Automatic Evaluation of Hindustani Learner's SARGAM Practice

Gurunath Reddy M and K. Sreenivasa Rao
Indian Institute of Technology, Kharagpur, India
{mgurunathreddy, ksrao}@sit.iitkgp.ernet.in

***Abstract*—In this paper, Hindustani learner's SARGAM practice evaluation method is proposed. A SARGAM is a collection of notes or the swars in Indian art music. In this work, an automatic SARGAM evaluation method is proposed to detect the notes which deviates from the predefined musical ratio rendered by the learner. The method involves initially recording the SARGAM sequence from the learner, followed by note boundary detection by finding their onsets in the spectral domain. The fundamental frequency in each note is obtained by finding the glottal closure instants of the vocal source. The note deviation is computed as the absolute musically relevant cent scale frequency deviation between the notes rendered by the learner and the ideal note frequencies. The correctness of the proposed method is evaluated by the time domain waveforms, spectrograms and objective evaluations.**

***Index Terms*—Onsets, SARGAM, Notes, ZFF, Hindustani Music, Note Frequency, Cent Scale.**

I. Introduction

The first step in the Hindustani classical music learning process is the SARGAM practice [1]. The SARGAMS's are the basic notes in the traditional Indian art music. In other words, SARGAM is a collection of musical notes or the swars of the scale such as Sa, Re, Ga, Ma, Pa, Dha, Ni as shown in Table I. The pure or natural notes are called as shudh notes [2] which are symbolically represented as SA, RE, GA, MA, PA, DHA, NI. The notes RE, GA, DHA, and NI can be either shudh (natural) or komal (flatter version) i.e., re, ga, dha, ni as shown in Table I [3]. The note Ma can be either shudh or tivra (sharp). The notes SA and PA are called as immovable notes (once SA is selected as base note). The SARGAM practice is also called as singing or vocal exercise in which the various combination of note sequences are sung in succession. The Guru (teacher) renders the SARGAM's in various combinations based on the reference note. The reference note can be any note such as SA, RE and so on. An example of the SARGAM sequence can be as simple as the sequence of notes SA, RE, GA, MA, PA, DHA, NI, SA' or SA', NI, DHA, PA, MA, GA, RE, SA. Where the later SARGAM started with higher octave SA' and descended to the lower octave SA. The Shishya (learner) is made to repeat the same sequence of notes until all notes follows a predefined ratio. Since the process of SARGAM practice involves both learner and the tutor, the presence of the tutor is mandatory to give the feedback of the sung SARGAM. Also, it involves repetition of the same sequence of SARGAM for many times, until rendered in correct sequence, which is a tedious process for both learner and the tutor. Hence, in this paper we propose a automated SARGAM learning method, which can be used by individual at any point of time to practice the SARGAM without the physical presence of the tutor. Also, we have developed a SARGAM practice application, which plays the pre-recorded SARGAM's from the teacher and the learner is asked to repeat the same sequence of notes until the deviation of the notes is within the tolerance range. The system flags the deviation indicator for those notes which are in out-of-tune. Thus, giving a chance to the learner to correct the notes which are sung with wrong pitch (scale).

TABLE I
Relative and absolute frequency relationship between the notes in the SARGAM

Pure Tuning Shruti Name	Pure Tuning Ratios and Fractions	Pure Tuning Cent Scale
SA	1.00	0000
re	1.0666 = 16/15	111.308
RE	1.125 = 9/8	203.910
ga	1.2 = 6/5	315.641
GA	1.25 = 5/4	386.314
ma	1.333 = 4/3	498.043
MA	1.406 = 45/32	590.224
PA	1.5 = 3/2	701.995
dha	1.6 = 8/5	813.686
DHA	1.666 = 5/3	884.357
ni	1.8 = 9/5	1017.596
NI	1.875 = 15/8	1088.269
SA	2.0000	1200

II. Proposed SARGAM Evaluation Method

The block diagram of the proposed learner's SARGAM evaluation is shown in Fig. 1. The input SARGAM sequence is transformed to spectral domain by applying the Short-Time Fourier Transform (STFT). The note onsets are detected from the normalized spectral change energy detection function of the magnitude spectrogram. Further, the spurious note onsets are detected and removed by note frequency deviation criterion to eliminate the duplicate onsets which are detected within the note and between the notes. The note frequency is computed from the glottal closure instants of the vocal source. The note frequencies in Hertz are convert to musically relevant cent scale. The note deviation is computed against to the stored benchmark SARGAM's. The note is flagged as an out-of-tune note if its relative cent note value exceeds the predefined

threshold. The steps in the proposed learner's SARGAM evaluation is briefly explained in the following subsections.

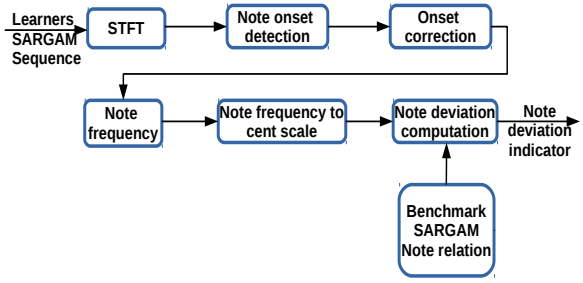

Fig. 1. *Proposed SARGAM evaluation method.*

A. Spectral Transformation

The SARGAM sequence digitized at 44.1 KHz sampling rate recorded from the learner is transformed to frequency domain by applying the STFT of 40 ms frame size and 3 ms frame shift. A relatively small frame size of 3 ms is chosen to resolve the sharp boundaries of note onsets given by

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

where $x(n)$ is the sampled time domain SARGAM sequence, $w(n)$ is the Hamming window, $N = 2048$ is the number of Fourier frequency points, $k = 0, ..., N-1$ are Fourier frequency bins.

An example of the SARGAM sequence and its spectrogram is shown in Fig. 2. The time domain waveform of SARGAM sequence SA RE GA ma PA DA NI SA' is shown in Fig. 2(a), and its spectrogram which shows clear distinction between note and non-note regions is shown in Fig. 2(b).

B. Note Onset Detection

The normalized Euclidean distance [4] between the spectral frames of the magnitude spectrogram of the SARGAM sequence $X(l,k)$ is computed to obtain the onset detection function (the peak locations in the onset detection function indicates the candidate note onsets) given by

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

$$E_x(l,k) = X_m(l,k) - X_m(l-1,k) \tag{3}$$

where $X_m(l,k)$ is the magnitude spectrum of $X(l,k)$, and $E_{df}(l)$ is the onset detection function. The distance measure is normalized to obtain the onset detection function whose peaks corresponds to the note onsets given by

$$E_{ndf}(l) = \frac{E_{df}(l)}{\sum_{k=f1}^{f2} X_m(l-1,k)^2} \tag{4}$$

The noisy regions in the onset detection function $E_{ndf}(l)$ which leads to multiple onset detection is smoothed without

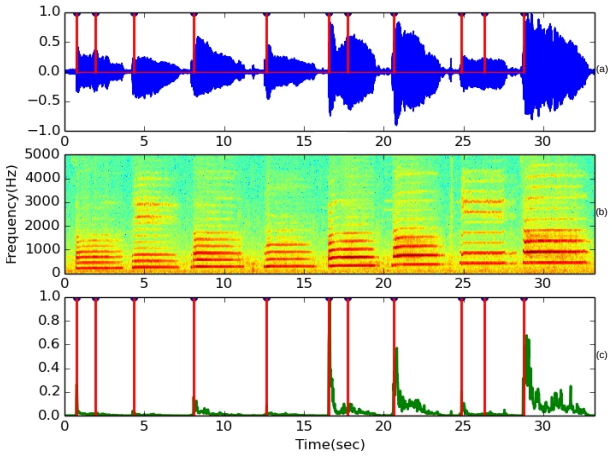

Fig. 2. *(a) Time domain SARGAM note sequence and the overlaid detected note onset markers, (b) corresponding spectrogram, and (c) onset detection function and the overlaid detected onset markers (vertical red markers).*

blurring the onset peaks by a sharp cutoff low pass filter. In time domain, the low pass like filtering is performed by taking the difference between the current frame and the contribution of exponentially weighted previous frames of detection function given by

$$y(l) = E_{ndf}(l) - \sum_{a=1}^{A} \frac{E_{ndf}(l-a)}{l} \tag{5}$$

The location of onsets in the onset detection function of Eq. 5 is obtained by peak picking heuristics as follows: The l^{th} frame is considered as onset, if the onset detection function fulfills the following conditions

$$y(l) = max(y(l-w)) \tag{6}$$

$$y(l) >= mean(y(l-w:l+w)) + \delta \tag{7}$$

$$l - l_{lastonset} > w \tag{8}$$

The values for w and δ are empirically chosen as 100 and 0.05 respectively after analyzing the note duration and detection function distributions.

The process of SARGAM note onset detection is illustrated in Fig. 2. The time domain waveform of the SARGAM sequence SA RE GA ma PA DA NI SA' and the overlaid detected onset markers (red vertical markers) is shown in Fig. 2(a). The corresponding frequency domain spectrogram of the SARGAM sequence is shown in Fig. 2(b). The onset detection function and the onset locations obtained after peak picking heuristics is shown in Fig. 2(c).

C. Note Frequency Detection

The note frequency is obtained by finding the glottal closure instants (GCI) [5] in each note. Since the note frequency varies drastically from one note to the other, the GCI locations

within each note is obtained by adaptive Zero Frequency Filtering (ZFF) [6], [7]. The GCI locations in each note region obtained in the previous subsection is Zero Frequency Filtered with the resonance frequency obtained by the Two-way-miss-match algorithm [6], [7]. The reciprocal of the time difference between the successive GCI locations is computed to obtain the frequency in Hertz. The frequency in Hertz is converted to musically relevant cent scale as

$$F = 1200 \log_2 \left(\frac{f}{f_{ref}} \right) \tag{9}$$

where f is the frequency in Hz, the reference frequency f_{ref} = 55 Hz and F is the frequency in cent scale.

An illustration of the GCI based note frequency detection is shown in Fig. 3. The time domain SARGAM sequence waveform is shown in Fig. 3(a). The frequency (blue contour) of each note region is shown in Fig. 3(b).

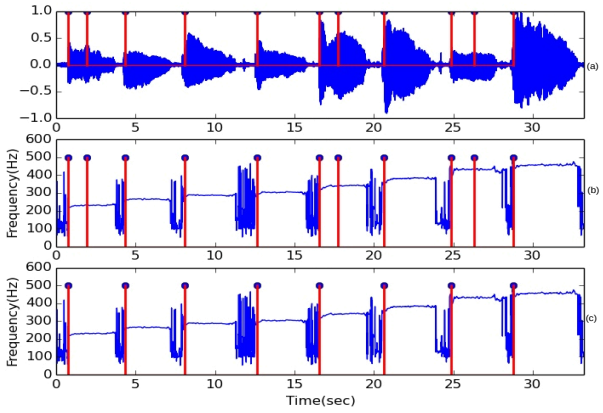

Fig. 3. *Illustration of the SARGAM note onset correction from the note frequency. (a) Time domain SARGAM note sequence and the overlaid detected note onset markers, (b) the absolute frequencies of the note sequence and the overlaid onset markers, and (c) the absolute frequencies of the note sequence and the overlaid correct onset markers (vertical red markers).*

D. Onset Correction

The spectral magnitude changes within each note due to the missing higher harmonics, fading of the higher harmonics from high energy to low energy and energy fluctuations within each note causes the onset detection function to have multiple significant peaks within a note. Further, the onset detection function shows significant peak magnitude in the non-note regions due to sudden impulsive like noise caused by various uncontrolled environmental factors during SARGAM recording through microphone. These spurious note onsets which are detected along with the true onsets detected in subsection II-B are eliminated by observing the fact that the frequency within the note remains almost constant or stable and the frequency between the notes (non-note regions) varies randomly because of no pitch information within this regions. From Fig. 3(b), we can observe that the note detection method has detected spurious note onsets within the notes 1 (approximately the note duration is from 0.5 sec to 4.5 sec), 5 (16 sec to 21 sec), and 6 (25 sec to 29 sec) shown as vertical marker within each stable note region. The spurious onsets are mainly due to the significant peak magnitude in the detection function which is comparable with that of the peak magnitudes of the true onsets. From Fig. 3(b), we can also observe that, if the onset is a valid onset, the frequency contour after the onset is almost constant (stable), and the pitch contour before the onset is highly variable or unstable. Further, if the detected onset is the spurious onset, two cases are possible i.e., the detected onset may be within the note or it may be between the notes. If the spurious onset occurs within the note, the pitch contour before and after the note onset will be stable or almost constant. Here, stable region is identified as the note region whose frequency variance is less than 80 cents. The spuriously detected onset between the notes shows highly variable pitch contour before and after the onset. The algorithm for the spurious onset detection and elimination is given below. If the spurious onset is detected between the notes (non-note region), then the frequency variance after and before the onset is computed and eliminated by Algorithm 1. If the spurious onset is within the note, then the frequency variance after and before the onset is computed and eliminated by Algorithm 2.

Result: V: vector containing the true onset locations
V = vector with onset locations containing both true and spurious onsets ;
F = vector containing the frequency in cents ;
$L = 100$ is total number of frames considered for computing the variance ;
for $i = 1 : length(\boldsymbol{V})$ **do**
  $l = \mathbf{V}[i]$;
  $\overline{x} = \frac{1}{L} \sum_{k=l}^{l+L} F[k]$;
  stdDevBegFram = $\sqrt{\frac{1}{L-1} \sum_{k=l}^{l+L} (F[k] - \overline{x})^2}$;
  $\overline{x} = \frac{1}{L} \sum_{k=l-L}^{l} F[k]$;
  stdDevEndFram = $\sqrt{\frac{1}{L-1} \sum_{k=l-L}^{l} (F[k] - \overline{x})^2}$;
  if *stdDevBegFram* $>$ *80 and stdDevEndFram* $>$ *80* **then**
    $V[i] = 0$;
  else
    $V[i] = \mathbf{V}[i]$;
  end
end

Algorithm 1: Detection and elimination of spurious onsets between the notes.

An illustration of the spurious onsets detected within the note and between the notes by the note onset detection method described in subsectionII-B is shown in Fig. 3. From Fig. 3(b), we can observe the spurious note onsets detected within the notes (shown as vertical red markers). From Fig. 3(c), we can observe that the spurious notes are eliminated after applying the note frequency deviation criterion as explained previously.

Result: V the vector containing the true onset locations
V = vector with onset locations containing both true and spurious onsets ;
F = vector containing the frequency in cents ;
$L = 100$ is total number of frames considered for computing the variance ;
for $i = 1 : length(\boldsymbol{V})$ **do**
  $l = \mathbf{V}[i]$;
  $\overline{x} = \frac{1}{L}\sum_{k=l}^{l+L} F[k]$;
  stdDevBegFram = $\sqrt{\frac{1}{L-1}\sum_{k=l}^{l+L}(F[k]-\overline{x})^2}$;
  $\overline{x} = \frac{1}{L}\sum_{k=l-L}^{l} F[k]$;
  stdDevEndFram = $\sqrt{\frac{1}{L-1}\sum_{k=l-L}^{l}(F[k]-\overline{x})^2}$;
  if *stdDevBegFram* $>$ *80 and stdDevEndFram* $<$ *80* **then**
    $V[i] = 0$;
  else
    $V[i] = \mathbf{V}[i]$;
  end
end

Algorithm 2: Detection and elimination of spurious onsets within the note.

E. Note Frequency Assignment

A single absolute frequency value in cents is assigned to each note obtained in the previous subsection. The note frequency in cents is identified as the median of the frequencies in the 100 frames (which corresponds to 300 ms) in the stable region of the note. Assuming that the note sustains for atleast 300 ms in the stable region. The stable region is identified as the note region whose frequency variance is less than 80 cents. An illustration of the note frequency detection from the stable regions is shown in Fig. 4(a). The colored contours on the blue SARGAM frequency contour are the stable regions from where the note frequencies are computed.

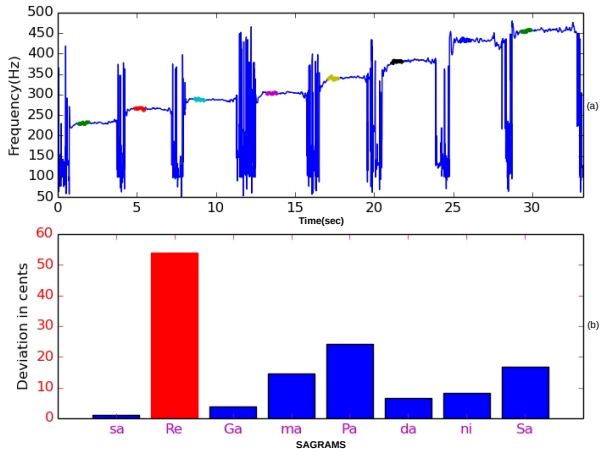

Fig. 4. *The SARGAM note deviation indicator. (a) The absolute frequency values of the note sequence and the region of the note used for calculating the absolute note value, and (b) note frequency deviation indicator in cent scale. The red bar indicates the note which is in out-of-tune.*

F. Note Deviation Computation

The note deviation factor (explained later with an example) is computed as the absolute deviation between the relative frequency difference in cent scale from the base note to the remaining notes in the learner's SARGAM sequence and the ideal relative frequency difference between the same set of notes computed from the Table I. Table I shows the frequency relationship between the base note (SA) and the subsequent notes in the SARGAM sequence. The fist column in Table I shows the Suddh and Komal notes. The second column shows the relative pure tuning ratios for the base note 'SA'. The column three shows the relative frequency difference in cent scale for the base note 'SA'. The third column shows that if the note 'SA' is the base or the reference note with the zero cents, the note 're' should be 111.308 cents away from the base note 'SA', 'RE' should be 203.910 cents away from 'SA' and so on.

An example of note deviation computation is explained below. For example, if the learner has sung the SARGAM sequence SA RE GA ma PA DA NI SA' as shown in Fig. 3(a) with the corresponding note frequencies obtained in Hz is
$S_{fHz} = [229, 266, 287, 303, 339, 381, 432, 455]$
and its cent scale equivalent frequency computed as described in subsection II-E
$S_{fCent} = [2475, 2731, 2864, 2958, 3151, 3352, 3571, 3658]$.
The relative cent frequency difference is computed as

$$S_{diff} = |S_{fCent} - S_{fCent}[0]| \tag{10}$$

where $|.|$ indicates the absolute difference and $S_{fCent}[0]$ is the base note frequency $(2475cents)$. For the given example, the relative cent frequency difference is
$S_{diff} = [0, 256, 389, 483, 676, 877, 1096, 1183]$.
Now, by looking at the Table I, we can get the ideal relative cent scale frequency differences for the notes SA RE GA ma PA DA NI SA' as
$S_{iCent} = [0, 203, 386, 498, 701, 884, 1088, 1200]$.
The absolute cent scale deviation between the sung notes and the ideal notes is computed as

$$nDev = |S_{diff} - S_{iCent}| \tag{11}$$

where the vector $nDev$ will contain the note deviation values. For the above example, $nDev = [0, 53, 3, 13, 25, 7, 8, 17]$. The notes which deviates above 50 cents (i.e., half a semitone) is marked as out-of-tune notes. An illustration of the out-of-tune note detection is shown in Fig. 4 (b). The x-axis shows the SARGAM note labels and the y-axis shows the note deviation in cents. From Fig. 4(b), we can observe that the note 'RE' is in out-of-tune with approximately 55 cents, which is displayed as a red vertical bar.

III. Evaluation and Discussion

The performance of the proposed SARGAM evaluation method is assessed by five female semi-professional (SP) Hindustani singer's who are practicing singing from their childhood. The benchmark SARGAM's (instructor SARGAM's)

are recorded from the professional Hindustani vocalist. The SARGAM's recorded from the vocalist are (i) Sa, Re, Ga, ma, Pa, Da, Ni, Sa', (ii) Sa, ni, Da, Pa, Ma, Ga, Re, Sa, (iii) Sa Sa, Re Re, Ga Ga, Ma Ma, Pa Pa, Dha Dha, Ni Ni, Sa' Sa', (iv) Sa Sa, Ni Ni, Dha Dha, Pa Pa, Ma Ma, Ga Ga, Re Re, Sa Sa, and (v) Sa Re Ga Re Sa, Re Ga Ma Ga Re, Ga Ma Pa Ma Ga, Ma Pa Dha Pa Ma. The ideal note deviations of SARGAM's from the base note ('Sa') for each SARGAM is obtained from the Table I and stored in the database. The semi-professional singer's are asked to repeat each SARGAM sequence in five sessions after listening from the benchmark SARGAMS's at their own will. Thus, there were a total of 125 (5 (SARGAM's) x 5 (singers) x 5 (sessions)) test SARGAM's. Since the SP singer's already well trained in SARGAM practice, they were asked to rate the proposed method on a five point (1-5) scale ranging from very bad to very good. Also, the SP singer's are instructed to intentionally sing some of the notes in out-of-tune to assess the reliability of the proposed method. The singer's rated the proposed method with an average of 4.5 rating saying that they are impressed with the performance of the automatic system for its accuracy. They also appreciated the method for accurately detecting even very slight changes in the deviations of the SARGAM's notes. The singer's also expressed that the proposed method will be much more beneficial to the learner's if it can correct the out-of-tune sung notes on its own and play back the correct in-tune notes as feedback to the learner's, which we have kept it as a future work.

IV. Summary and Conclusions

An automatic SARGAM evaluation method is proposed to detect the notes which deviates from the predefined musical ratio. The method involves initially finding the note onsets to determine the SARGAM note boundaries in the spectral domain. Further, the spurious note onsets are eliminated by observing the stable frequency nature of the notes in the SARGAM's. The note frequency in musically relevant cent scale is computed by finding the glottal closure instants of the vocal source. Further, the out-of-tune notes are determined by finding the frequency deviation factor from the ideal relative note ratios. In future, the authors would like to evaluate the proposed method by designing objective measures. Also, would like to evaluate with more number of learner's including the beginners.

References

[1] "Hindustani classical music," http://raag-hindustani.com/LearningTools.html, last accessed: 2017-03-13.

[2] "Svara," https://en.wikipedia.org/wiki/Svara, last accessed: 2017-03-13.

[3] "Know your raga," http://www.knowyourraga.com/ragagyan/?docname=swar, last accessed: 2017-03-13.

[4] 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.

[5] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[6] 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.

[7] 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.