Audio Segmentation Using A Priori Information in the Context of Karnatic Music

Ashwin Kalyan V, Sreecharan Sankaranarayanan and Sumam David S
Music and Audio Research Group
Department of Electronics and Communication Engineering
National Institute of Technology Karnataka, Surathkal, India
Email: {asaavashwin, sreecharan93, sumam}@ieee.org

***Abstract*—Karnatic Music (KM) is distinct because of the prevalence of gamaka - embellishments to musical notes in the form of frequency traversals. Another important aspect of KM is that the performance style is mostly extempore. Hence, Music Information Retrieval (MIR) tasks in the context of KM are highly challenging. This paper deals with the task of Audio Segmentation and its application to MIR challenges of KM at various levels. This work presents a method that incorporates a priori knowledge about the music system and the audio track at hand for segmenting the audio into its constituent notes. The method uses amplitude and energy based features to train a neural network and an accuracy of 95.2% has been achieved on KM audio samples. The paper also elucidates the application of the method to important MIR tasks such as Music Transcription and Score-Alignment in the context of KM.**

***Index Terms*—Music Information Retrieval, Karnatic Music, Audio Segmentation, Onset Detection**

I. Introduction

Karnatic music (KM) is characterized by phrases and consists of embellishments in the form of frequency traversals and oscillations called gamaka.[1] Because of the usage of gamaka, individual notes have significant frequency digressions from intended pitch; sometimes up to 4-6 semitones.The same note is typically associated with multiple pitch tracks due to different embellishments. Hence, notes cannot be distinguished based on pitch alone, which is generally the case with Western music. Another factor which complicates the analysis of KM is the extempore form of the music. A significant part of a performance is extempore while only a small portion follows previously composed or fixed music. This fixed component of KM also varies from performer to performer owing to different schools of thought, geographical influences and individual creativity. Because of this diversity and complexity, information retrieval methods for KM need to incorporate some prior knowledge about the music system to achieve reasonable results.

This work focusses on the task of audio segmentation in the context of KM. In most MIR tasks, it is necessary to segment the audio file before proceeding to further processing stages; making audio segmentation a task of utmost importance. The segmentation required can be of different levels depending on the goal of the task. For example, music archival tasks require the identification of various stages like alapana, krithi, neraval and kalpana-swara. In this case, the segmentation task is of a high level and the entire audio may be segmented into only a few parts. Automation of this task is highly useful in developing structured archives and search engines for KM. Tasks like identifying musical patterns require cycle or phrase-level resolution. The next level of segmentation is at the note-level and this resolution is suitable for tasks like score alignment and music transcription. Figure 1 shows the hierarchy of audio segmentation required in MIR tasks in the background of KM.

In this work, methods to segment audio at the note-level are explained. Making use of a priori information available about the target audio, techniques to improve the performance of the segmentation methods are discussed.

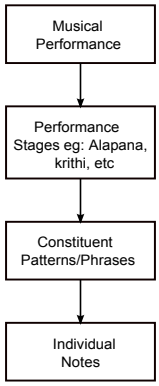

Fig. 1: Hierarchy of audio segmentation tasks in KM

II. Background

The first part of this section concerns the explanation of the musical terms used in the paper and the second part presents the overview of the paper.

A. Musical Terms

The *raga* is the fundamental melodic structure in KM and this is generally specified using a scale. The scale lists the order of notes that is to be followed while ascending and descending an octave. Every *raga* has a unique set of musical notes and patterns. Incorporating this knowledge into the MIR

system simplifies the task at hand to a great extent.
A typical rendition in KM consists of both fixed and extempore components. The most common stages of a KM rendition are as follows:

1) **Alapana:** This stage is extempore and consists of phrases that are not set to any rhythmic structure. This stage usually makes use of a single raga and the purpose of this stage is to set an ambience for the rendition.
2) **Krithi:** This stage consists of a previously composed song and it generally has an underlying fixed structure. The performer adds to this structure based on his creative potential.
3) **Neraval:** In this stage, the performer chooses a particular line from the song for creative elaboration.
4) **Kalpana-Swara:** In this stage, the performer renders a string of notes and arrives at a specific point of the song. This stage usually consists of interesting mathematical patterns and progressions. Here, the word swara stands for note.

The *Krithi* and the *Kalpana-swara* stages are amenable to a note-level analysis while the raga and the neraval stages need attention at the phrase-level.

This paper focusses on note-level segmentation and so, the audio samples used for the validation of the methods in this paper are extracted from the *krithi* and *kalpana-swara* stages of the performance.

B. Organization

The organization of the paper is as follows: Section III elucidates the method for note-level audio segmentation. Section IV concerns the application of this method to the task of music transcription and Section V deals with the task of score-alignment. Section VI states the conclusions and possible future improvements in this regard.

III. Note-Level Segmentation

This section explains the method used to segment audio into its constituent notes.Monophonic music of vocal and violin were recorded at a sampling rate of 44.1kHz. The music samples used were performed at different rhythms to validate the method for varying speeds.

A. Background

When a new note corresponds to a new syllable (as in the case of Kalpana-swara), every note corresponds to a new attack or onset. In such cases, the task of audio segmentation reduces to audio-onset detection.

Previous methods to detect onsets make use of derivative-based techniques [2] which suffer from the problem of setting thresholds. To overcome this, classifiers [3] have been used to detect onsets based on extracted amplitude and energy features. Our method also uses a classifier to detect onsets but differs in the process of feature extraction by using a priori information.By incorporating a priori information, the accuracy of the classifier can be improved.
This method involves the following steps which will be explained in detail in the following sub-sections:

1) Amplitude Envelope Extraction
2) Wavelet Transform
3) Feature Extraction and Classification
4) Post-processing

B. Amplitude-Envelope extraction

The amplitude envelope is extracted from the audio data using an intuitive method that requires minimal computation. An analysis window of 10ms is generally sufficient to identify onset locations. For every frame, the maximum value taken by the audio signal is chosen as the value of the envelope at that instant.The so extracted envelope can be spline interpolated to the length of the audio signal.

C. Wavelet Transform

The task of detecting onsets requires detection of peaks in the amplitude envelope. This stage helps to remove the spurious peaks that affect the performance of the method. This stage uses a priori information in the form of rhythm parameter. The rhythm parameter is defined as the approximate duration of the shortest note in the audio.In KM parlance, this stands for the least count in the tala (rhythmic structure in KM) i.e. akshara. The duration of the notes in the song is an integral multiple of the rhythm parameter approximately.

The wavelet transform of the amplitude envelope is computed with a scale that corresponds to the rhythm parameter. Intuitively, the wavelet transform behaves like a low-pass filter removing all the intra-note variations and magnifying the note-level changes as illustrated in Fig. 2. It can also be observed in the figure that the local minima correspond to the onsets. The rhythm parameter can be visualized to be controlling the cut-off frequency of this filter. This view can be seen mathematically by viewing the wavelet transform as a correlation, which is illustrated in the following discussion. [4] The Continuous wavelet transform (CWT), X_{CWT} of a signal $x(t)$ is given by

$$X_{CWT}(p,q) = \frac{1}{\sqrt{p}} \int_{-\infty}^{\infty} x(t)\psi^*(\frac{t-q}{p})dt \tag{1}$$

where

$$\psi_(p,q) = \frac{1}{\sqrt{p}}\psi(\frac{t-q}{p}) \tag{2}$$

is the mother wavelet, $*$ represents the complex conjugate, p is the scaling parameter and q, the time-shift parameter.This definition of the CWT when compared with the definition of correlation,

$$R(q) = \int_{-\infty}^{\infty} x_1(t)x_2(t-q)dt \tag{3}$$

we get,

$$X_{CWT}(p,q) = x(t) \otimes \psi^*_{p,q}(t) \tag{4}$$

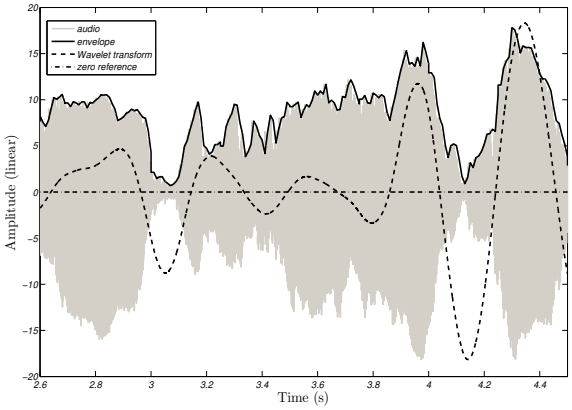

Fig. 2: Audio signal along with its amplitude envelope and corresponding wavelet transform.

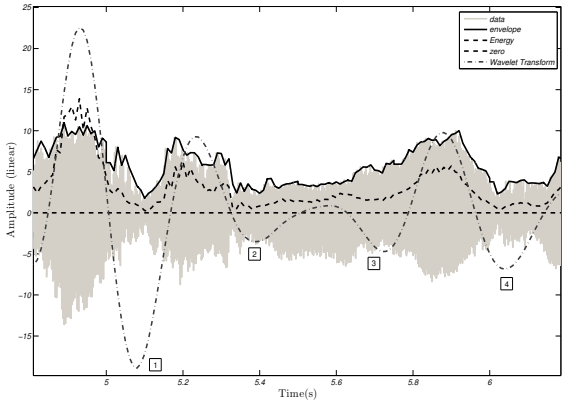

Fig. 3: Audio signal along with amplitude and energy features illustrating false-positives in onset detection

where $\otimes$ symbolizes correlation.The mother wavelet used in our works is the *reverse bi-orthogonal 4.4* as this gives comparatively better performance, owing to its symmetric and structural properties.Figure 2 shows the audio signal, amplitude envelope and its wavelet transform for a better understanding of the motivation for using wavelets.

D. Feature Extraction and Classification

The times corresponding to local minima of the wavelet transform are taken to be potential candidates for note-onset. The spline-interpolated value of the amplitude envelope and the local minima value of the wavelet transform constitute the amplitude-features. To get the energy based features, the energy of the audio signal is calculated for a frame-length of 10ms and its wavelet transform is computed as explained in the previous section. The energy and its wavelet transform values constitute the energy-features. Using both the amplitude and energy features, a single-layer neural network classifier was trained and an accuracy of 91.6% was obtained with 10-fold cross-validation.

E. Post-processing

This stage makes use of the rhythm parameter to improve the accuracy of onset detection.The definition of the rhythm parameter requires that all onsets must occur at integral multiples of the rhythm-parameter within a certain error ϵ. To determine the suitable scaling for the wavelet transform, an approximation of the rhythm factor is sufficient. This approximation is bettered in this stage by updating the value of the rhythm parameter based on the prediction of the trained neural network. The post-processing stage consists of the following steps:

1) The rhythm parameter ρ is initialized to the approximate value ρ_o when the post-processing stage begins from the start of audio.
2) The next onset is position is expected at a time τ which is an integral multiple of ρ. If this criterion is not satisfied, then this point is re-classified as a false onset.
3) When a true onset which satisfies the timing criterion is identified, the rhythm parameter is updated. Let the new onset be at time τ' while the previous onset is at τ.As the timing criterion is satisfied, the next onset occurs approximately after $k\rho$, where k is an integer. The new rhythm parameter ρ' is updated as

$$\rho' = \frac{1}{2}\Big(\rho + \frac{\tau' - \tau}{k}\Big) \tag{5}$$

This stage increases the precision of onset detection as false-positives are re-classified into false-negatives. Figure 3 illustrates the elimination of false onsets through this post-processing stage. Points 1 and 4 marked as potential onset locations are true-positives. However, points 2 and 3 which are false-positives arise due to a decay stage after the attack. These are re-classified properly during the post-processing stage as they do not satisfy the timing criterion. The accuracy increased from 91.6% to 95.2% for the audio samples considered, after this stage.

IV. Music Transcription

Automatic Music Transcription is the process of converting music in the form of audio signals into musical notation. In Western Music, notes typically have a direct mapping to frequency. However in KM, the concept of absolute pitch does not exist and all notes are calculated relative to the tonic which can be fixed at any frequency . Hence, methods using chromagram based features [5] to transcribe music do not hold in the case of KM. Figure 4 shows the block-diagram of the music transcription method which is explained in the following sub-sections.

A. Overview

Due to the prevalence of gamaka, one note has various pitch tracks associated with it, depending on the musical context. The context is determined by the raga and aesthetics. Figure 5 shows four variation for the note 'madhyama' in the raga Kambodhi. In KM transcription, it is necessary to transcribe

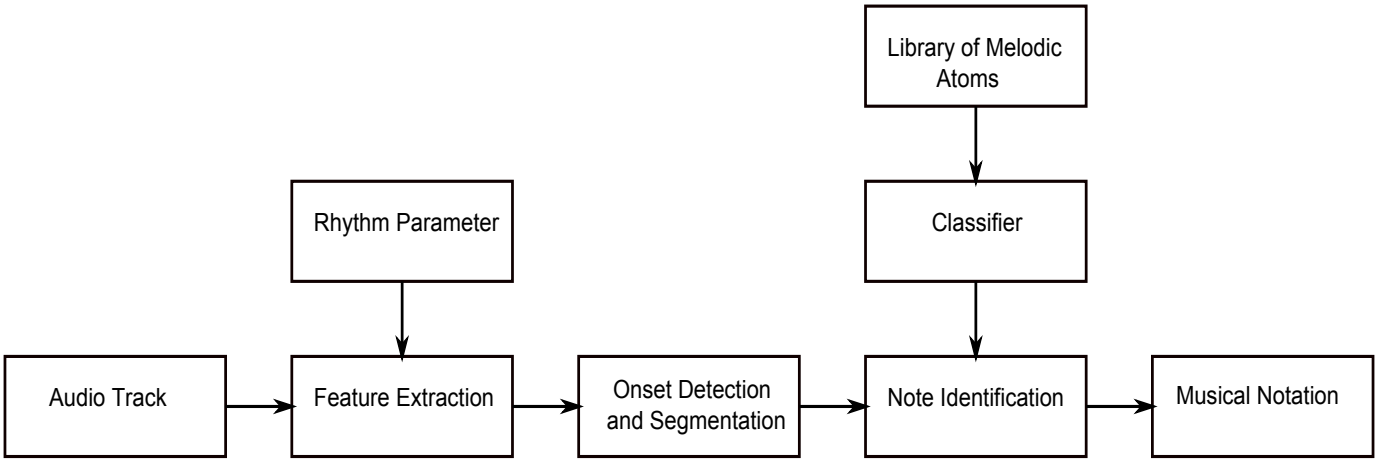

Fig. 4: Block-diagram view of the Music Transcription process. The note-level audio segmentation is a key stage.

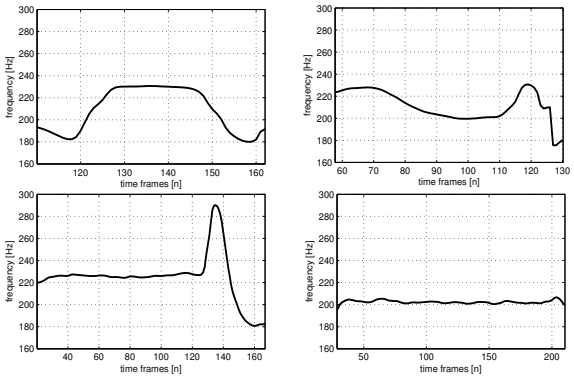

Fig. 5: Four different gamaka for the note 'madhyama'in the raga Kambodhi.

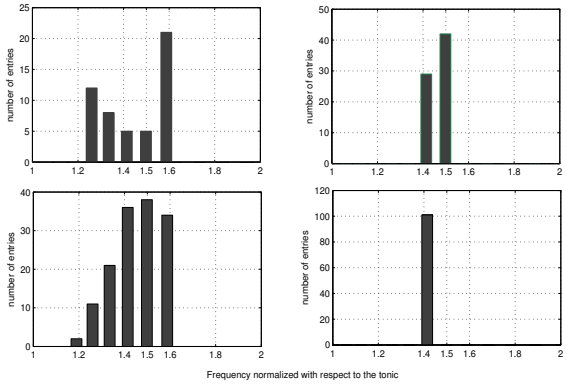

Fig. 6: The note histogram feature of the 4 gamaka shown in Fig. 5. For brevity, the x-axis spans one octave corresponding to 12 bins or notes.

these seemingly different frequency tracks into the same note, making the task challenging. To transcribe KM, a library of the melodic atoms [6] that are aesthetically correct in the context needs to be built. It should be evident that building an exhaustive library is a tough task to accomplish due to the vast number of variations that are possible. To overcome this problem, only the frequently used gamaka are accounted for, in the development of the library. The library-development can be effectively structured by adopting a raga-specific strategy, implying that the raga is needed for transcription as a priori information.

B. Music Transcription Process

First, the audio is segmented using the onset-detection algorithm mentioned in section III. Following this, the pitch track is estimated [7] and suitable features are extracted for the classification stage which classifies the segment into appropriate notes. The following features were used by us:

- **Mean pitch**
- **Variance**
- **Part Mean** the value of the mean of the signal is sampled at different instants
- **Note Histogram** a feature similar to the chromagram that makes use of the pitch track of the segment. The pitch is calculated relative to the tonic.

Mean and variance capture the basic statistics of the pitch track while the Part-Mean accounts for the time dependency. The note-histogram highlights the distribution of the pitch track between the adjacent notes (see Fig.6).One can verify the importance of these features by analyzing their effect on the gamaka shown in Fig. 5. A Naive Bayes classifier was implemented and a transcription accuracy of 92.6% was achieved. The method was tested on monophonic violin and vocal audio samples that were recorded by the authors. The *raga mohana* was used to develop the library of *gamaka*.

V. Score-Alignment

Aligning notation with audio is an extremely important MIR task of great relevance to music education and automatic

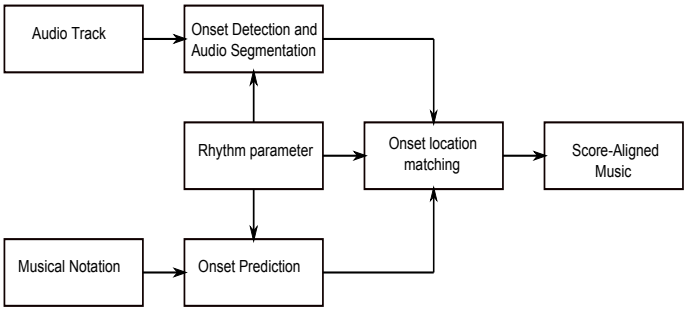

Fig. 7: Block-diagram level overview of Score-Alignment

music generation. Many self-learning modules use Score-Alignment [8] to help the learner to synchronize with the target music through visual prompt in the form of notation. In KM, musical forms taught initially have a simple structure with generally one syllable per note. For such compositions, the audio segmentation proposed by us in the previous section can be used to synchronize the notation and a stored audio of the music. The rhythm parameter is also used to predict a set of onsets from the notation. Then, these onset locations are matched to align the score and the music. In the case of compositions with lyrics, better results were obtained by matching the detected onsets with the syllables of the lyric. Figure 7 shows a block-diagram level overview of the score-alignment method.

VI. Conclusion and Future Work

The present method works well for audio with simple musical structures with an onset detection accuracy of 95.2%. This can be used to develop various KM-educational tools aimed at a beginner level.

To be able to work with complex musical forms, a method that incorporates data from the frequency track effectively needs to be developed. This needs to be raga-specific and must rely on an extensive library of allowed melodic atoms.

References

[1] T. Krishna and V. Ishwar, "Carnatic music: Svara, gamaka, motif and raga identity," in *Proc. of the 2nd CompMusic Workshop*. Istanbul,Turkey: Universitat Pompeu Fabra, Jul. 2012.

[2] J. Ricard, "An implementation of multi-band onset detection," in *Extended abstract of the 1st Annual Music Information Retrieval Evaluation eXchange (MIREX)*, London, U.K, 2005.

[3] C. Chuan and E. Chew, "Audio onset detection using machine learning techniques: the effect and applicability of key and tempo information," *Computer Science Department Technical Report*, no. 08-895, 2008.

[4] S. Narasimhan, N. Basumallick, and S. Veena, *Introduction to wavelet transform: a signal processing approach*. Alpha Science International, Ltd, 2011.

[5] M. Mauch and S. Dixon, "Approximate note transcription for the improved identification of difficult chords." in *ISMIR*, 2010, pp. 135–140.

[6] A. Krishnaswamy, "Melodic atoms for transcribing carnatic music," in *ISMIR*, 2004.

[7] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," vol. 17, no. 1193, pp. 97–110, 1993.

[8] R. B. Dannenberg and C. Raphael, "Music score alignment and computer accompaniment," *Communications of the ACM*, vol. 49, no. 8, pp. 38–43, 2006.