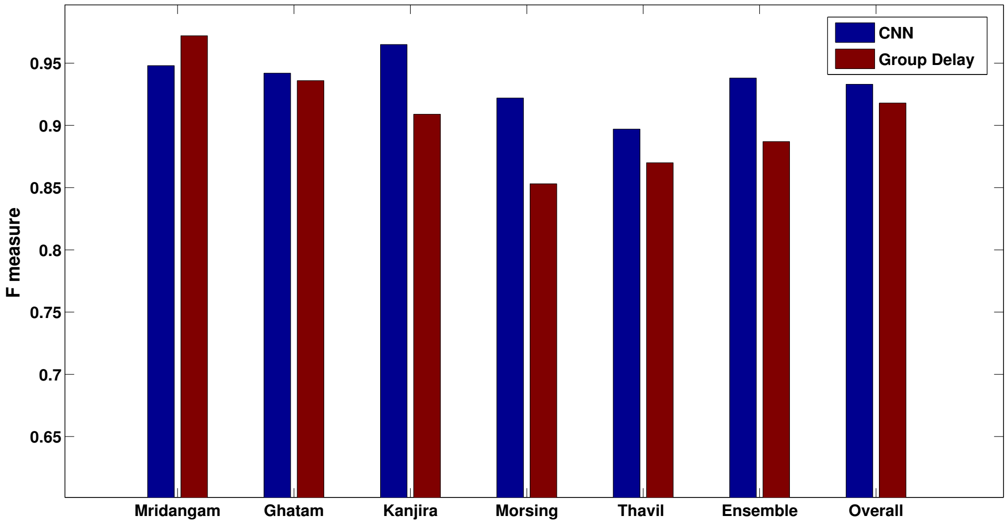


Figure 4.6: Comparison of F measures with Convolutional Neural Network (state of the art) based Onset Detection algorithm

the high resolution properties of group delay signals.

The following section discusses two modifications/improvisations to the proposed algorithm. Firstly, multi-resolution analysis is introduced, which reduces the parameter dependency. Although this modification has been evaluated with only the proposed algorithm, it can be applied in all applications where group delay processing is employed. Secondly, the possibility of extending group delay functions to a more generic onset detection algorithm is explored. Through various existing onset detection implementations, it is suggested that the group delay of the original *detection function* is a better *detection function*.

4.5 Window Scale Factor - Multi Resolution

Almost every approach to onset detection involves a threshold-based peak picking on the detection function. The choice of threshold greatly influences the performance of the algorithm, and most commercial implementations (Aubio : [7], onsetDS : [48], Note Onset Detector : [14]) leave it to the user to set the threshold. The algorithm proposed in this work consists of two parameters - Window Scale Factor *WSF* (adjusts the windowing of $x(t)$) and Threshold. While the dependence of Threshold was less straightforward to understand, an attempt is made to remove *WSF* from the algorithm.

WSF determines the number of samples to be retained in $x(t)$ (Algorithm 1). Hence, higher the *WSF*, lesser number of samples are retained, and a lower cut-off for filtering the group delay function.

$$\begin{aligned} x_1(t) &= x(t); \quad t < \frac{N_{FFT}}{WSF} \\ &= 0; \qquad t \geq \frac{N_{FFT}}{WSF} \end{aligned} \tag{4.11}$$

Hence an optimum *WSF* is required to achieve a trade-off between a noisy and over-smoothed detection function. Fig. 4.7a and 4.7b illustrate both the phenomena respectively. A low *WSF* implies a large cut-off frequency for the low pass filter, resulting in large number of false positives.

Note that some of the peaks get smoothed out when a higher *WSF* is used. However, a lower *WSF* may potentially give rise to false positives which render setting an optimum threshold a difficult task. Previous implementations of group delay processing which computed the minimum-phase group delay of short term energy (STE) functions [34, 25], have retained *WSF* as a parameter for each dataset and set peak picking threshold to 0. Later, [45] considered slightly larger windows for STE computation. Hence, STE was more smoothed out and the dependence on *WSF* was brought down to within an order of magnitude (3-8).

In the previous section, both *WSF* and Threshold for peak were optimised. Here, the dependence on *WSF* is removed completely. The smoothness of a higher *WSF* and the higher resolution of a lower *WSF* are combined by simply adding the group delay response element-wise from *WSF*$= 5$ to *WSF*$= 45$. The choice of upper and lower bounds on the *WSF* are made so as to cover the maximum possible variation in the group delay response due to *WSF*. The resultant response is observed to capture most of the true positives from the lower *WSF* while not giving rise to false positives.

An example using a mridangam excerpt is illustrated in Figure 4.7. The minimum-phase group delay response corresponding to the lower and upper bounds of *WSF* are plotted in Figures 4.7a and 4.7b. Note that *WSF* plays a significant role in the smoothing of group delay response. In Figure 4.7c, group delay responses for different *WSF* are plotted, and the final response which is a sum of the individual responses, is plotted in 4.7d.

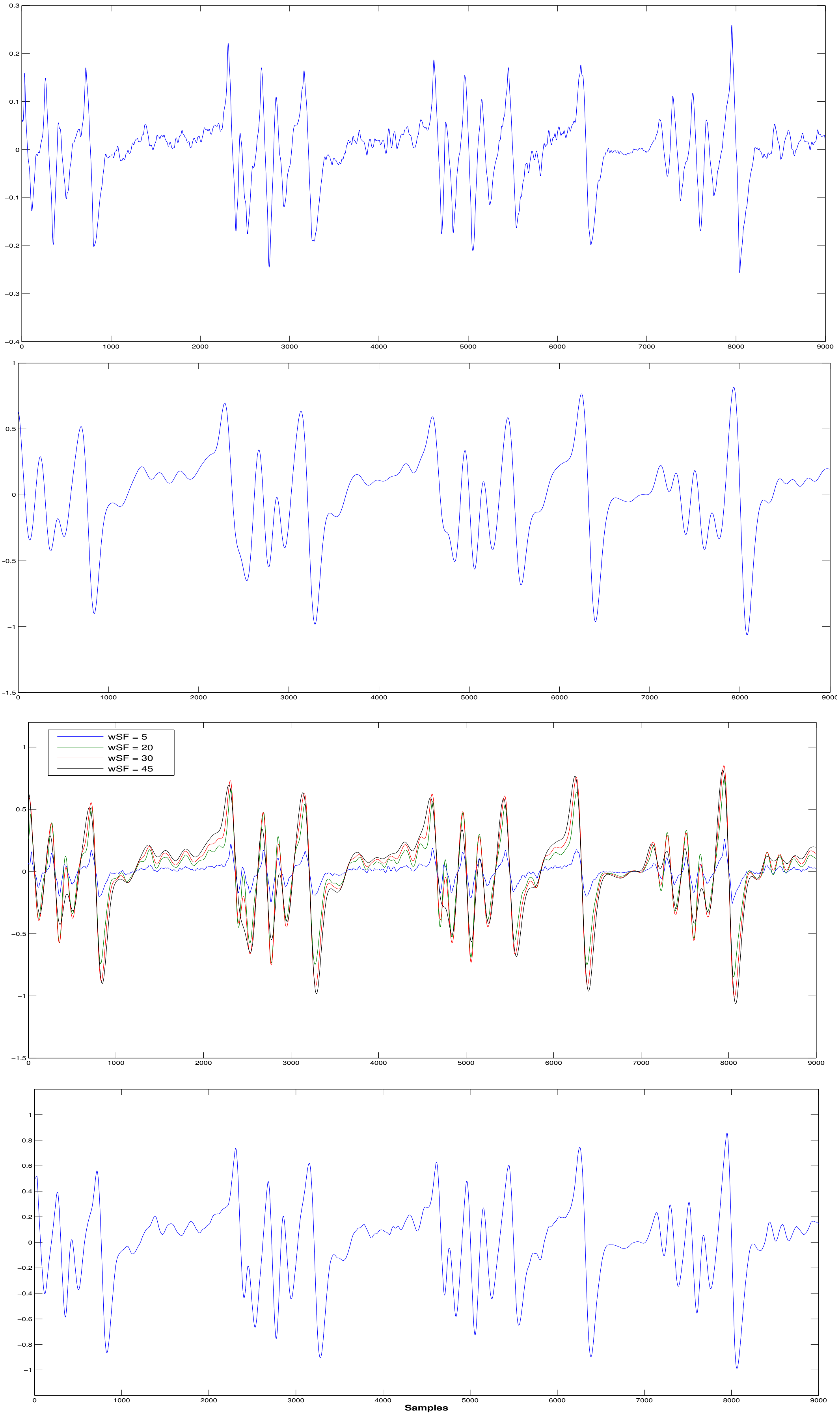


Figure 4.7: Detection functions computed with different Window Scale Factors

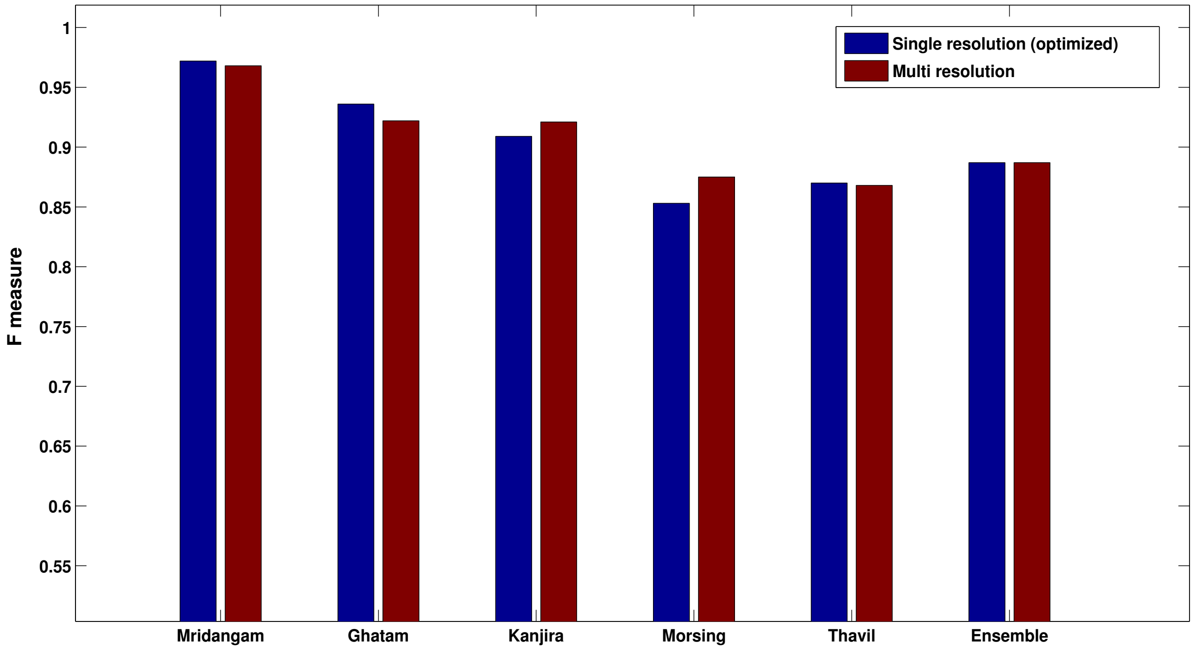


Figure 4.8: Comparison of performance on Carnatic dataset

For the purpose of evaluation, the same datasets and performance measures are used, except that optimisation is carried out for only the Threshold, as opposed to both *WSF* and Threshold done previously. Overall F-measure is found to remain constant with slight variations within instruments. This algorithm, with a multi-resolution approach is argued to be robust in comparison to the single resolution approach.

4.6 Group Delay as Post-Processing

In the algorithm proposed in this thesis, the minimum phase group delay is agnostic to the AM-FM envelope, in the sense that no assumptions are made on the envelope other than it being a positive signal. (The envelope is assumed to be the magnitude spectrum of a hypothetical signal, hence the positivity constraint). The high resolution of group delay functions could improve the temporal resolution of any suitable *detection function*. Hence, the possibility of applying group delay as a post-processing step to existing *detection functions* is explored in this section. Since the motive is to extend group delay processing as a generic algorithm, another dataset consisting of western instruments has been considered.

The western dataset was introduced in [5] and contains a mixture of various western instruments and genres. The audio is combined with datasets used in [3], [22] and [16]. By far the biggest onset-annotated dataset in literature, it is subdivided into 321 files and consists of 27,774 onsets. The total runtime length of the dataset is $\approx$ 102 minutes.

Implementations of onset detection algorithms based on energy, spectral difference, HFC, complex domain, phase and Kullback-Leibler methods are taken from the Aubio library [7]. The implementations are made available as Vamp plugins, and the Sonic Annotator [8] tool is used for extraction of the *detection functions*. The evaluation setup is as follows - First, the original Aubio implementation is optimised separately for each of the six methods mentioned. Next, each *detection function* is taken and made symmetric so as to resemble magnitude spectrum of a real signal. The inverse Fourier transform is computed to give a real signal that is minimum-phase. Group delay of this signal is computed and treated as the new *detection function*. A thresholding step is performed to identify the onsets. Results for the both approaches are presented in Figure 4.9 using the F-measures.

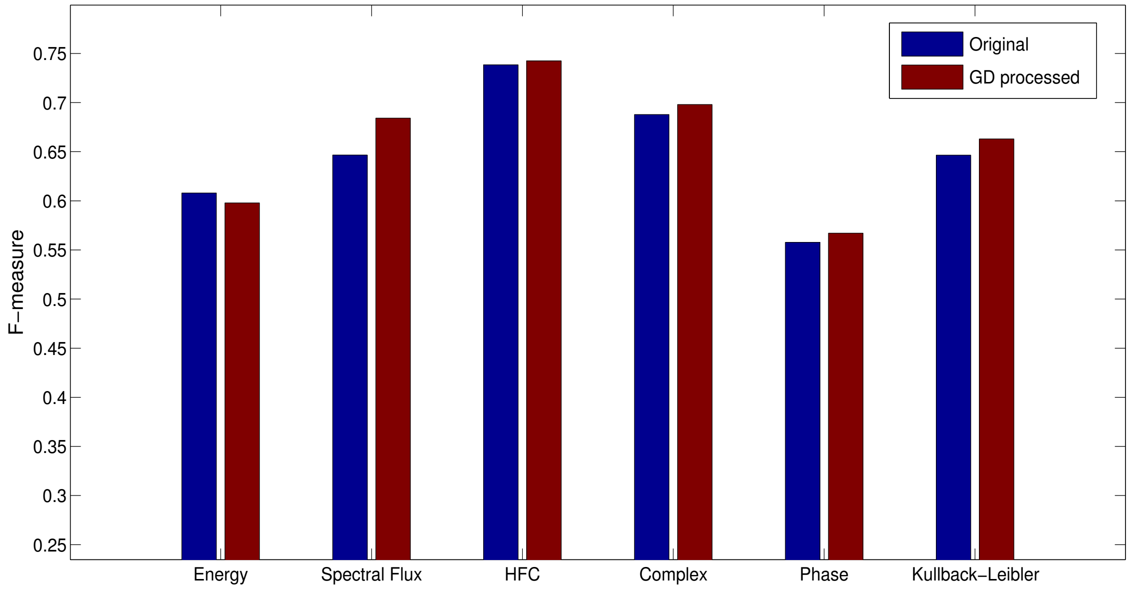


Figure 4.9: Group Delay as a post processing step - Comparison of original and GD processed implementations

The overall F-measures (averaged across the six methods) are presented in Table 4.2. In both cases, the F-measures are lower in comparison to the Carnatic dataset. This is expected due to the presence of softer onsets and vocal onsets, which do not possess a sharp attack. However, there is clearly an improvement when group delay is introduced as a post-processing step. The improvement is consistent in all methods except energy based. This shows that even though group delay processing does not add any new information to the existing *detection function*, it improves the temporal resolution enabling an easier thresholding step.

Table 4.2: Overall F-measure before and after application of GD post-processing

	Original	**GD processed**
F-measure	0.6475	0.6588

4.7 Summary

In this Chapter, a new algorithm for musical onset detection is proposed. It employs the resemblance of Carnatic percussion instruments to a generic amplitude-frequency modulated waveform. A technique is developed that extracts information about onsets into the envelope of the music signal. The extracted envelope is treated as the magnitude spectrum of a real signal, which is necessarily minimum phase. Group delay response of this signal is treated as the *detection function* in this work. The algorithm is observed to perform better on the mridangam which contains a considerable number of silent strokes. Later, two improvisations for reducing the parameter dependency of the algorithm and to extend it as a generic technique are proposed.

CHAPTER 5

Summary and Future Work

5.1 Summary

Traditional processing of audio signals mostly considered only the magnitude spectrum, and ignored the phase spectrum. Experiments were then performed to show the importance of the phase component. These experiments drew inference from the human perception with and without the presence of noisy phase. The wrapped nature of phase has made it difficult to process in it's direct form. However, the negative frequency derivative of phase, known as group delay function was found to be a successful measure for various tasks due to it's high resolution property and the additive property of phase.

In this thesis, a proof for the high resolution property is presented in the first part of the thesis. It is shown via a single-resonator system that, irrespective of the pole location, the group delay function is always sharper than the magnitude spectrum at the ndB bandwidth. Better peakedness is argued to assist the discriminating power of group delay functions in tasks like pitch estimation, formant estimation and syllable segmentation. The proof is extended to multi-pole systems using numerical measures that characterise the peakedness of a function. The result is found to be consistent for all possible pole configurations.

Next, a new algorithm is proposed for the task of musical onset detection, one which employs the high resolution property of group delay functions discussed in the first part of this thesis. The music waveform is treated as an amplitude-frequency modulated waveform, and information about the onsets is proposed to lie in the message signals which have been modulated. Using Hilbert transform, the message signals are emphasised in the envelope of the music signal. Group delay processing is applied on the smoothed envelope to detect onsets. The proposed algorithm is shown to perform comparably with a state of the art machine learning approach. Two improvisations are proposed, one removes the parameter dependency, and the other extends group delay processing to a more generic onset detection algorithm.

5.2 Criticisms of the Work

- A theoretical proof was proposed to show that the strength of group delay function at the ndB bandwidth of the magnitude spectrum is always lesser. The proof applies only to a single-resonator system, since it is not possible to derive generic expressions for the ndB bandwidth without assumptions about the pole locations and pole bandwidths. Hence, numerical measures were used to justify the peakedness.
- The proposed algorithm is found to degrade in performance when applied on softer onsets (eg. string and bowed instruments, vocal onsets). The reason being such examples were significantly different from an amplitude-frequency modulated waveform. The extracted envelope was found to be inadequate to discriminate the onsets from steady-state regions.
- Although group delay as a post processing step was found to improve performance on a much larger dataset, the F-measures are still way behind the state of the art techniques.

5.3 Scope for Future Work

- The accuracy of the proposed method by working in tandem with a proven machine learning algorithm, say (Deep) neural networks ([16, 30, 29]) can be explored. These methods have been observed to work well in the presence of large training data. The necessity of training data can be removed by correction with a data-agnostic signal processing technique like the proposed algorithm.
- Many MIR tasks can be carried forward with a good onset prediction - isolated style training/testing of stroke models for recognition purposes, detection of salient events in a *tani-avarthanam* based on cues from stroke instants, $t\bar{a}la$ classification based on stroke instants and stroke transcriptions and so on.

APPENDIX A

Carnatic Percussion Instruments

A.1 Mridangam

Mridangam is the most widely used percussion accompaniment in Carnatic music. It resembles a two sided drum with tightly stretched membranes on either side. The two sides are unequal in size and a right handed artist positions the smaller side to his/her right. The *tonic* is defined as the base pitch, which is used as a reference for all other higher harmonics. The strokes can be categorized based on the side of the mridangam being played (all left side strokes are tonic independent, while some right side strokes are tonic dependent) and the position, manner and force with which the membranes are struck. The two sides allow composite strokes (one from the left side and one from the right side at the same instant) to be created which from an MIR perspective ought to be treated as one, although they sometimes appear as separate strokes while performing the onset detection analysis. The first study on mridangam carried out by Nobel prize winning scientist Raman [42] and later by [47] analyzed the harmonics of the strokes and more recently, [2] employed a non negative matrix factorization to classify the strokes. Details about the stroke nomenclature can be found in [2] and [43].

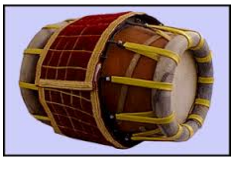
A.2 Ghatam

The ghatam is a hollow pot that is placed on the lap of the artist and struck with the palm and fingers. The mouth is positioned facing towards the artist, although the direction is reversed occasionally. It is made of specially burnt clay with metallic powder for strength and care is taken that the walls of ghatam are of equal thickness. Distinct ghatam strokes count lesser in number than mridangam. Tuning of the pitch is possible to a limited extent by application of *play-doh*, but mostly another ghatam is chosen to achieve significant variations. Ghatam strokes also produce a characteristic sound when struck on the neck of the pot. The artist modulates the sound by altering the size of the mouth during the performance, by partly or fully closing the area of the mouth with palms.

A.3 Morsing

Known as *Jew's Harp*, the morsing is a wind percussion instrument. It resembles a metallic clamp with a thin film (the *tongue*) in between them. The instrument is caught by the hand and placed in the mouth of the artist, the teeth firmly holding it in place. Sound is produced inside the mouth of the artist by triggering the tongue of the instrument with the index finger. The artist's tongue is also used to produce morsing notes. The instrument pitch cannot be varied significantly and the artists prefer to carry morsings of different dimensions for the purpose of fine tuning.

A.4 Thavil



The thavil is similar to the mridangam in the sense that it is a two sided barrel, with both sides participating in sound production. However, a right handed artist positions the larger side to his right, unlike the mridangam. The left side is struck with a stick while the artist plays the right side with fingertips covered with *thumb caps*. The *thumb caps* are mostly made of hardened rice flour and give rise to sharp, cracking sounds. The thavil is played in festivities as much as in professional concerts, and almost certainly accompanies the *nadaswaram* (a wind instrument). Variations in pitch are achieved by tightening the left side of the instrument. Distinct strokes exist, based on the side of the instrument struck and the number of fingers involved in production (for the right side).

Specifically, *Ta* and *Di* involve four fingers, but are still treated as a single stroke.

A.5 Kanjira

The kanjira is a one-sided percussion instrument and is small enough to be held with one hand. The instrument is made of monitor lizard belly skin stretched across a circular wooden frame made from the jackfruit tree. High pitched sound is produced by striking the circular face with the palm and fingers of the free hand. Unlike the mridangam, the face of the kanjira is not loaded with any paste. The pitch can be varied to an extent by applying pressure on the face using the hand holding the kanjira or by sprinkling water on the kanjira skin from behind.

GLOSSARY

Carnatic Concert A typical concert consists of a number of segments (*items*). It begins with a *varnam* - the opening song, and followed by *compositions*, where the vocalist renders a solo performance initially followed by the melodic and rhythmic accompaniments. The last parts feature a solo percussion (*tani avarthanam*) which forms the focus of the onset detection task. A more detailed discussion about a concert can be found in [43].

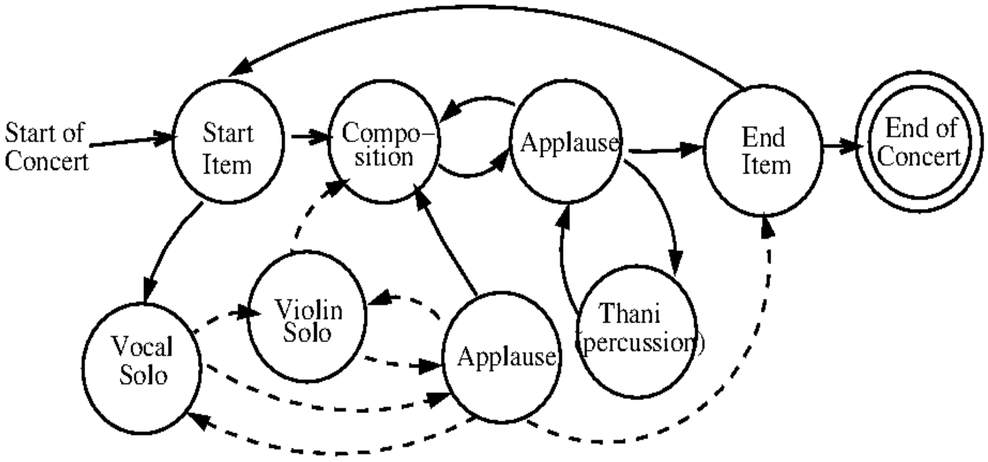


Figure A.1: General structure of a Carnatic music concert. Solid lines indicate paths that are mandatory and dotted lines indicate optional paths which are not mandatory. (From [43])

Tani avarthanam The artistic skill of the percussionist is displayed in this segment towards the end of the concert. The rhythmic cycles conform to a *tala*, which determine the number of beats per cycle.

Mohra Special patterns characteristic to the *tala* which signify the end of a *tani avarthanam*. Having a dictionary of *mohra* can help in identification of *tala* provided a good stroke recognition system is in place.

REFERENCES

[1] **Alsteris, L. D.** and **K. K. Paliwal** (2007). Short-time phase spectrum in speech processing: A review and some experimental results. *Digital Signal Processing*, **17**(3), 578 – 616. ISSN 1051-2004.

[2] **Anantapadmanabhan, A.**, **A. Bellur**, and **H. A. Murthy**, Modal analysis and transcription of strokes of the mridangam using non-negative matrix factorization. *In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. 2013.

[3] **Bello, J. P.**, **L. Daudet**, **S. Abdallah**, **C. Duxbury**, **M. Davies**, and **M. B. Sandler** (2005). A tutorial on onset detection in music signals. *Speech and Audio Processing, IEEE Transactions on*, **13**(5), 1035–1047.

[4] **Bellur, A.** and **H. A. Murthy**, A novel application of group delay functions for tonic estimation in carnatic music. *In Eurpoean Conference on Signal Processinge*. 2013. ISSN 2219-5491.

[5] **Böck, S.**, **F. Krebs**, and **M. Schedl**, Evaluating the online capabilities of onset detection methods. *In Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2012)*. 2012.

[6] **Bozkurt, B.** and **L. Couvreur**, On the use of phase information for speech recognition. *In European Conference on Signal Processing*. 2005.

[7] **Brossier, P. M.** (2006). *Automatic annotation of musical audio for interactive applications*. Ph.D. thesis, Queen Mary, University of London.

[8] **Cannam, C.**, **M. O. Jewell**, **C. Rhodes**, **M. Sandler**, and **M. d'Inverno** (2010). Linked data and you: Bringing music research software into the semantic web. *Journal of New Music Research*, **39**(4), 313–325.

[9] **DeCarlo, L. T.** (1997). On the meaning and use of kurtosis. *Psychological methods*, **2**(3), 292.

[10] **Dixon, S.**, Onset detection revisited. *In Proc. of the 9th Int. Conference on Digital Audio Effects (DAFx) pp. 133-137*. 2006.

[11] **Dubnov, S.** (2004). Generalization of spectral flatness measure for non-gaussian linear processes. *Signal Processing Letters, IEEE*, **11**(8), 698–701.

[12] **Duncan**, **H. A. Murthy**, and **B. Yegnanarayana** (1989). A nonparametric method of formant estimation using group delay spectra. *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, **1**, 572–575.

[13] **Duxbury, C.**, **J. P. Bello**, **M. Davies**, **M. Sandler**, *et al.*, Complex domain onset detection for musical signals. *In Digital AudioEffects Workshop (DAFx) No. 1, pp. 6-9*. 2003.

[14] **Duxbury, C.**, **J. P. Bello**, and **C. Landone**, Note onset detector. 2007. URL `http://vamp-plugins.org/plugin-doc/qm-vamp-plugins.html`.

[15] **Duxbury, C.**, **M. Sandler**, and **M. Davies**, A hybrid approach to musical note onset detection. *In Digital AudioEffects Workshop (DAFx,'02) (pp. 33-38)*. 2002.

[16] **Eyben, F.**, **S. Böck**, **B. Schuller**, and **A. Graves**, Universal onset detection with bidirectional long short-term memory neural networks. *In Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2010)*. 2010.

[17] **Gabor, D.** (1946). Theory of communication. *The Journal of the Institution of Electrical Engineers*, **93**(26), 429–457.

[18] **Goto, M.** and **Y. Muraoka**, Beat tracking based on multiple-agent architecture a real-time beat tracking system for audio signals. *In Proc. Second International Conference on Multiagent Systems*. 1996.

[19] **Handel, S.**, *Listening: An Introduction to the Perception of Auditory Events*. MIT Press, 1989.

[20] **Hegde, R. M.**, **H. A. Murthy**, and **V. R. R. Gadde**, Continuous speech recognition using joint features derived from the modified group delay function and MFCC. *In INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004*. 2004.

[21] **Hegde, R. M.**, **H. A. Murthy**, and **V. R. R. Gadde** (2007). Significance of the modified group delay features in speech recognition. *IEEE Transactions on Audio,Speech and Language Processing*, **15**, 190–202.

[22] **Holzapfel, A.**, **Y. Stylianou**, **A. Gedik**, and **B. Bozkurt** (2010). Three dimensions of pitched instrument onset detection. *Audio, Speech, and Language Processing, IEEE Transactions on*, **18**(6), 1517–1527. ISSN 1558-7916.

[23] **Jayant, N. S.** and **P. Noll**, *Digital Coding of Waveforms, Principles and Applications to Speech and Video*. Prentice-Hall, Englewood Cliffs NJ, USA, 1984, 688. N. S. Jayant: Bell Laboratories; ISBN 0-13-211913-7.

[24] **Johnston, J. D.** (1988). Transform coding of audio signals using perceptual noise criteria. *Selected Areas in Communications, IEEE Journal on*, **6**(2), 314–323.

[25] **Kamakshi., P. V.**, **N. T.**, and **H. A. Murthy** (2004). Automatic segmentation of continuous speech using minimum phase group delay functions. *Speech Communications*, **42**, 429–446.

[26] **Lakshmi, A.** and **H. A. Murthy**, A new approach to continuous speech recognition in indian languages. *In National Conference on Communication*. 2008.

[27] **Levine, S. N.** (1998). *Audio Representations for Data Compression and Compressed Domain Processing*. Ph.d. dissertation, Department of Electrical Engineering, CCRMA, Stanford University. URL `http://www-ccrma.stanford.edu/ scottl/thesis.html`.

[28] **Lim, J. S.**, Spectral root homomorphic deconvolution system. *In IEEE Trans. Acoust., Speech, Signal Processing*. 1979.

[29] **Marchi, E.**, **G. Ferroni**, **F. Eyben**, **L. Gabrielli**, **S. Squartini**, and **B. Schuller**, Multi-resolution linear prediction based features for audio onset detection with bidirectional LSTM neural networks. *In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*. 2014.

[30] **Marchi, E.**, **G. Ferroni**, **F. Eyben**, **S. Squartini**, and **B. Schuller**, Audio onset detection: A wavelet packet based approach with recurrent neural networks. *In Neural Networks (IJCNN), 2014 International Joint Conference on*. 2014.

[31] **Masri, P.** (1996). *Computer Modeling of Sound for Transformation and Synthesis of Musical Signals*. Ph.D. thesis, University of Bristol, UK.

[32] **Murthy, H. A.** and **G. V. R. Rao** (2003). The modified group delay function and its application to phoneme recognition. *ICASSP*, 1.68–1.71.

[33] **Murthy, H. A.** and **B. Yegnanarayana** (2011). Group delay functions and its application to speech processing. *Sadhana*, **36**(5), 745–782.

[34] **Nagarajan, T.**, **H. A. Murthy**, and **R. M. Hegde**, Segmentation of speech into syllable-like units. *In Proceedings of EUROSPEECH*. Geneva, Switzerland, 2003.

[35] **Nagarajan, T.**, **V. K. Prasad**, and **H. A. Murthy**, The minimum phase signal derived from the magnitude spectrum and its applications to speech segmentation. *In SPCOM*. 2001.

[36] **Nagarajan, T.**, **V. K. Prasad**, and **H. A. Murthy** (2003). Minimum phase signal derived from the root cepstrum. *IEE Electronics Letters*, **39**, pp.941–942.

[37] **Oppenheim, A.** and **J. Lim** (1981). The importance of phase in signals. *Proceedings of the IEEE*, **69**(5), 529–541. ISSN 0018-9219.

[38] **Padmanabhan, R.**, **S. H. K. Parthasarthi**, and **H. A. Murthy**, Robustness of phase based features for speaker recognition. *In Proceedings of Int. Conf. Spoken Language Processing*. 2009.

[39] **Parthasarathi, S.**, **R. Padmanabhan**, and **H. A. Murthy** (2011). Robustness of group delay representations for noisy speech signals. *International Journal of Speech Technology*, **14**(4), 361–368.

[40] **Prasad, V. K.**, **T. Nagarajan**, and **H. A. Murthy**, Automatic segmentation of continuous speech using minimum phase group delay functions. *In Speech Communications*, volume 42. 2004.

[41] **Rajesh M Hegde** (2005). *Fourier transform phase based features for speech recognition*. PhD dissertation, Indian Institute of Technology Madras, Department of Computer Science and Engg., Madras, India.

[42] **Raman, C.**, The Indian Musical Drums. *In Proceedings of the Indian Academy of Sciences-Section A*, volume 1. Springer, 1934.

[43] **Sarala, P.** (2014). *Automatic Segmentation of Continuous Audio Recordings of Carnatic Music Concerts into Items for Archival*. Ph.d. dissertation, Department of Computer Science and Enigneering, IIT Madras.

[44] **Schlüter, J.** and **S. Böck**, Improved Musical Onset Detection with Convolutional Neural Networks. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*. Florence, Italy, 2014.

[45] **Shanmugam, S. A.** and **H. A. Murthy**, A hybrid approach to segmentation of speech using group delay processing and hmm based embedded reestimation. *In INTERSPEECH*. 2014.

[46] **Shi, G.**, **M. Shanechi**, and **P. Aarabi** (2006). On the importance of phase in human speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, **14**(5), 1867 –1874.

[47] **Siddharthan, R.**, **P. Chatterjee**, and **V. Tripathi**, A study of harmonic overtones produced in indian drums. *In Physics Education*. 1994.

[48] **Stowell, D.** and **M. D. Plumbley**, Adaptive whitening for improved real-time audio onset detection. *In Proceedings of the International Computer Music Conference (ICMC'07)*, volume 2. 2007.

[49] **Tribolet, J. M.** (1979). A new phase unwrapping algorithm. *IEEE Trans. Acoustics Speech and Signal Processing*, **ASSP-**(2), 170–179.

[50] **Wu, S. L.**, **E. D. Brian**, **E. D. Kingsbury**, **N. Morgan**, and **S. Greenberg.**, Incorporating information from syllable-length time scales into automatic speech recognition. *In Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, volume 2. 1998.

[51] **Yegnanarayana, B.** (1979). Formant extraction from linear prediction phase spectra. *Acoustical Society of America*, **63**, 1638–1640.

[52] **Yegnanarayana, B.** and **H. A. Murthy** (1992). Significance of group delay functions in spectrum estimation. *IEEE Trans. Signal Processing*, **40**(9), 2281–2289.

[53] **Ying, L.** (2006). Phase unwrapping. *Wiley Encyclopedia of Biomedical Engineering*.

LIST OF PAPERS BASED ON THESIS

1. **Manoj Kumar** and Jilt Sebastian and Hema A Murthy "Musical Onset Detection on Carnatic Percussion Instruments" *National Conference on Communications 2015 (NCC-2015)*, Mumbai, India, Feb 28 - Mar 2 (2015).

2. **Manoj Kumar**, Jom Kuriakose, Jilt Sebastian, Sridharan Sankaran and Hema A Murthy "Onset detection and stroke recognition for percussion instruments in Carnatic music" *Musical Rhythm: Cross-Disciplinary and Multi-Cultural Perspectives (*Invited Paper*)*, Abu Dhabi, Oct 12-15 (2014).

3. **Manoj Kumar**, Jilt Sebastian and Hema A Murthy "High Resolution Property of Group Delay function" *Electronics Letters (Submitted)*

OTHER PAPERS

1. Jilt Sebastian and **Manoj Kumar** and Hema A Murthy "Pitch Estimation From Speech Using Grating Compression Transform on Modified Group-Delay-gram" *National Conference on Communications 2015 (NCC-2015)*, Mumbai, India, Feb 28 - Mar 2 (2015).