Evaluation of Learner's SARGAM Practice

Report submitted to
Indian Institute of Technology Kharagpur
in partial fulfillment for the award of the degree of
Bachelor of Technology (in)
Computer Science and Engineering Department

By Manikya Singh
(14CS10032)

Under the supervision of
Prof. K. Sreenivasa Rao

Computer Science and Engineering Department
Indian Institute of Technology Kharagpur
Spring Semester, 2017-18
May 7, 2018

Contents

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR
KHARAGPUR - 721302, INDIA

CERTIFICATE

This is to certify that the project report entitled "Evaluation of Learner's SARGAM Practice" submitted by Manikya Singh (Roll No. 14CS10032) to Indian Institute of Technology Kharagpur towards partial fulfilment of requirements for the award of degree of Bachelor of Technology in Computer Science and Engineering Department is a record of bonafide work carried out by him under my supervision and guidance during Spring Semester, 2017-18.

Date: May 7, 2018

Kharagpur - 721302, India

Prof. K. Sreenivasa Rao
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur

Abstract

Name of the student: Manikya Singh
Roll No: 14CS10032
Degree for which submitted: Bachelor of Technology
Department: Computer Science and Engineering Department
Thesis title: Evaluation of Learner's SARGAM Practice
Thesis supervisor: Prof. K. Sreenivasa Rao
Month and year of thesis submission: May 7, 2018

In Hindustani music, a 'swar' or note is represents the pitch which is being sung or played by a learner. In this work, as automatic SARGAM practice evaluation method is proposed which can be used by beginner learners to practice their SARGAM and get feedback.
In Hindustani music, the Sargam starts with swar 'Sa' and every note relative to Sa is fixed. In the evaluation proposed, the first note is taken as Sa and deviation of every other note with respect to that is calculated. Method involves initially recording Sargam and then finding notes onset after converting music signal to spectral domain. To find error in a note, deviation of the fundamental frequency of note being sung from the ideal fundamental frequency of the note with respect to Sa is compared.

Acknowledgements

I take this opportunity to express my profound gratitude and deep regards to my guides Prof. K. Sreenivasa Rao for this exemplary guidance, monitoring and constant encouragement throughout the course of this work. The blessing, help and guidance given by them time to time shall carry me a long way in the journey of life on which I am about to embark. I wish to acknowledge the encouragement received for initiating my interest in this topic and also for his unparallelled help and motivation round the clock to carry out my project work. Lastly, I would like to thank all professors, lab administrators and friends for their help, moral support and wishes.

Introduction

In music, a note is the pitch and duration of a sound, and also represent a pitch class. In general, notes are the building blocks of music. Twelve notes namely SA re RE ga GA ma MA PA dha DHA ni Ni represent different pitch sounds. Each note has a fundamental frequency, higher harmonics of the note however are also present when someone sings or plays a note. Set of these twelve notes make an octave. Same notes of different octaves are different harmonics of same fundamental frequency.

SARGAM is a collection of musical notes or the swars of a scale. Frequency of each note relative to SA is fixed. The cent is a logarithmic unit of measure used for musical intervals. Twelve-tone equal temperament divides the octave into 12 semitones of 100 cents each. A learner needs to keep practicing SARGAM until he is able to sing all notes correctly. This process involves both learner and teacher to be present. Hence, an automated method for Sargam evaluation which can be used by learner to practice Sargam is proposed.

The difference, in Cent, between two notes of frequencies a and b is given by:

$$n = 1200 * log\left(\frac{a}{b}\right)$$

For evaluation of learner's Sargam, after determining the fundamental frequency in Note region in the recording, the pitch is converted into musically relevant Cent scale. The block diagram of Sargam's evaluation is shown in Fig. 1.

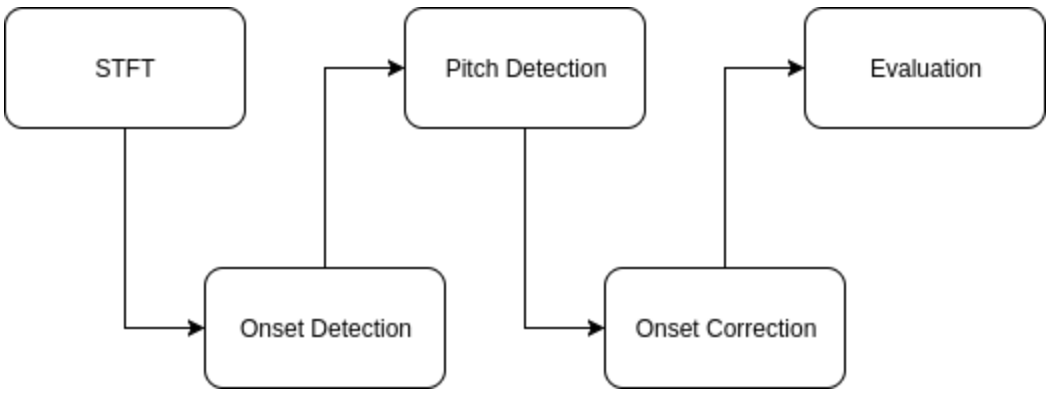

Fig. 1: Block diagram for Sargam's evaluation

Spectral Transformation

Spectrogram of Sargam sequence (digitized at 44.1KHz sampling rate) is obtained by applying STFT with frame with 40ms and frame shift 3ms.

$$X(l,k) = \sum_{n=0}^{N-1} x(n)w(n-l)e^{-j2\pi kn/N}$$

Fig. 2, 3 and 4 shows three sargam recording samples(referred as S1, S2 and S3 in subsequent sections) in and their respective spectrograms. The first audio sample is a noiseless recording of virtual piano and hence is closest to ideal inputs, the second and third are a learner's sargam recording but in the third one the notes sung are continuous without any gap in between them which causes many compilations discussed in each of the following sections.

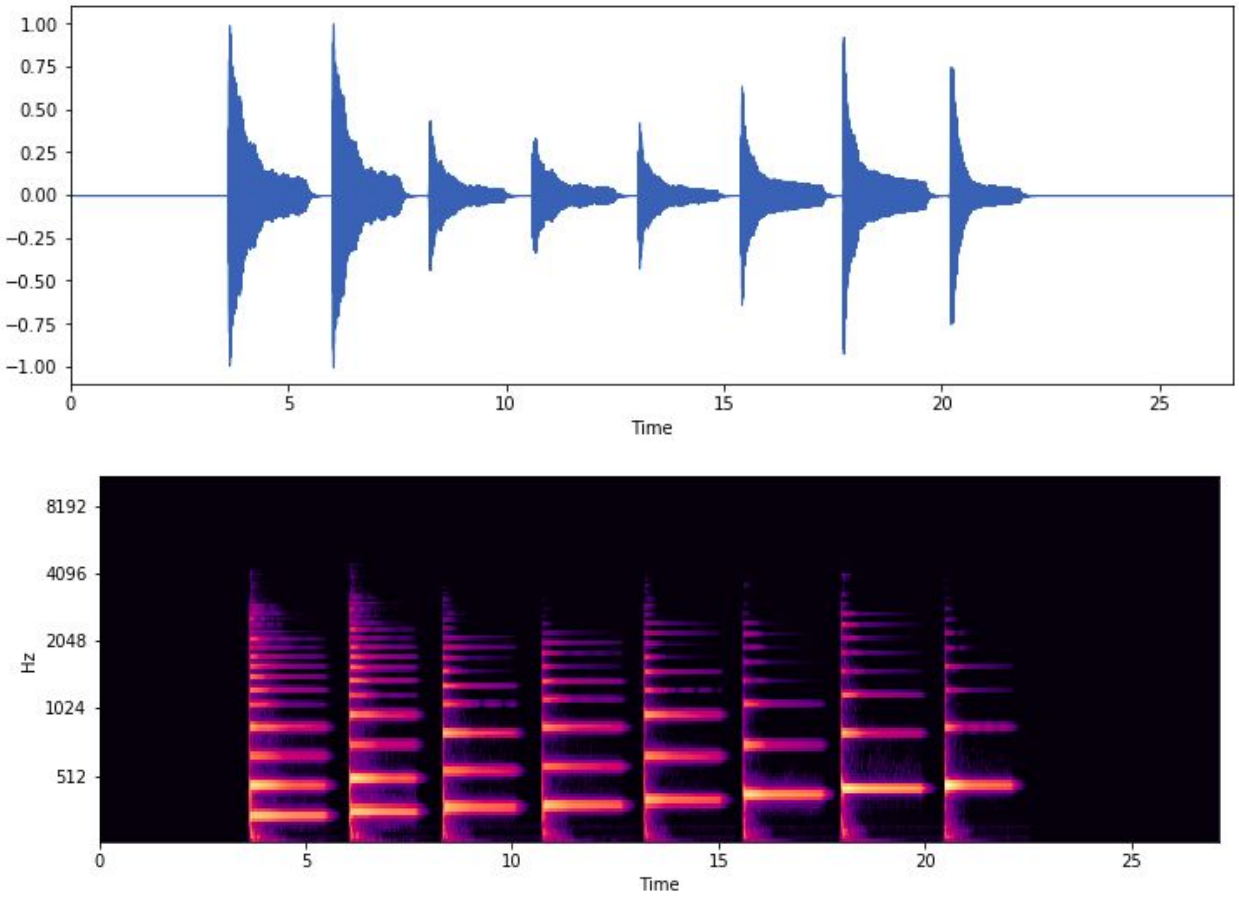

Fig. 2: noiseless recording of virtual piano and it's spectrogram

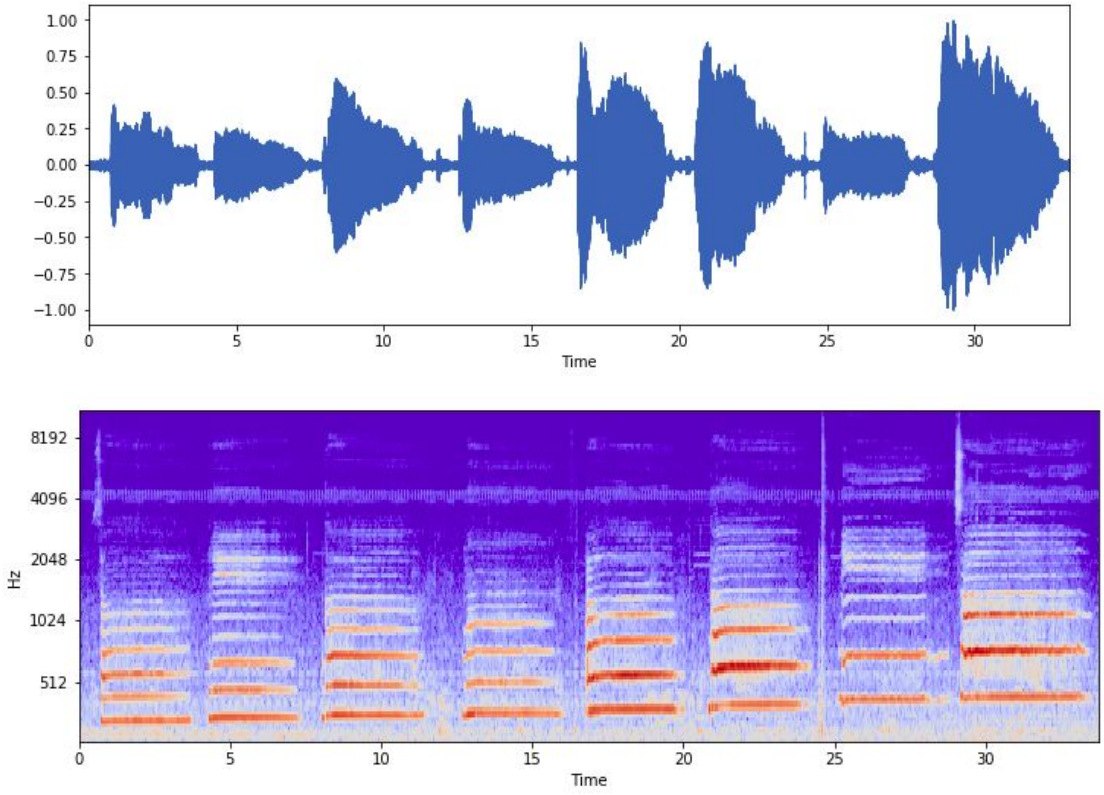

Fig. 3: recording of learner

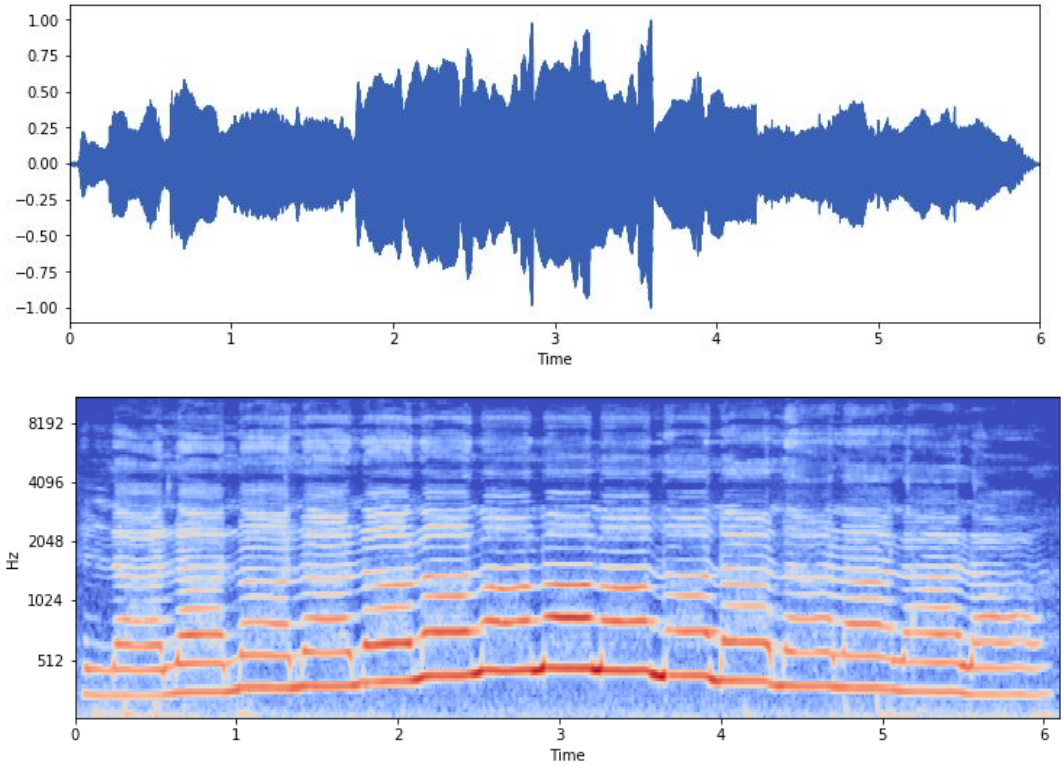

Fig. 4: Fast ascent and descent of Sargam of learner and it's spectrogram

Note Frequency

Fundamental frequency of a note can be determined by existing pitch detection algorithms.

Glottal Closure Instants Based

Glottal Closure instants (GCI) are found in each note using adaptive Zero Frequency Filtering[1][2]. The GCI locations are then zero frequency filtered with resonance frequency obtained by two way mismatch algorithm. The reciprocal of the time difference between the successive GCI locations is computed to obtain the frequency in Hertz which is converted into cent scale.

Spectral YIN

YIN-FFT[3] algorithm estimates the fundamental frequency given the spectrum of a monophonic music signal. It is an optimised version of YIN[4] algorithm which operates in time domain. This algorithm returns fundamental frequencies along with their confidence value. Fig. 7 shows median filtered, low-confidence clipped(frequencies corresponding to confidence values < 0.9 are clipped) Spectral YIN output for the sargams S1,S2 and S3.

For experiments and tests we have used frequencies obtained by YIN FFT algorithm.

Note Onset Detection

Detecting notes onset correctly are a crucial part for evaluating learner's Sargam practice as a spurious onset will lead to falsely identified note resulting in evaluation errors.

Note onsets are characterised by increase in spectral energy of the fundamental frequency of note being played and it's harmonics. Based on this approach we can detect notes onset by analysis of change in spectral energies between STFT frames. Another approach is based on confidence of pitch values.

Normalized spectral energy change detection

Normalized Euclidean distance[5] between spectral frames of the magnitude spectrogram of the Sargam sequence X(l,k) contains noisy regions which leads to multiple onsets is smoothed without blurring onset peaks by a sharp cutoff low pass filter to get the onset detection function.

$$E_x(l,k) = X_m(l,k) - X_m(l-1,k)$$

$$E_{df}(l) = \sum_{k;E_x(l,k)>0} E_x^2(l,k)$$

$$E_{ndf}(l) = E_{df}(l) / \sum X_m^2(l-1,k)$$

Low pass filtering in time domain is performed by exponentially weighted previous frames.

$$y(l) = E_{ndf}(l) - \sum_{a=1}^{A} E_{ndf}(l-a)/l$$

Onsets are detected by the following peak picking heuristics:

$$y(l) = max(y(l-w : l+w))$$
$$y(l) >- mean(y(l-w : l+w)) + \delta$$
$$l - l_{last\ onset} > w$$

Fig. 5 shows the onsets detected for the Sargams S1, S2 and S3.

Normalization of Euclid Distance function E_{df} by division of energy of previous frame results in relative peak enhancement at note onset as their is only recording noise in between notes contributing to low energy compared to frames which contains notes which lead to high energy resulting in spurious peak suppression which is desirable. This method however fails if notes are continuous.

It can be seen from Fig. 5 that the onset detection fails when the notes are continuous or their is no significant gap in time between onset of a note and offset of it's previous note. This error comes due to normalization as division by energy of preceding spectral frame results in significant suppression in onset peak if the previous frame has high energy which is the case if previous frame also contains a note.

Cosine Similarity Based

[6]Rather than calculating distances between successive frames, this method finds similarity between frames by dot product of spectral magnitude vectors.
The aim is to detect increase in energy of particular harmonics which are onset on a frame while ignoring decrease in energy associated with ending of previous notes. The onset detection function is given by:

$$y(l) = 1 - \widehat{X}(l-1).\widehat{X}(l)$$

Where $\widehat{X}$ represent unit vector of the frequency vector of a frame.

Figure 6 shows the onset detected by this function for recordings S1, S2 and S3. While this onset detection function works for perceivably continuous sequences of notes but fails for simple cases. This is due to the fact that that similarity between two random frames is random giving rise to a lot of spurious peaks.

Pitch Confidence Based

The YIN-FFT algorithm returns frequencies along with their confidence values. Fig. 7 shows median filtered, low-confidence clipped(frequencies corresponding to confidence values < 0.9 are clipped) Spectral YIN output for the sargams S1,S2 and S3.

For Sargams S1 and S2 where notes haves gap in between them, this method gives both cent scale values of notes being sung and their respective value per frame. All the non zero values correspond to a note region.

A similar approach could be done using errors of the frequency detected by two way mismatch algorithm for pitch detection. A note region would correspond to a minima region of the error function.

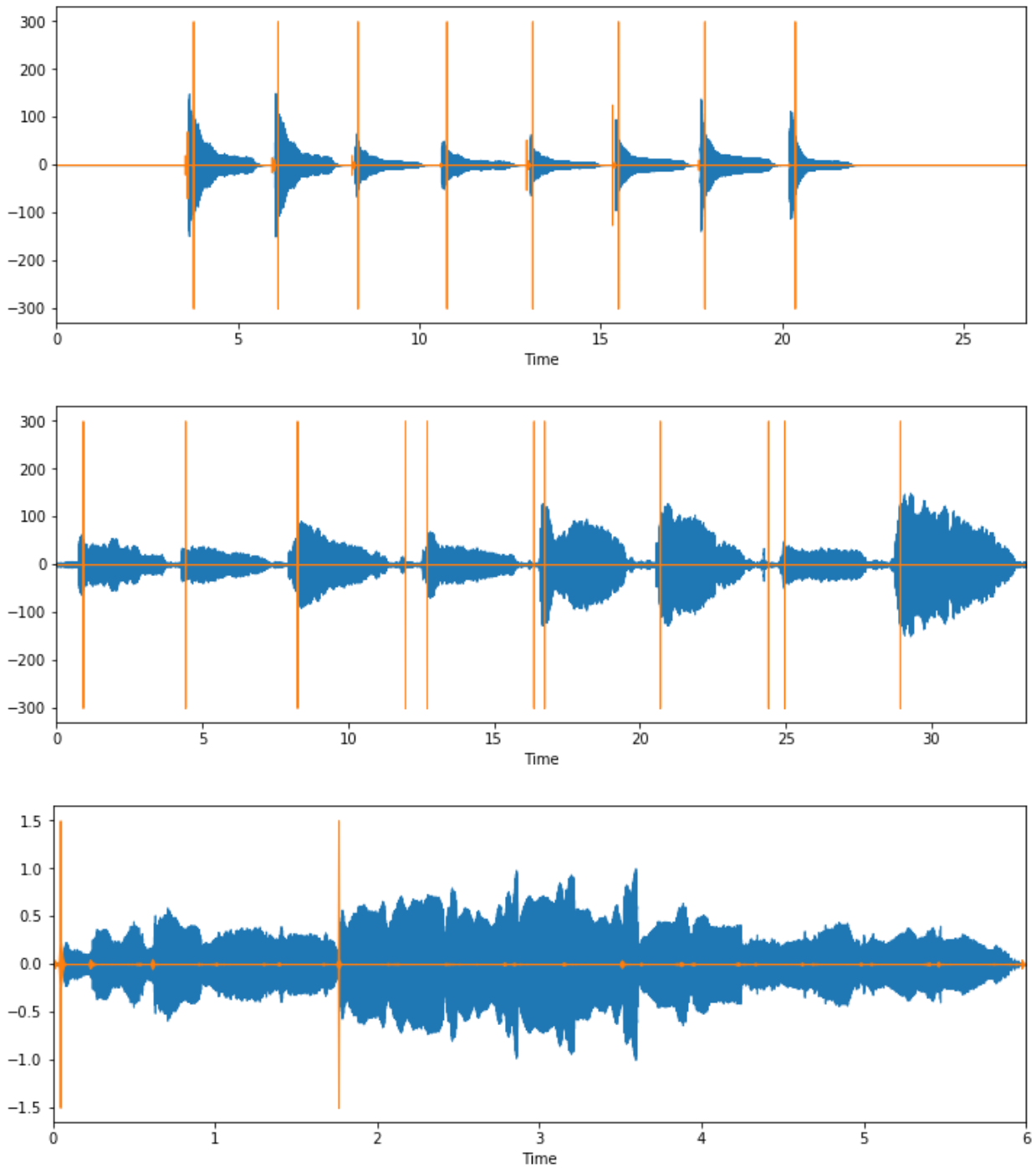

Fig. 5: Onsets detected by normalized spectral distance detection function for the three sargam recordings

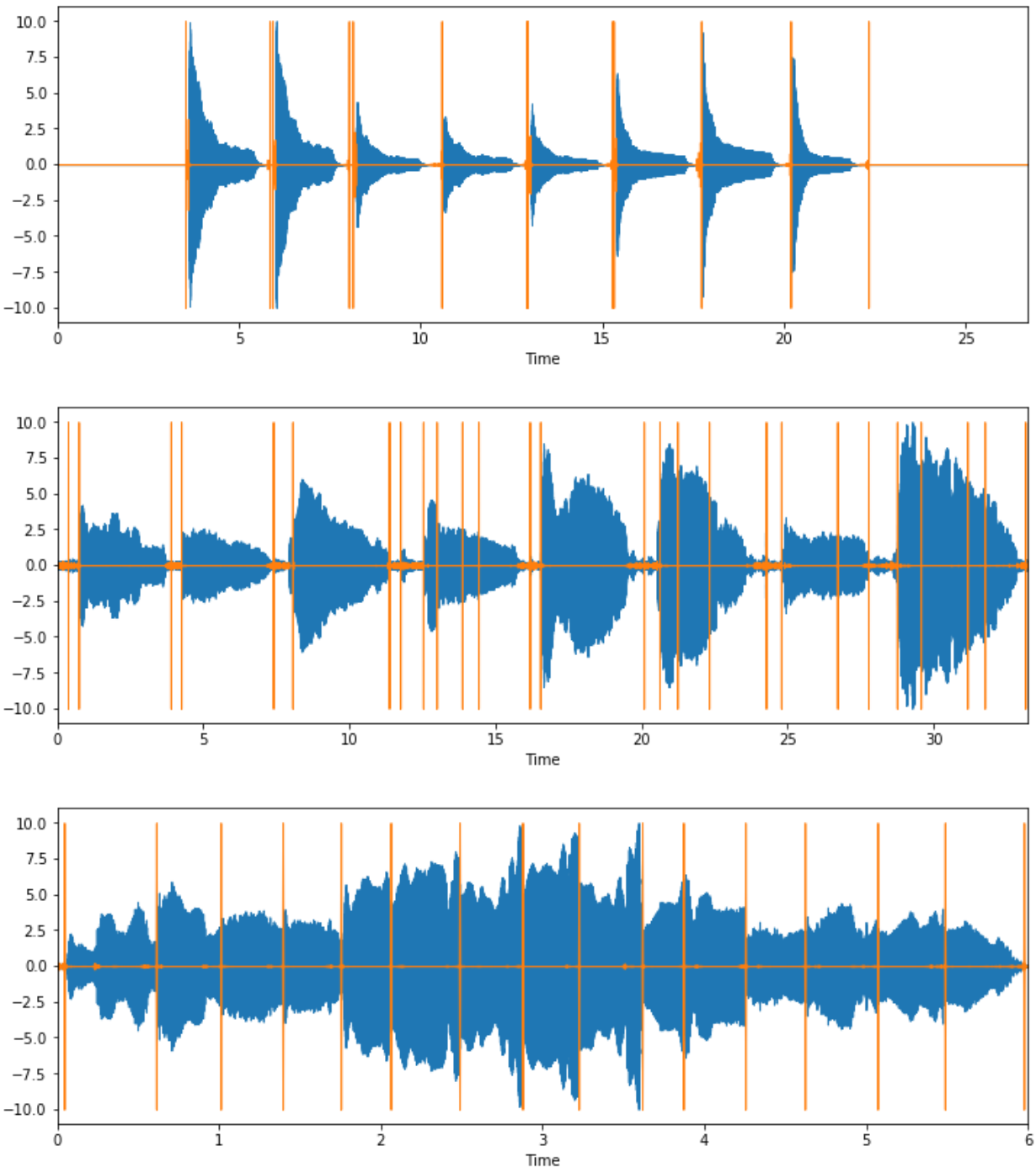

Fig. 6: Onsets detected by Cosine Product based onset detection function

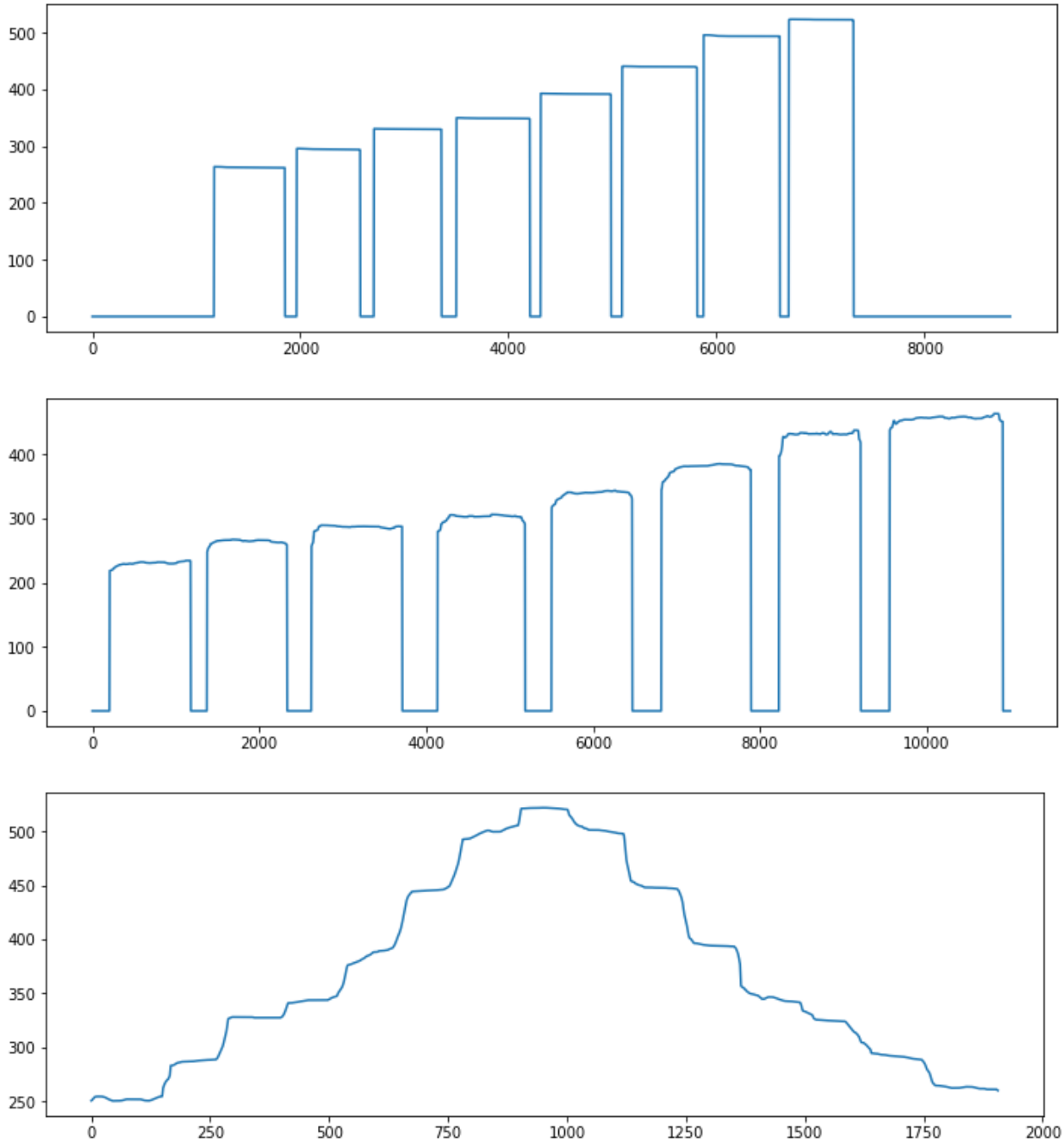

Fig. 7: Note regions detected by frequency confidence based onset detection

Onset Correction

Onset Marker Based

Spurious onsets may occur within note due to energy fluctuations of higher harmonics, or in non-note regions due to environmental factors.If an onset location is correct then the region before it is a non-note region - has high frequency variance and the region after it is note region - has low frequency variance. To detect spurious onsets in between two notes-

To detect spurious onsets within a note-
For each onset previously detected:

1. Standard deviation in frequency of following 100 frames is calculated (sd next)
2. Standard deviation in frequency of previous 100 frames is calculated (sd prev)
3. if(sd next < 80 and sd prev < 80) onset is within a note and is incorrect

To detect spurious onsets in between two notes-
For each onset previously detected:

1. Standard deviation in frequency of following 100 frames is calculated (sd next)
2. Standard deviation in frequency of previous 100 frames is calculated (sd prev)
3. if(sd next > 80 and sd prev > 80) onset is in between 2 notes and is incorrect

Stable note region is a region whose frequency is less than 80 cents. Each note is assigned a single frequency which is the median of note frequencies in the stable region of the note.

Pitch Confidence Based

To detect spurious notes interval, non-zero intervals corresponding to a note are thresholded based on their length.

1. Note corresponding to each frame and corresponding error having non-zero frequency(after median filtering) is calculated.
   Note Number = int(Frequency in Cent/100)
   //Note Number 0 corresponds to Sa , 1 to re, 2 to Re etc.
   Note error(in Cent) = Frequency in Cent % 100
2. Notes having length less than a threshold are eliminated.
3. Cent frequencies corresponding to same notes are averaged to get note pitch.

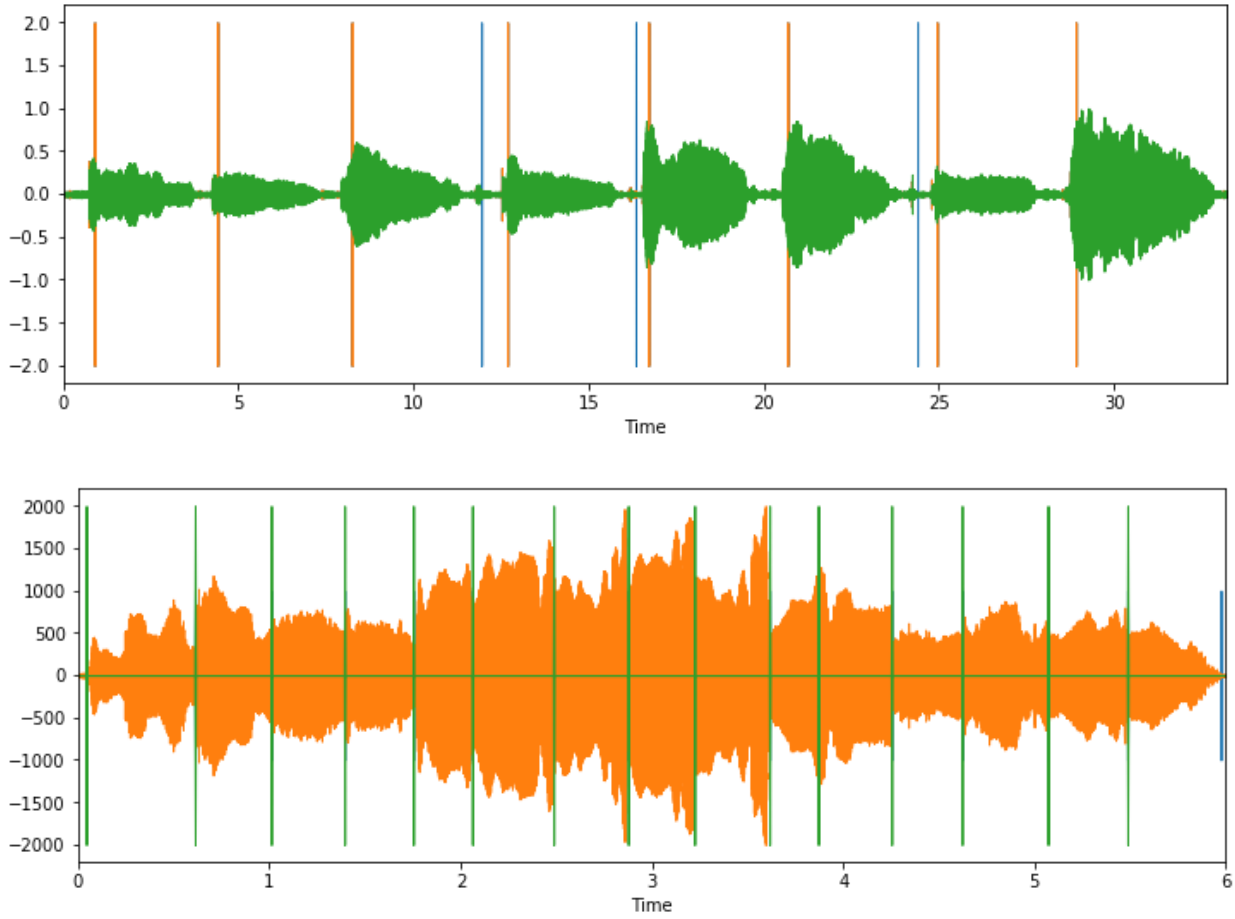

Fig. 8: Corrected Onsets for learner's Sargam

Evaluation

Since every note is fixed relative to Sa we assume first note of the Sargam sequence to be Sa. The frequencies of remaining notes are computed with respect to Sa.

$$F_{relative} = F - F_{Sa}$$

Where F = frequency of note in cent

F_{Sa} = frequency of Sa in cent = Frequency of first note

$F_{relative}$ = relative frequency of each note

Since we do not know which notes of the Sargam learner is going to practice we will predict the note to closest to the one he is singing.

Note = int(Note Frequency/100)

Where Note = 0 corresponds to Sa, 1 to Re etc…

Note Error = Note Frequency%100

Fig. 9: Evaluation result of Sargam S3(using Cosine Similarity Based Offset Detection)

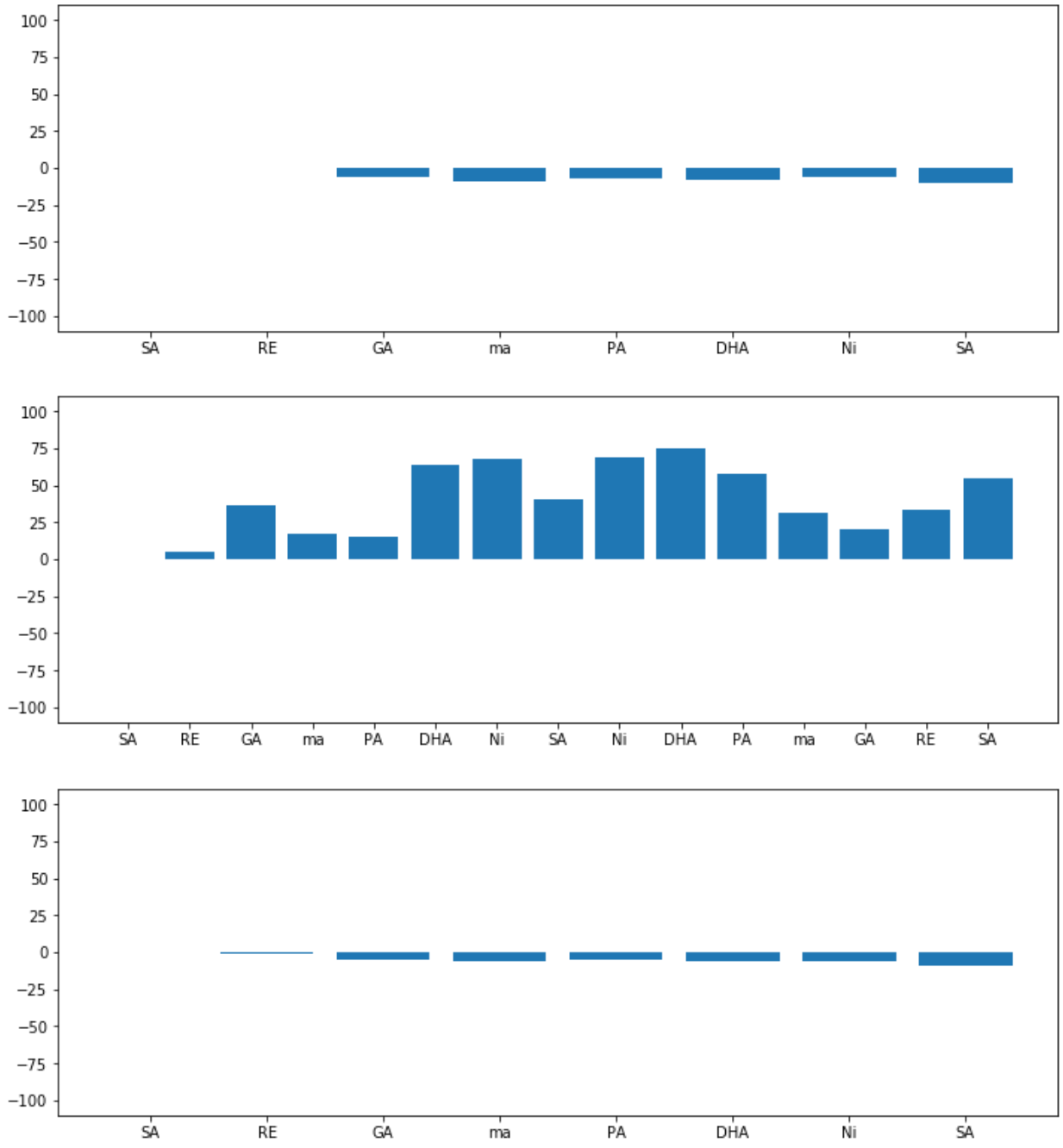

Fig. 10 Evaluation result of Sargam S1(1. Using Spectral Energy based onset detection, 2. Using frequency confidence based note detection)

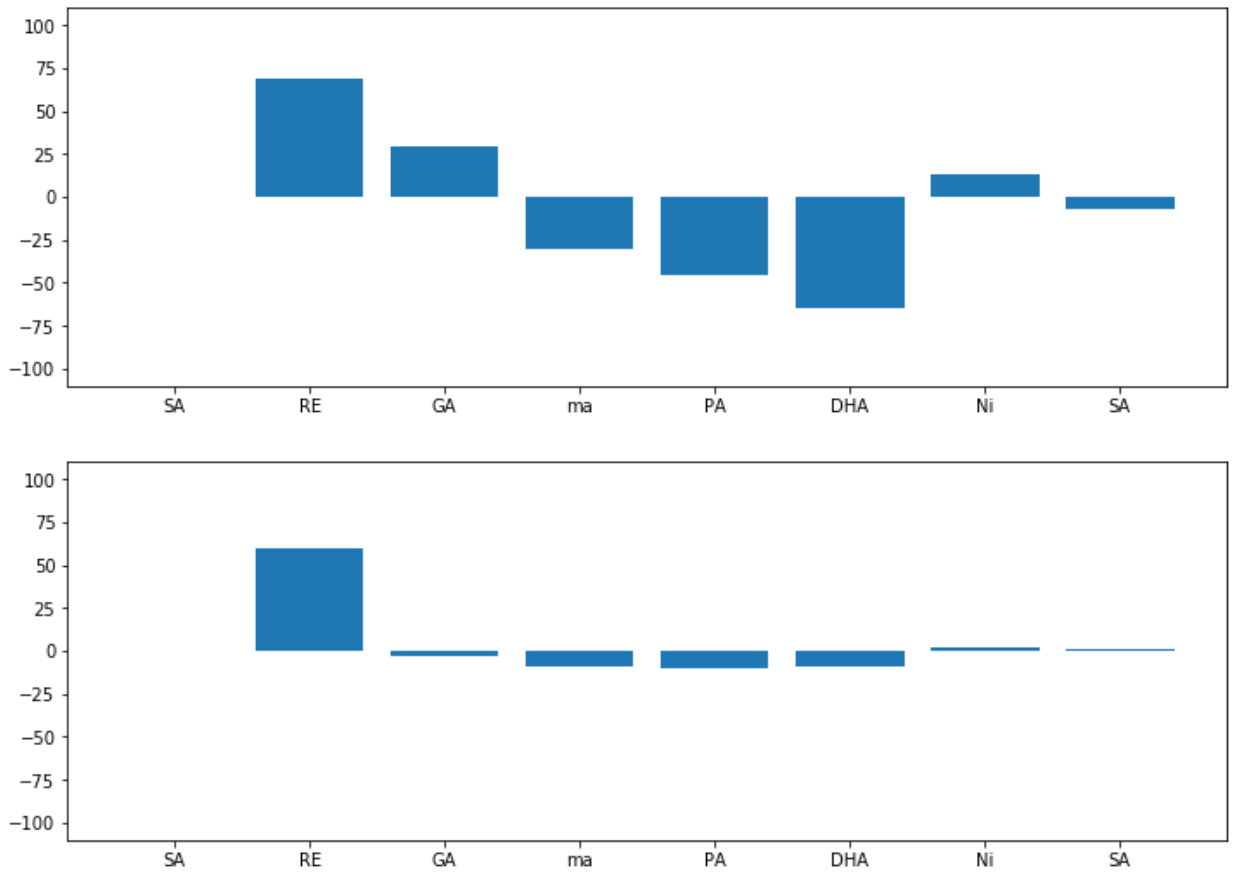

Fig. 11 Evaluation result of Sargam S2(1. Using Spectral Energy based onset detection, 2. Using frequency confidence based note detection)

References

[1] M. G. Reddy and K. S. Rao, “Predominant melody extraction from vocal polyphonic music signal by combined spectro-temporal method,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, 2016, pp. 455–459.

[2] G. Reddy and K. S. Rao, “Enhanced harmonic content and vocal note based predominant melody extraction from vocal polyphonic music signals,” Interspeech 2016, pp. 3309–3313, 2016.

[3] P. M. Brossier, "Automatic Annotation of Musical Audio for Interactive Applications,” QMUL, London, UK, 2007.

[4] Alain de Cheveign´e and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. Journal of the Acoustical Society of America, 111(4), 2002.

[5] C. Duxbury, M. Sandler, and M. Davies, “A hybrid approach to musical note onset detection,” in Proceedings of the International Conference on Digital Audio Effects (DAFX), 2002, pp. 33–38.

[6] Foote, J. and S. Uchihashi (2001). The beat spectrum: a new approach to rhythm analysis. In Proc. Int. Conf. on Multimedia and Expo (ICME).