Real-time Onset Detection in Music Signals

EE779 - ATSP Course Project

Yash Bhalgat (13D070014) | Kalpesh Patil (130040019)

November 18, 2016

Abstract

Due to the recent progress in score-following and beat-tracking in various music systems used worldwide, it has become very important to segment music signals temporally using points called 'onsets'. Score following using real-time onset detection is still a very active area of research in AI, Pattern Recognition and Signal Processing. Our project reviews some standard techniques in onset detection and then explore additional techniques, namely Linear Prediction and Sinusoidal Modelling for real-time onset detection. We have tried to implement the techniques on some standard available signals and have included comprehensive results demonstrating the effectiveness of each of the techniques described in our report.

Introduction

Score following systems use detected note events to interact directly with a live performer. And beat-synchronizing systems group detected notes into beats and then use this knowledge to improve an underlying analysis process. Onset-detection plays a vital role in temporal segmentation which is used in both these systems. There are several published techniques for onset detection which majorly involve formulation of various ODFs (Onset-detection functions) and peak extraction for the localization of onsets. We first discuss some of the standard ODFs, namely energy ODF, spectral difference ODF and the complex domain ODF. In the later part of the report, we discuss the ways to improve on these techniques using linear prediction and a technique based on sinusoidal modelling.

Definitions

Onsets: An instant marking the beginning of a transient or a note.

Transient: Short interval during which the signal behaves in a relatively unpredictable way.

Attack: The time interval during which the amplitude envelope increases

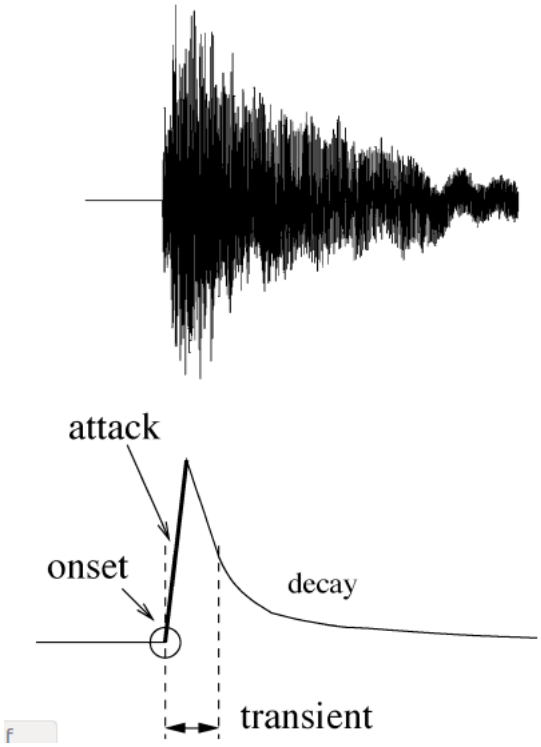


Figure 1: onset and attack demo

Multi-step approach

Onset Detection Functions

Onset detection function is primarily an undersampled version of the original music signal. As shown in the preprocessing section, we divide the signal into partially overlapping frames and the ODF consists of one value for each frame. By the definition of an onset, we can say that onset detection is the process of identifying which parts of a signal are relatively unpredictable. Hence, each value in an ODF should give a good indication as to the measure of the unpredictability of that frame. The vector of these values (obtained using methods discussed later) is passed to the peak-detection algorithm for onset detection.

Peak Detection

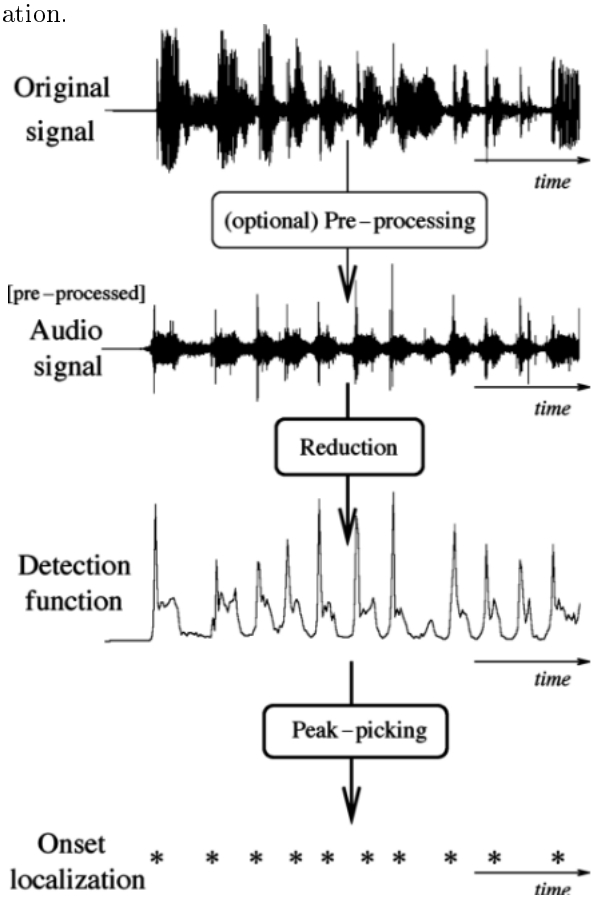
Appropriate ODFs give identifiable peaks at the onsets or abrupt events. In this process, we identify the local maxima, also called peaks in the ODF. The first stage in peak-detection algorithm is to identify the local maxima using the neighbouring values. As we cannot 'look ahead', the previous value needs to be saved until we reach the next frame. This leads to a latency equal to the buffer-size every time. And in the second step, the peak is recorded as an onset location if its value is greater than a certain threhold.

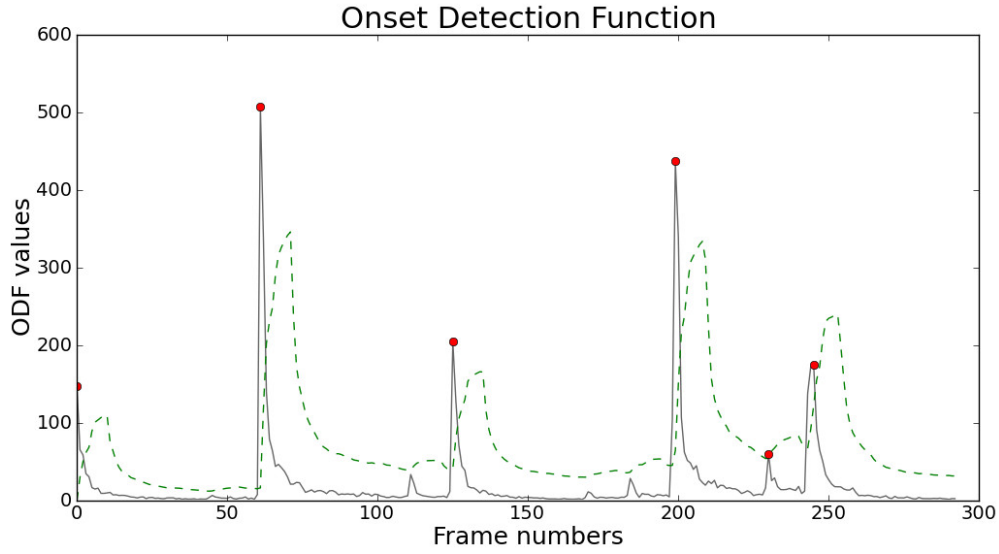
Smooth Thresholding

Since we want to avoid the false peaks and the onset-detection is real-time, we use calculate the thresholds using a slight variation of the median/mean function for each frame.

$$\sigma_n = \lambda \times median\left(O\left[n_m\right]\right) + \alpha \times mean\left(O\left[n_m\right]\right) + N$$

Here, the $O[n_m]$ is the previous m values of the ODF of frame n.
Every ODF peak that is above this threshold (see figure below) is taken to be a note onset location.





Types of ODFs

We will discuss some of the standard ODF techniques and then go on to discuss the LP method and the Sinusoidal Modelling method for significants improvements on these techniques. In each of the following methods, the ODFs retum one value for every frame, corresponding to the likelihood of that frame containing a note onset.

Energy method

This is the most simple and computationally efficient method. In this method, we assumt that the onsets correspond to a higher energy component than the steady state notes in the music signal. So, for each frame the energy is defined as:

$$E\left(n\right) = \sum_{k=0}^{N} x\left(k\right)^2$$

Now, larger abrupt changes in the amplitude envelope of the energy signal are expected to coincide with onset locations. Hence, the energy ODF is defined as:

$$ODF_E\left(n\right) = |E(n) - E(n-1)|$$

Spectral Difference method

Spectral difference method is successful in detecting onsets in polyphonic signals (multiple notes simultaneously) and 'soft' onsets created by instruments which do not have a percussive attack (unlike the energy method). This is achieved by identifying time-varying changes in a

frequency domain representation of an audio signal, i.e. we calculate the frame-to-frame in the Short Time Fourier Transform (STFT).

$$X(k,n) = \sum_{m=0}^{N-1} x(m)w(m)e^{\frac{-2j\pi mk}{N}}$$

$$ODF_{SD}(n) = \sum_{m=0}^{N/2} ||X(k,n)| - |X(k,n-1)||$$

Complex domain method

Instead of making predictions only on the magnitudes like the spectral difference method, the complex domain ODF attempts to improve the prediction for the next value of a given bin using combined magnitude and phase information. The phase prediction is formed by assuming a constant rate of phase change between frames. The method is implemented using the following equations:

$$\hat{R}(k,n) = |X(k,n-1)|$$

$$\hat{\phi}(k,n) = princarg\left[2\phi(k,n{-}1) - \phi(k,n{-}2)\right]$$

$$\Gamma(k,n) = \sqrt{\hat{R}(k,n)^2 + R(k,n)^2 - 2R(k,n)\hat{R}(k,n)\cos(\phi(k,n) - \hat{\phi}(k,n))}$$

On obtaining the values for $\Gamma(k,n)$, we get the complex domain ODF as:

$$ODF_{CD}(n) = \sum_{k=0}^{N/2} \Gamma(k,n)$$

Linear Prediction method

In this method, we try to distinguish between the steady-state and transient regions of an audio signal by making predictions. We use an arbitrary number of previous values combined with LP to improve the accuracy of the estimate. The ODF is then the absolute value of the differences between the actual frame measurements and the LP predictions. The ODF values are low when the LP prediction is accurate, but larger in regions of the signal that are more unpredictable, which should correspond with note onset locations.
We use the Burg's method to calculate the LP coefficients. According to the literature, Burg method gives the most accurate and consistent results in real-time compared to the autocorrelation and the covariance method. Also, it has a minimum phase and estimates the coefficients on a finite support.

Burg Algorithm

Minimise the forward prediction error $f_m(n)$ and the backward prediction error $b_m(n)$.

Algorithm 0.1 Burg algorithm:
f ← x
b← x
a← x
for m←0 to p-1 do
fp←f without its first element
bp←b without its last element
γ← -2bp*fp/(fp*fp+bp*bp)
f← fp+γ*bp
b← bp+γ*fp
a←(a[0], a[1], ..., a[m], 0) + k (0, a[m], a[m - 1], ..., a[0])

$$f_0(n) = b_0(n) = x(n)$$

Recursion:

$$f_m(n) = f_{m-1}(n) - \gamma b_{m-1}(n-1)$$

$$b_m(n) = b_{m-1}(n-1) - \gamma f_{m-1}(n)$$

The expression for γ is obtained by minimising the cost of the backward and forward prediction error.

All the previously described methods can be combined with LP prediction to give better results.

Energy method with LP:

$$ODF_{ELP}(n) = |E(n) - P_E(n)|$$

Spectral difference with LP:

$$ODF_{SDLP}(n) = \sum_{m=0}^{N/2} ||X(k,n)| - |P_{SD}(k,n)||$$

Complex domain method with LP:

$$ODF_{CDLP}(n) = \sum_{k=0}^{N/2} ||\Gamma(k,n)| - |P_{CD}(k,n)||$$

Sinusoidal modelling approach

Fourier's theorem: Any periodic waveform can be modelled as the sum of sinusoids at various amplitudes and harmonic frequencies.
For stationary pseudo-periodic sounds, these amplitudes and frequencies evolve slowly with time. They can be used as parameters to control pseudo-sinusoidal oscillators, commonly referred to as partials.
Additive synthesis: adding together many sinusoidal components modu- lated by relatively slowly varying amplitude and frequency envelopes

$$y(t) = \sum_{i=1}^{N} A_i(t) \sin\left(\theta_i(t)\right)$$

$$and\ \theta_i(t) = \int_0^t \omega_i(t)dt + \theta_i(0)$$

Partial tracking: Calculating these parameters for each frame is referred to as peak detection, while the process of connecting these peaks between frames is called partial tracking.
Two methods are used for onset detection using Sinusoidal modelling: Online processing technique and Offline processing technique.

Offline processing technique

In this method, transient signals in the time domain can be mapped onto sinusoidal signals in a frequency domain using DCT. But this method is not suitable for real time since DCT frame length required is very large. A multi-resolution sinusoidal model is then applied to the signal to isolate the harmonic component of the sound. And finally, Onset Location is determined by abrupt increase in energy of the frame.

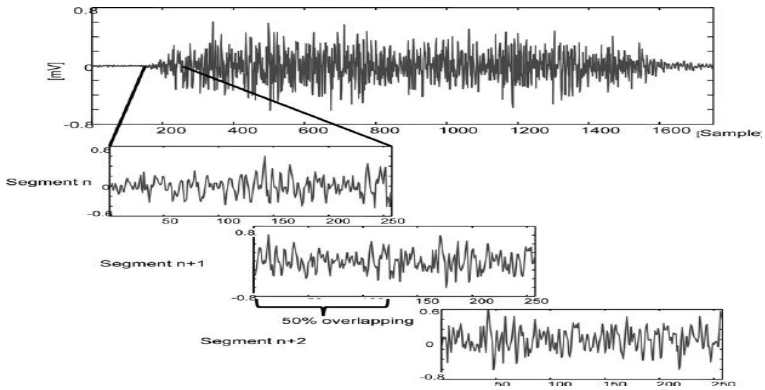
Online processing technique

This method was proposed in the main reference paper [1].
During the steady state of a musical note, the harmonic signal component can be well modelled as a sum of sinusoids, whose amplitude and frequency are slowly evovling in time. Absolute values of the frame-to-frame differences in the sinusoidal peak amplitudes and frequencies are quite low for steady state. Amplitudes of detected sinusoidal partials increases during attack region. These large amplitudes in the partials are inidicators of an Onset Location.

Data Preprocessing

- The signal is converted to overlapping, fixed-sized frames of audio, each having 4 buffers - 512 x 4 in duration.
- We used a 50% overlap for the frames and a samples.
- It consists of the most recent audio buffer which is passed directly to the algorithm, combined with the previous three buffers which are saved in memory.

- The time taken by the algorithm to process one frame of audio must be less than the duration of audio that is held in each buffer. Sampling rate = 44kHz and buffer size = 512 samples. So, the algorithm must be able to process a frame in 11.6 ms or less when operating.



Results

We collected samples of music signals corresponding to a guitar, drums and a saxophone. Ans ran all the ODF algorithms on these samples. As shown below, you can see the results corresponsing to the energy_ODF, LPSD and the peak-amplitude difference ODFs:

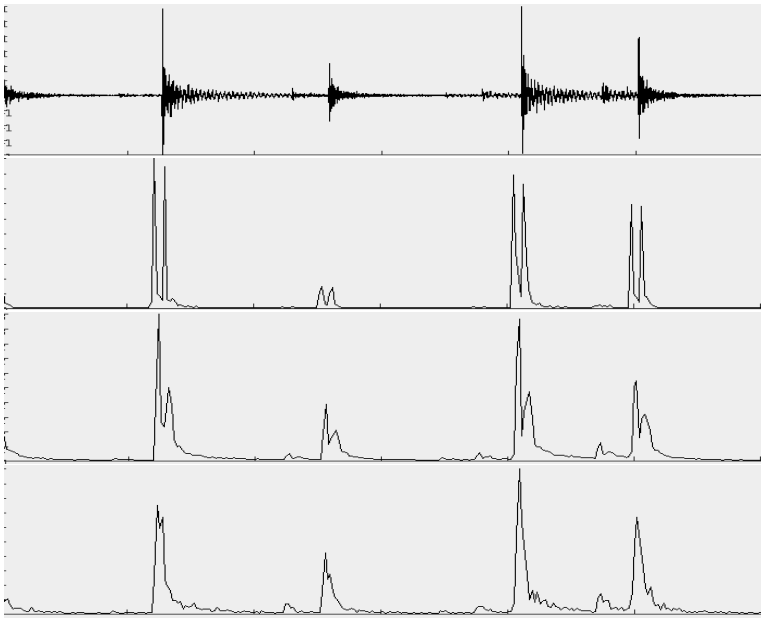


Figure 2: **drums-signal, energy-odf, LPSD, peak-amp (top to bottom)**

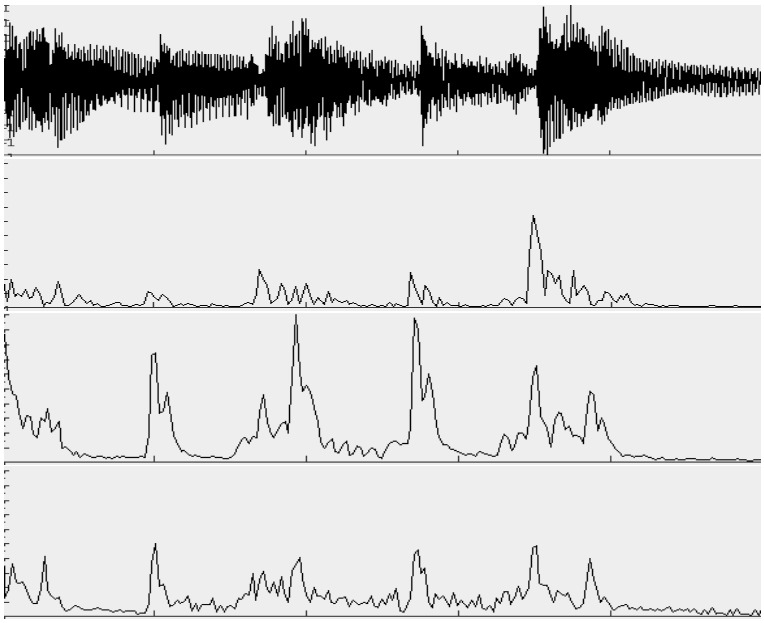


Figure 3: **guitar-signal, energy-odf, LPSD, peak-amp**

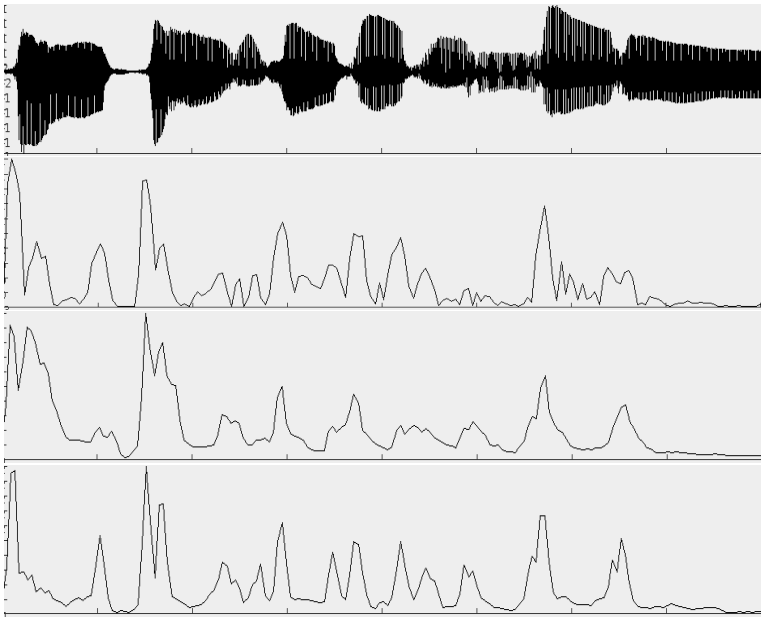


Figure 4: **saxophone-signal, energy-odf, LPSD, peak-amp**

Database results

The Modal library also provides a database of reference samples. The ground-references were obtained through crowd-sourcing and to be marked as 'correctly detected', the onset must be located within 50 ms of a reference onset. We have analysed the results in two ways:
1) Precision, recall and f-measure across all the methods
2) Precision, recall and f-measure v/s number of parameters in Linear prediction SD method

Precision, recall and f-measure across all the methods

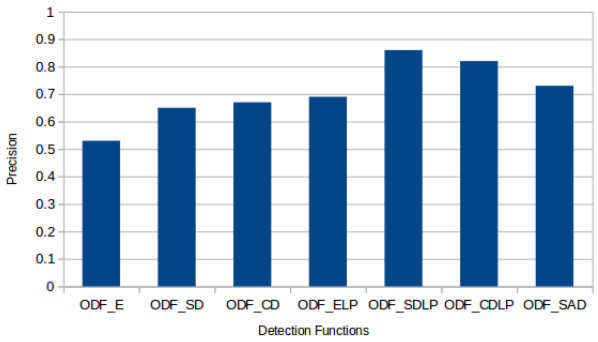


Figure 5: precision for each odf [1]

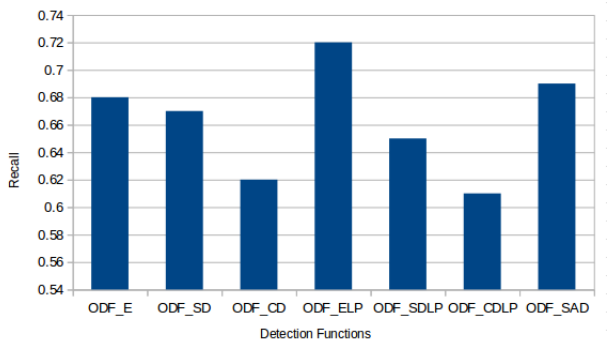


Figure 6: recall for each odf [1]

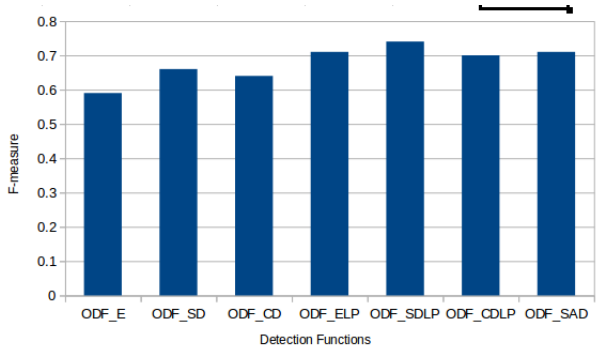


Figure 7: f-measure for each odf [1]

Precision, recall and f-measure v/s number of parameters in LP

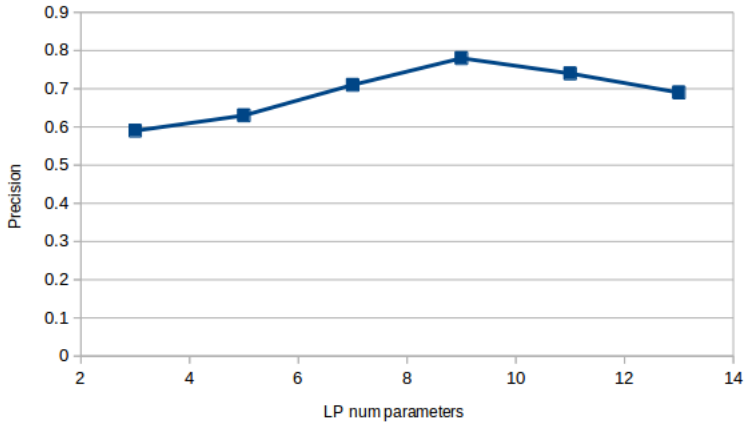


Figure 8: Precision v/s LP num parameters

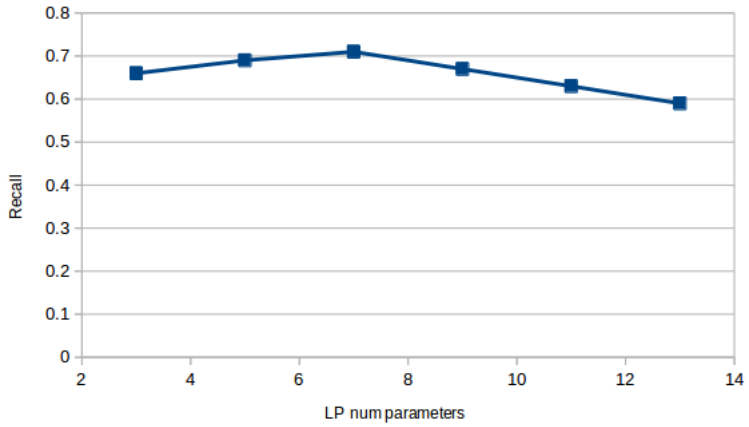


Figure 9: Recall v/s LP num parameters

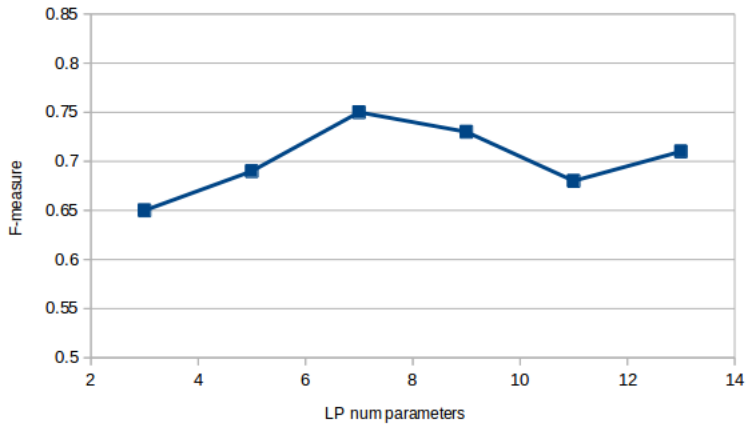


Figure 10: F-measure v/s LP num parameters

Conclusion

In this project, we discussed the traditional approaches as well as two new approaches, namely LP and sinusoidal modelling. It was noteworthy how the performance improved in the peak-amplitude difference method compared to all the methods. It should also be noted from the results that using too many parameters for LP does not necessarily guarantee better results as the f-measure and precision peak at num_parameters = 7. It was shown to provide more accurate results than the well-established complex domain method with noticeably lower computation requirements.

Bibliography

[1] John Glover, Victor Lazzarini, Joseph Timoney - Real-time detection of musical onsets with linear prediction and sinusoidal modeling, EURASIP '11

[2] Robert McAulay, Thomas Quatieri - Speech Analysis/Synthesis based on sinusoidal representation, IEEE Transactions '86

[3] Sinusoidal Modelling - http://www.music.mcgill.ca/ ich/classes/dafx book.pdf

[4] Juan Pablo Bello, Laurent Daudet et al - A Tutorial on Onset Detection in Music Signals, IEEE Transactions '05