

Onset Detection

An interesting approach to Query by Humming systems

Soham Bonnerjee (BS1609)

Subhrajyoty Roy (BS1613)

Ritwik Bhaduri (BS1616)

April 30, 2019

B.Stat., Indian Statistical Institute, Kolkata

Table of contents

1. Introduction
2. Building the Database
3. Onset Detection Methods
4. Power Calculation
5. Searching the database
6. Experimental Results
7. Conclusion

Introduction

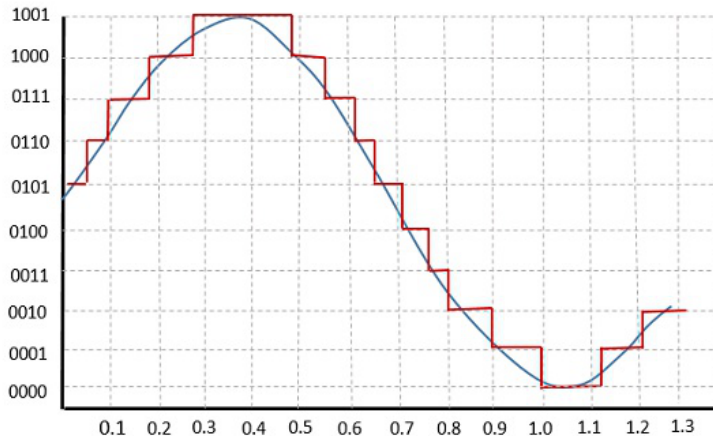
1. **THE STORY:** Where you might need it?

1. **THE STORY:** Where you might need it?
2. **3 Stage Procedure:**
 - 2.1 Building the database of songs.
 - 2.2 Detection of onsets (beats in layman's terms) of hummed song.
 - 2.3 Return the song which matches the most.

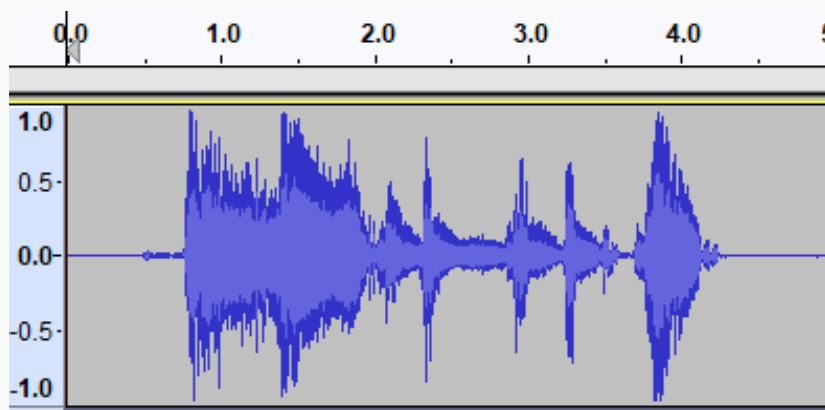
How is Digital Audio stored?

- **Sampling:** The binning in x-axis. Generally 44100 samples per second.
- **Quantization:** The binning in y-axis. Generally, ranges from $-(2^{15} - 1)$ to $+(2^{15} - 1)$ for 16 bit machine.

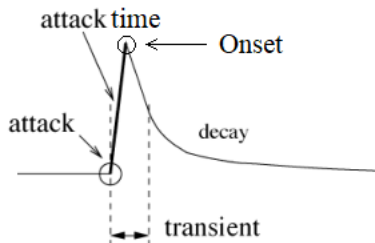
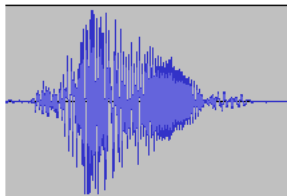
How does Digital Audio look like? (Toy Example)



How does Digital Audio look like? (Real Example)



What is Onset?



Discrete Fourier Transform (DFT)

- Let digital audio is denoted by a finite sequence or equivalently a vector of numbers, $\mathbf{x} = x[0], x[1], \dots, x[N-1]$, where N is the number of samples taken in the whole audio.
- Define the inner product: $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{k=0}^{N-1} x[k] \bar{y}[k]$
- The pure digital tones of order N of frequency f is the following complex N length vector;

$$\mathbf{v}_f = \frac{1}{\sqrt{N}} \left(1, e^{2\pi i f/N}, e^{4\pi i f/N}, e^{6\pi i f/N}, \dots, e^{2\pi i f(N-1)/N} \right)$$

- DFT of \mathbf{x} is \mathbf{y} , where;

$$y[k] = \langle \mathbf{x}, \mathbf{v}_k \rangle = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x[n] e^{-2\pi i f n/N}$$

for $k = 0, 1, \dots, (N-1)$.

The STDFT of a signal \mathbf{x} at a position n is given by;

$$X_k(n) = \sum_{m=-\omega/2}^{\omega/2} x[n+m]e^{-2\pi i k m / \omega}$$

where ω is the window length. This $X_k(n)$ is called Short Time Fourier coefficient of discrete frequency k at position n , representing the frequency content of frequency $k/\omega \times S_f$ during t_n to $t_{n+\omega}$.

Building the Database

1. For our experiment, we use a database of 10 songs.
2. For each song, the onsets are determined from the instrumental notations.
3. A typical element of database is a vector of size 20 to 35, containing the onset times of first verse of the song.

10 Chosen songs in Database

1. Jana Gana Mana.
2. Ekla Cholo Re.
3. Ore Grihobasi.
4. Sa Re Jahan Se Accha.
5. Esho Shyamolo Shundoro.
6. Jingle Bells.
7. My Heart will go on.
8. Fur Elise.
9. Hain Apna Dil.
10. Jindegi Ek Safar.

How the vector of onsets are created?

For example, consider *sargam* of *Ekla Cholo Re*

|NaNaNa|Sa—Sa|ReRe—|Pa—|Ma—|Ga—|ReSa—|ReGaGa|ReSa—|

gets converted to the vector;

|1, 2, 3, |4, 6, |7, 8, |10, |13, |16, |19, 20, |22, 23, 24, |25, 26, |

Onset Detection Methods

3 Stage Procedure

1. **Preprocessing:** It allows the raw vector of signal to be transformed slightly in order to improve the performance of the subsequent analysis. This step is optional and highly depends on the type of signal you are analyzing.

3 Stage Procedure

1. **Preprocessing:** It allows the raw vector of signal to be transformed slightly in order to improve the performance of the subsequent analysis. This step is optional and highly depends on the type of signal you are analyzing.
2. **Detection Function:** A detection function is a statistic which sufficiently reduce the data in a more compact form keeping the necessary information about the presence or the strength of the signal in a local neighbourhood. A detection function is applied to the signal through a sliding window (or moving window), allowing only the neighbourhood signal to be summarized.

3 Stage Procedure (Contd.)

1. For example, let ω is the size of the sliding window, and let $T(\cdot)$ be the statistic to be applied. Then, this detection function computes the value of the statistic for each of the moving window;

$$T[n] = T(x[n], x[n+1], \dots, x[n+\omega-1]) \quad \forall n = 0, 1, 2, \dots, (N-\omega)$$

The statistic $T(\cdot)$ is chosen in a way so that the onset at time t_n results in a comparatively higher value of $T[n]$ rather than its other values.

3 Stage Procedure (Contd.)

1. For example, let ω is the size of the sliding window, and let $T(\cdot)$ be the statistic to be applied. Then, this detection function computes the value of the statistic for each of the moving window;

$$T[n] = T(x[n], x[n+1], \dots, x[n+\omega-1]) \quad \forall n = 0, 1, 2, \dots, (N-\omega)$$

The statistic $T(\cdot)$ is chosen in a way so that the onset at time t_n results in a comparatively higher value of $T[n]$ rather than its other values.

2. **Peak Detection:** The resulting detection function is desired to produce local maximums at the time of the true onsets. Therefore, a peak detection algorithm is run at the end to identify the peaks in the vector of detection function. The times corresponding to these peaks are finally identified as possible onsets.

We consider the detection of presence of a signal in the n^{th} time-point.

$$H_0 : x[n] = w[n] \quad n = 0, 1, \dots, \omega - 1$$

$$H_1 : x[n] = s[n] + w[n] \quad n = 0, 1, \dots, \omega - 1$$

where $s[n]$ is deterministic and *completely unknown*, and $w[n]$ is WGN (White Gaussian Noise) with variance σ^2 .

Energy Detector (Contd.)

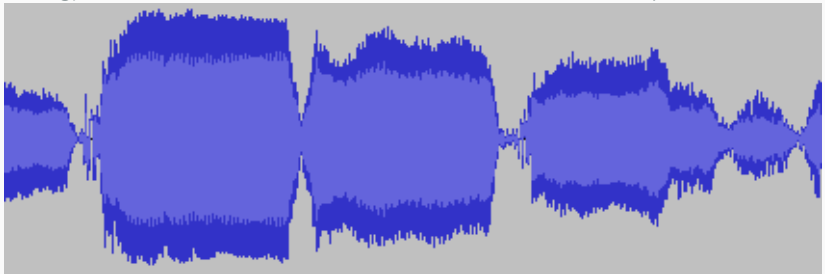
Using Likelihood ratio test, we reject H_0 in favour of H_1 if

$$T(\mathbf{x}) = \sum_{n=0}^{\omega-1} x^2[n] > \gamma' \quad (1)$$

This detector computes the *local energy* in the received data and compares it to a threshold. Hence this statistic is known as *Energy Detector*.

Problem with Energy Detector

Energy Detector is unable to detect immediate changes in local energy. Hence, unable to detect onsets if *Meends*¹ are present.



¹Meends are sliding from one note to another

$$T(n) = \sum_k (|X_k(n)| - |X_k(n-1)|) \mathbf{1}_{\{|X_k(n)| > |X_k(n-1)|\}}$$

1. Detect immediate changes in whole range of frequency spectrum.
2. Focus on onsets rather than offsets.

Dominant Spectral Dissimilarity

$$H_0 : x[n] = w[n] \quad n = 0, 1, \dots, \omega - 1$$

$$H_1 : x[n] = \begin{cases} w[n] & n = 0, \dots, n_0 - 1, n_0 + M, \dots, \omega - 1 \\ A \cos(2\pi f_0 n + \phi) + w[n] & n = n_0, n_0 + 1, \dots, n_0 + M - 1 \end{cases}$$

where $w[n]$ is WGN with known variance σ^2 , and A, f_0, ϕ are *Amplitude, Frequency, and Phase* respectively. These parameters might be unknown.

Dominant Spectral Dissimilarity (Contd.)

Likelihood Ratio test reject H_0 in favour of H_1 if;

$$T(x) = \frac{4}{\omega} \max_{f_0} \frac{1}{\omega} \left| \sum_{n=0}^{\omega-1} x[n] \exp(-i2\pi f_0 n) \right|^2 > \gamma$$

equivalently, if the maximum of the magnitudes of the DFT frequency contents is large enough.

Combining this, and idea of detecting instantaneous changes;

$$T[n] = \left(\max_k |X_k(n)|^2 - \max_k |X_k(n-1)|^2 \right) \times \mathbf{1}_{\{\max_k |X_k(n)| > \max_k |X_k(n-1)|\}}$$

Peak Detection: Features of a peak

1. Due to randomness in noise, it might happen that the detection function shows a peak at a location where there is only noise. Such peaks would be of relatively smaller height than a peak where onset has occurred. Therefore, we must choose a threshold parameter so that the peaks below that threshold parameter is completely ignored.
2. A peak should be of a higher value than its neighbouring values of detection function. Therefore, we must compare its value of detection function with that of its predetermined neighbours.
3. Two peaks should not be too close to each other. Two consecutive sounds must be at least $1/10$ -th of a second apart in time to be heard as distinguished sounds. Therefore, a reasonable peak detection algorithm should merge two onsets into a single one if they are less than 0.1 second apart.

Power Calculation

Basic Power Calculation

$$x[k] \sim N(0, \sigma^2)$$

$$k = 0, 1, \dots, k^* - 1$$

$$x[k] \sim N \left(A e^{-\lambda \frac{k - k^*}{S_f}} \cos \left(2\pi f_0 \frac{k - k^*}{S_f} \right), \sigma^2 \right)$$

$$k = k^*, k^* + 1, \dots, N - 1$$

Two main Results

Result 1

$$P(\hat{k} \text{ is an outputted onset}) \\ \geq \min \left\{ 0, \sum_{n=(-r)}^r P(T[\hat{k}] > T[\hat{k} + nh]) + P(T[\hat{k}] > \alpha) - 2r \right\}$$

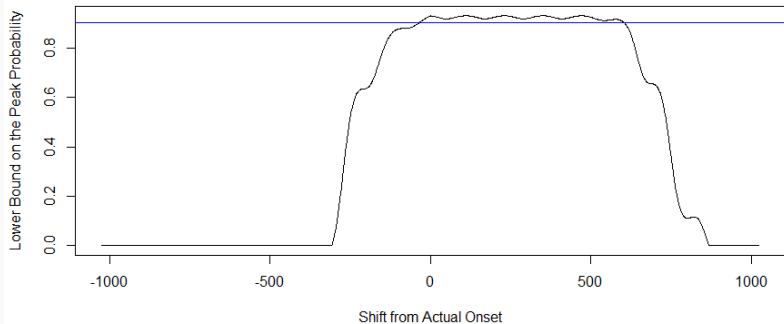
Helps to say with confidence, a true onset is detected.

Result 2

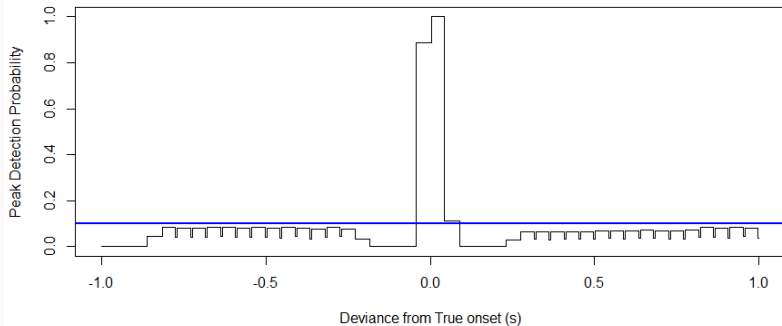
$$P(\hat{k} \text{ is an outputted onset}) < \frac{1}{2} \left(P(T[\hat{k}] > \alpha)(2 - P(T[\hat{k}] > \alpha)) \right)$$

when k is at least ω samples away from k^* . Helps to bound false positive probability.

Power of Energy Detector

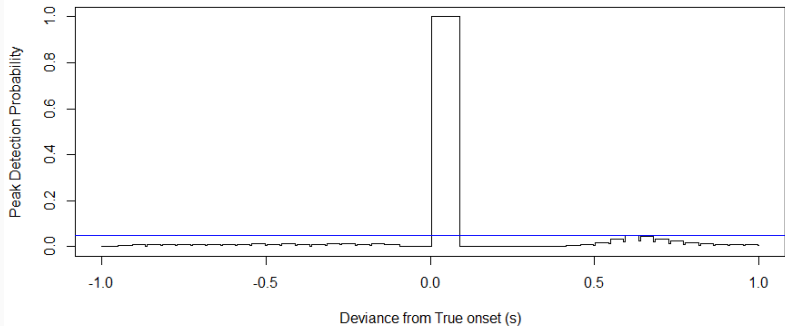


Power of Spectral Dissimilarity



Blue line is $y = 0.1$.

Power of Dominant SD

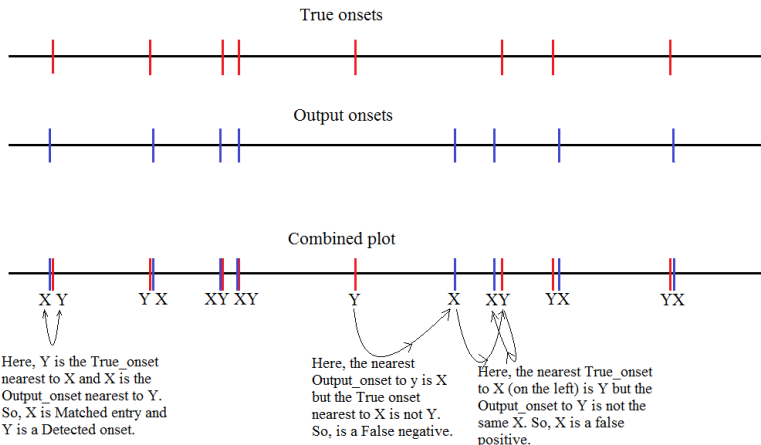


Blue line is $y = 0.05$.

Searching the database

Subset Matching

Assumption: Both the songs are assumed to be in the same unit with respect to time. It means that the same verse is present in both the songs and the time taken to sing both the parts is equal.



Correlative Matching

1. **Assumption:** First and last true onsets are detected.
2. Find all possible matches of First and last true onsets.
3. For each matchings, compute the linear transformation.
4. Perform subset matching on transformed onsets.
5. Compute score for each transformation:

$$\text{Score} = \text{Cor}(\text{matched_output}, \text{matched_transformed_true_onset}) \\ \times \left(1 - \frac{\#False_positive}{\text{output_length}}\right) \times \left(1 - \frac{\#False_negative}{\text{true_onset_length}}\right)$$

Experimental Results

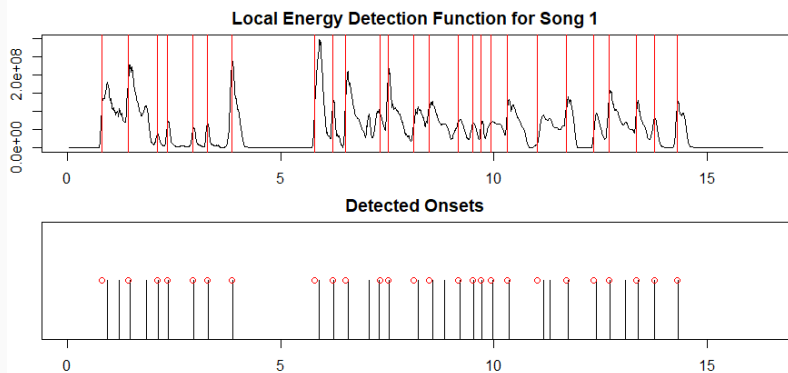
1. 3 hummed song.
 - 1.1 *Sa Re Jahan Se Accha*; for finding the optimal hyperparameter tuning.
 - 1.2 *Ekla Cholo Re*; to study effect of meends.
 - 1.3 *Jingle Bells*; to study effect of changing pitches.
2. 10 songs in database.

Optimal Hyperparameter Setup

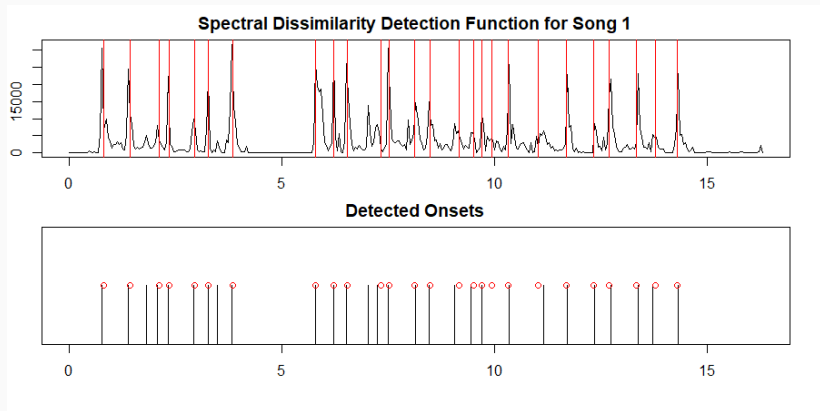
1. Local Energy Detector, window length of 4096 samples, hopsize of 512 samples. For peak detection procedure, it computes a peak with its 8 neighbouring points to both sides, and use a mean based thresholding criterion.
2. Spectral Dissimilarity Detector, window length of 4096 samples, hopsize of 2048 samples. For peak detection procedure, it computes a peak with its 4 neighbouring points to both sides, and use a mean based thresholding criterion.
3. Dominant Spectral Dissimilarity Detector, window length of 4096 samples, hopsize of 2048 samples. For peak detection procedure, it computes a peak with its 2 neighbouring points to both sides, and use a mean based thresholding criterion.

Song 1: Sa Re Jahan Se Accha

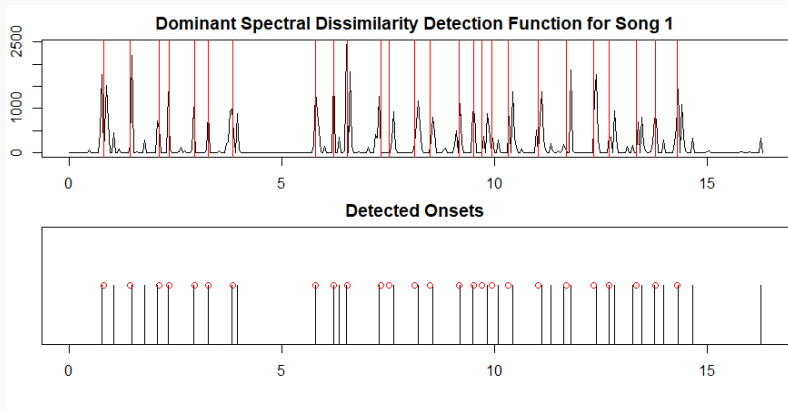
Performance of Local Energy Detector



Performance of Spectral Dissimilarity



Performance of Dominant Spectral Dissimilarity



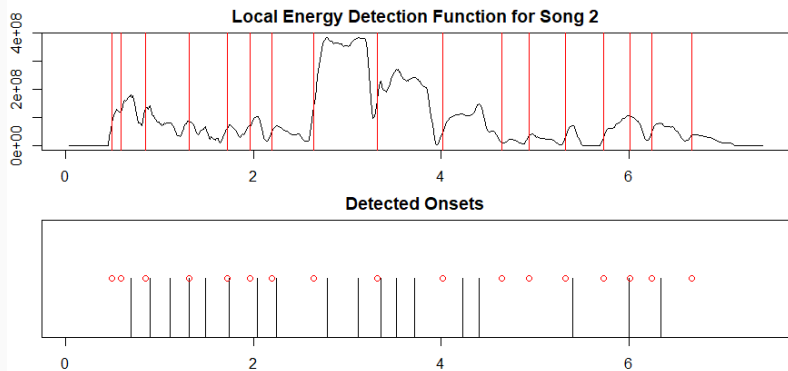
Performance of Searching Algorithm

Table 1: Details of Searching Output using hummed version of *Sa Re Jahan Se Accha*

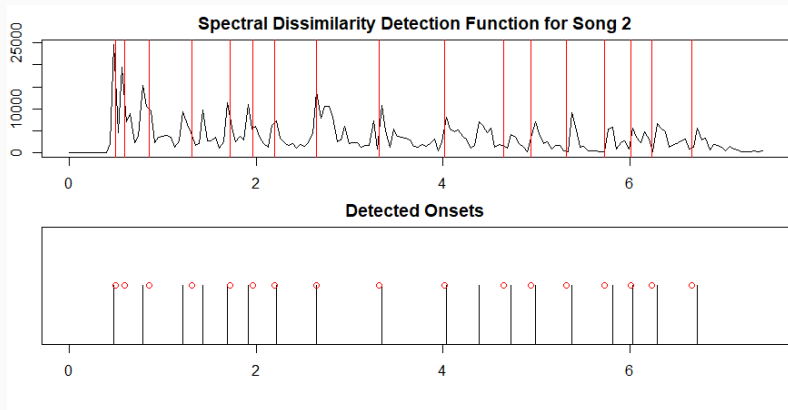
Ranks	Energy Detector		Spectral Dissimilarity		Dominant SD	
	Song	Score	Song	Score	Song	Score
1	Sa Re Jahan Se Accha	0.75	Sa Re Jahan Se Accha	0.791	Sa Re Jahan Se Accha	0.649
2	Jingle Bells	0.573	Jingle Bells	0.595	Fur Elise	0.623
3	Jana Gana Mana	0.568	Jana Gana Mana	0.585	Jingle Bells	0.544
4	Hain Apna Dil	0.53	Jindegi Ek Safar	0.578	Jana Gana Mana	0.541
5	Jindegi Ek Safar	0.506	My Heart will go on	0.578	Ore Grihobasi	0.493

Song 2: Ekla Cholo Re

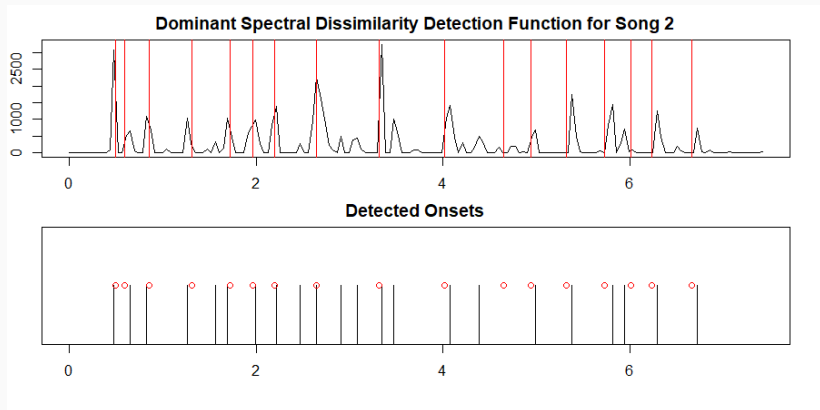
Performance of Local Energy Detector



Performance of Spectral Dissimilarity



Performance of Dominant Spectral Dissimilarity



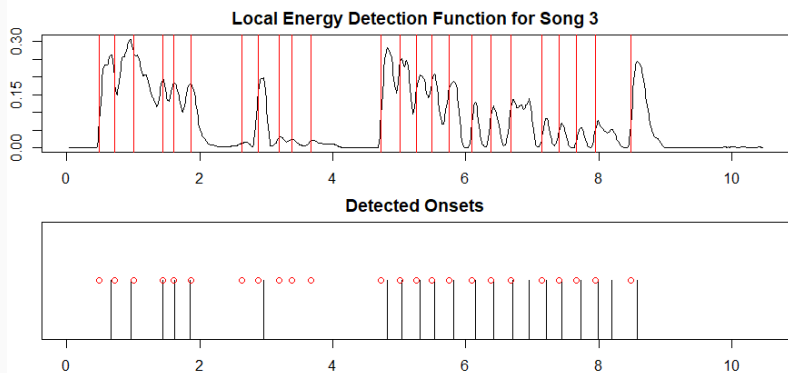
Performance of Searching Algorithm

Table 2: Details of Searching Output using hummed version of *Ekla Cholo Re*

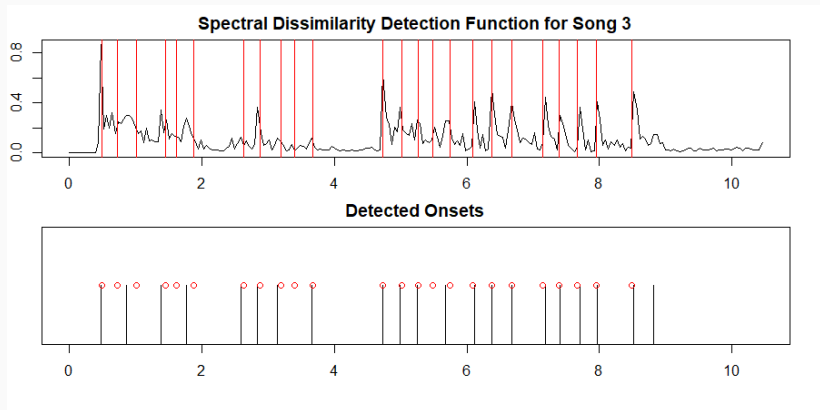
Ranks	Energy Detector		Spectral Dissimilarity		Dominant SD	
	Song	Score	Song	Score	Song	Score
1	Ekla Cholo Re	0.64	Ekla Cholo Re	0.735	Ekla Cholo Re	0.684
2	Hain Apna Dil	0.552	Hain Apna Dil	0.551	Jana Gana Mana	0.668
3	Jana Gana Mana	0.522	Jana Gana Mana	0.509	My Heart will go on	0.657

Song 3: Jingle Bells

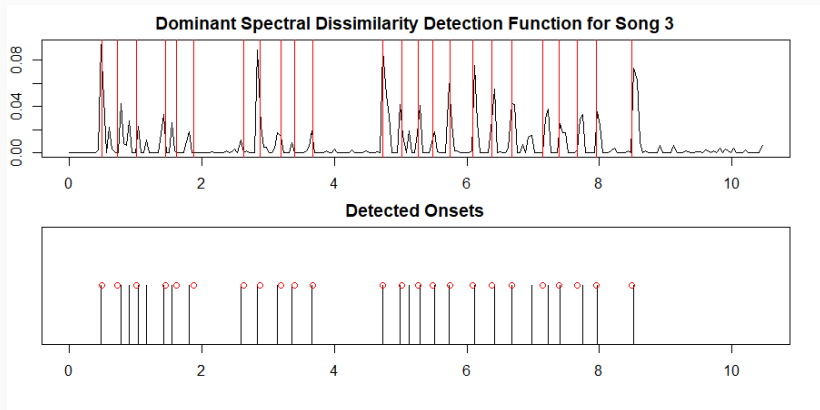
Performance of Local Energy Detector



Performance of Spectral Dissimilarity



Performance of Dominant Spectral Dissimilarity



Performance of Searching Algorithm

Table 3: Details of Searching Output using hummed version of *Jingle Bells*

Ranks	Energy Detector		Spectral Dissimilarity		Dominant SD	
	Song	Score	Song	Score	Song	Score
1	Jindegi Ek Safar	0.608	Hain Apna Dil	0.716	Jingle Bells	0.72
2	Hain Apna Dil	0.549	Jindegi Ek Safar	0.716	Jindegi Ek Safar	0.644
3	Ekla Cholo Re	0.548	Jingle Bells	0.687	Hain Apna Dil	0.607

Conclusion




1. Local Energy detector does not perform well under different note variants (or Alankars, e.g. Meend).
2. Spectral Dissimilarity detection suffers from the changes in the noise pattern and in pitches.
3. Dominant Spectral Dissimilarity is better than the previous two approaches, as it does not skip a true onset. Although, it suffers from the detection of more false positives than other approaches, the system as a whole works better with this.




Advantages over other QBH system

1. Most of the typical methods for QBH use the pitch information in the hummed song. Our method relies more on the rhythm of the song. So when some input is somewhat inharmonious or off-key most of the traditional systems fail but our method performs rather well.
2. Our method compares the onsets of the hummed song with that of the songs in the database. This eliminates the need of storing the entire songs in the database. We just need to store the onsets of the song in the database. This results in a huge reduction in the storage space.
3. Most QBH systems require a song to be sung by a handful of singers so that the algorithm can compare the hummed song with the different versions of the same song. This entails a lot of human effort to generate the database. Our approach doesn't require any human singing but the onsets of the actual song which is accessible from online sources.

Future Scopes

1. We perform the analysis assuming the fact that the noise are independent and identically distributed according to normal distribution with mean 0 and constant variance σ^2 . However, in practice, this assumption might not be true, as there might be different correlation structure between these errors.
2. It would be better if some tighter bounds on the probability of type I error and type II errors are available for the three detection algorithms.
3. Usage of dynamic time wrapping method rather than simple correlation to improve the searching technique.
4. Our algorithm demands the user to sing the first verse of the song. However, the point where the first verse ends may not be known to the user. In such case, searching procedure also should be able to match prefix of the song.

-  Steven M. Kay, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*. Prentice-Hall, 1998.
-  J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies and M. B. Sandler, *A tutorial on onset detection in music signals*, in IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, pp. 1035-1047, Sept. 2005.
-  I. Kauppinen, *Methods for detecting impulsive noise in speech and audio signals*, 2002 14th International Conference on Digital Signal Processing Proceedings. DSP 2002 (Cat. No.02TH8628), Santorini, Greece, 2002, pp. 967-970 vol.2.

-  A. Holzapfel, Y. Stylianou, A. C. Gedik and B. Bozkurt, *Three Dimensions of Pitched Instrument Onset Detection*, in IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 6, pp. 1517-1527, Aug. 2010.
-  Uniservity of Oslo: Sound Processing Lecture Part I: *Fourier analysis and applications to sound processing*.
<https://www.uio.no/studier/emner/matnat/math/nedlagte-emner/MAT-INF2360/v12/part1.pdf>.
-  Amos Lapidoth. *A Foundation in Digital Communication* ETH Zurich. Swiss Federal Institute of Technology. 2009.

THANK YOU

QUESTIONS?