INTER AND INTRA ITEM SEGMENTATION OF CONTINUOUS AUDIO RECORDINGS OF CARNATIC MUSIC FOR ARCHIVAL

Padi Sarala
Computer Science and Engineering
Indian Institute of Technology, Madras
padi.sarala@gmail.com

Hema A.Murthy
Computer Science and Engineering
Indian Institute of Technology, Madras
hema@cse.iitm.ac.in

ABSTRACT

The purpose of this paper is to segment carnatic music recordings into individual items for archival purposes using applauses. A concert in carnatic music is replete with applauses. These applauses may be inter-item or intra-item applauses. A property of an item in carnatic music, is that within every item, a small portion of the audio corresponds to the rendering of a composition which is rendered by the entire ensemble of lead performer and accompanying instruments. A concert is divided into segments using applauses and the location of the ensemble in every item is first obtained using Cent Filterbank Cepstral Coefficients (CFCC) combined with Gaussian Mixture Models (GMMs). Since constituent parts of an item are rendered in a single *raga*, *raga* information is used to merge adjacent segments belonging to the same item. Inter-item applauses are used to locate the end of an item in a concert. The results are evaluated for fifty live recordings with 990 applauses in total. The classification accuracy for inter and intra item applauses is 93%. Given a song list and the audio, the song list is mapped to the segmented audio of items, which are then stored in the database.

1. INTRODUCTION

Indian classical music consists of two popular traditions, namely, Carnatic and Hindustani. In most of the carnatic music concerts audio recordings are continuous and unsegmented [1]. A concert in carnatic music is replete with applauses. As most Indian music is improvisational, the audience applauds the artist spontaneously. Thus, in a given performance, there can be a number of applauses even within a single item. The general structure of a concert is shown in Figure 1. As shown in Figure a concert is made up of a number of different items. Each item can optionally have a solo vocal, a solo violin, a solo percussion (referred to as Thani in the Figure) but a composition is mandatory. Generally, the audience applauds the artist at the end of every item. Owing to the spontaneity that can be leveraged by an artist in a carnatic music concert, the concert is more like a dialogue between the artist and the audience [10]. Aesthetic phrases are immediately acknowledged by the audience. Owing to this aspect, applauses can occur anywhere in the concert. Different types of items and their constituent segments are shown in Figure 2. An item can be i) a single composition, ii) a vocal followed by the violin and then a composition, iii) a composition followed by ThaniAvarthanam, iv) a ragam tanam pallavi (RTP) that consists of solo pieces of vocal , violin, and the pallavi, which is equivalent to a composition for the purpose of analysis. The composition itself can be rendered in tandem by a lead performer, followed by accompanying instruments. For better understanding of carnatic music terms, we refer the reader to the supplemental material [2].

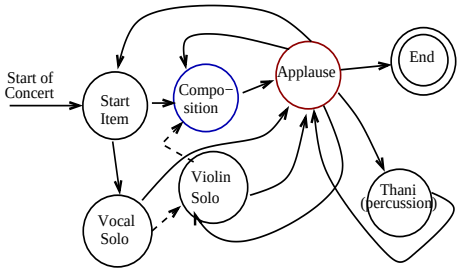

Figure 1: *General structure of carnatic music concert.*

From the Figure 2 it is clear that a composition segment is mandatory in an item. Furthermore an item can consist of one or more composition segments as shown in Figure 3. This is especially true for the Main Song in a concert, where different parts of the composition lend itself to significant improvisation. This leads to the composition being fragmented into a number of segments owing to applauses. A composition segment is also the last segment in an item. Since compositions are mandatory in every item, the begin and end of an item can be found in the vicinity of composition segments. To determine whether a sequence of composition segments belong to the same item or not, information about *raga* is required. In general, an item can be performed in single *raga*. The characteristics of *raga* can be detected by

[1] "http://www.sangeethapriya.org"

[2] "http://xa.yimg.com/kq/groups/2785355/2047532645/name/Carnatic+music+terminology.pdf"

melodic histograms [6]. So, to identify different items , melodic histograms can be used. This will not work when an item is intrinsically performed in multiple *ragas*. Such items are quite rare in a concert and seldom have multiple composition segments like *viruttam* or a vocal solo sung in isolation.

The first task for archival of carnatic music recordings requires that the audio should be segmented into individual items. A finer segmentation of the concert into Vocal solo, Violin solo, Song (Composition), and percussion solo, vocal-violin tandem, different percussion instruments in tandem may be of relevance to a keen learner, listener or researcher. In this paper, an attempt is made to determine the fine segments using applauses. The fine segments are first labeled using supervised Gaussian Mixture Models (GMM). Most of the segments in an item are rendered in single *raga*, *raga* information is used to merge the adjacent segments belonging to same item. Given an item list (text), the items of audio are aligned. This results in a database where the items can be queried using the text corresponding to that of the item.

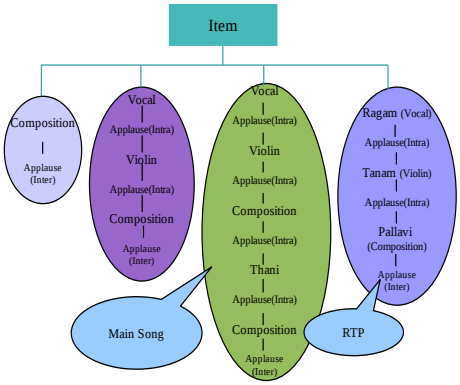

Figure 2: *Different types of items in a carnatic music concert.*

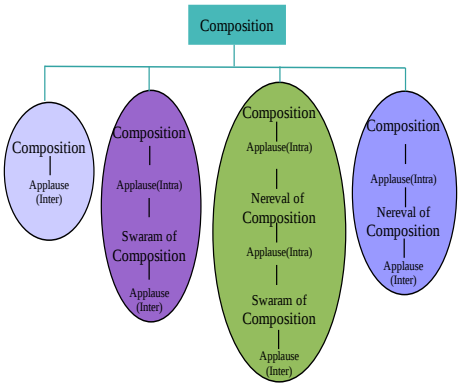

Figure 3: *Different types of compositions in a carnatic music concert.*

The remaining part of the paper is organised as follows. In Section 2 we discuss the process of segmentation of a concert and identification of different composition segments . Section 2 also discusses previous work for locating the applauses in a concert using different spectral domain features [11]. Based on composition segments in a concert usage of pitch histograms to merge the composition segments belonging to a same *raga* is discussed in Section 3. Section 4 discusses for a given meta-data in terms of text corresponding to items in a concert the meta-data is used to index the audio. In Section 5 we discusses about the database used for the experimental evaluation and the results obtained for segmentation and, inter and intra item classification. Finally, Section 6 concludes the work.

2. SEGMENTATION OF CARNATIC MUSIC CONCERT

In Subsection 2.1 we discuss identification of applause locations in a concert using different features. And Subsection 2.2 explains how these identified applause locations are used for segmentation of a concert for identifying the composition segments.

2.1 Applause Analysis

In a carnatic music concert applauses can occur at the end of an item or within an item, as explained in Section 1. The paper [11] discusses applause identification in carnatic music and its applications to archival of carnatic music. The paper also discusses the use of Cumulative Sum (CUSUM) [3] to highlight the important aspects of a concert. Spectral domain features like spectral flux and spectral entropy are used for detecting the applause locations for the given concert. Figure 4 shows applause locations (peaks) in a 3 hour concert using CUSUM of spectral entropy as the feature. This concert has 25 applauses and 9 items. The complete details about the applause identification and thresholds used for detection of applauses are explained in Section 5.

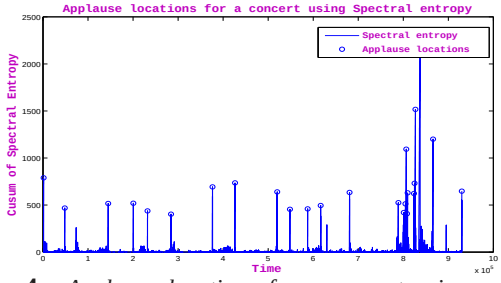

Figure 4: *Applause locations for a concert using spectral entropy feature.*

2.2 Identifying the Composition Segment Locations in a Concert

In order to find inter or intra item applauses we need to find change in *raga* among all the segments. As explained in Section 1, the end of an item is always after the composition segment. Therefore, composition segment locations probably indicates an end of an item. Owing to this, composition segments are identified in a concert. In order to find the composition segments in a concert, GMMs are built for four classes, namely a Vocal solo, a Violin solo, a Composition ensemble and ThaniAvarthanam using Cent Filterbank based Cepstral Coefficients (CFCC).

In carnatic music each singer renders every item with respect to a reference called Tonic (*sadja*) [1]. Any melodic analysis of Indian music therefore requires a

normalisation with the tonic. Within a concert the lead performer chooses the tonic and the accompanying instruments are also tuned it. Across concerts the same musician can have different tonic. Nevertheless, the tonic chosen for a concert is maintained throughout the concert using an instrument called the tambura that provides the drone. Therefore, the analysis of a concert depends on the tonic. CFCCs are proposed in this paper to perform normalisation on the spectrum. The fundamental difference between Mel Frequency Cepstral Coefficients (MFCC) and CFCC is that, in CFCC the frequency spectrum of the signal is mapped to the cent scale (after normalising with tonic) as opposed to the Mel scale. Extraction of CFCCs is detailed below:

1. The audio signal is divided into frames.
2. The short-time discrete Fourier is computed for each frame.
3. The power spectrum is then multiplied by a bank of filters that are spaced uniformly in the tonic normalised cent scale. The cent scale is defined as:

$$cent = 1200 \times \log_2 \left(\frac{f}{tonic} \right) \quad (1)$$

4. The energy in each filter is computed. The logarithm of filterbank energies is computed.
5. Discrete cosine transform (DCT-II) of log filter bank energies is computed.
6. The cepstral coefficients obtained after DCT computation are used as features for building the GMMs.

Alternatively, the chroma filterbanks can also be used . The chroma filterbanks discussed in [7], is primarily for the Western classical music where the scale is equitemperament and is characterised by a unique set of 12 semitones, subsets of which are used in performances. As indicated in [12], Indian music pitches follow a just intonation rather than an equitemperament intonation. In [9], it is shown that even just intonation is not adequate because pitch histogram across all *raga*s of carnatic music appears to be more or less continuous. To account for this, the chroma filterbanks include a set of overlapping filters. Furthermore, the filters can span less than a semitone, to account for the fact that in Indian music two adjacent *svaras* need not be separated by a semitone. Cepstral coefficients derived using the Chroma filterbanks are referred to as ChromaFCC.

Figure 5 shows MFCC, ChromaFCC and CFCC features for three types of segments, Vocal, Violin, and Composition. As shown in the Figure 5 the mel filterbanks emphasises mostly timbre, the chroma filterbanks emphasises the *svaras*, and the cent filterbanks emphasises both *svara* and the timbre around the melodic frequency. In this paper, we therefore use CFCC to distinguish different types of segments. To find CFCC cepstral coefficients we need to identify the tonic for every concert. To this extent pitch histograms are used to find the tonic as explained in [1]. Table 1 (Subsection 5.1) provides concerts for each singer and the tonic values identified using the pitch histograms. GMMs [2] are

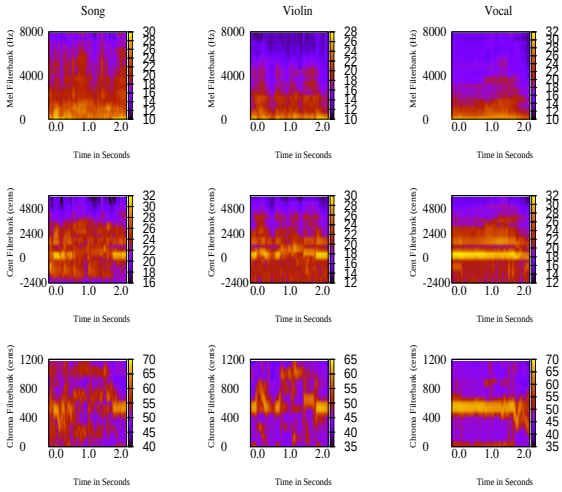

Figure 5: *Time frequency representations of Composition (Song), Vocal, Violin using MFCC, CFCC and ChromaFCC.*

frequently used for speaker verification and identification, and for segmentation tasks. In this paper, GMMs are used to segment the given concert using CFCC features. Complete details about the training of GMMs, segmentation of the concert and its accuracy is explained in Section 5.

3. INTER AND INTRA ITEM CLASSIFICATION

As explained in Subsection 2.2, composition segment locations might probably indicate the end of an item. Figure 3 shows that a composition can be fragmented into number of segments owing to applauses. As a result, an item can have one or more composition segments. We first explain how the adjacent composition segments belonging to same the *raga* are merged into a single item using the pitch histograms. Following it we provide an algorithm used for classifying the inter-item and intra-item applauses for finding the number of items in a concert for archival purpose.

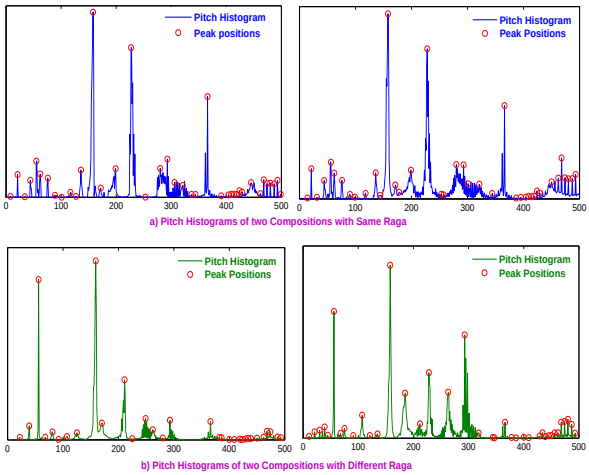

Figure 6: *Pitch histograms for same raga and different raga segments.*

Pitch is an auditory sensation in which a listener assigns musical tones to relative positions on a musical

scale based primarily on the frequency of vibration. *Raga* is one of the melodic modes used in Indian classical music. *Raga* uses a sequence of *svaras* upon which a melody is constructed [9]. However, the way the *svaras* are approached and rendered in musical phrases and the mood they convey are important in defining a *raga* than the *svaras* themselves. The note locations themselves are not absolute in carnatic music. Even for two *raga*s with the same set of notes, the rendered note locations can be distinct. As the purpose in this paper is not to identify the *raga*s but to detect the change in *raga*, pitch histograms are used [6]. Also, for the given task, it is irrelevant if the pitch corresponding to that of the lead artist or the accompanying artists is picked up.

Yin algorithm [5] is used for extracting the pitch for the segments. Pitch histograms for a segment show the relative notes sung by a musician [6]. Based on the pitch histograms, adjacent segments can be compared. Figure 6 (a) shows the pitch histogram for the adjacent segments belonging to the same *raga*, while Figure 6 (b) shows the pitch histogram for the adjacent segments that belong to different *raga*s. It can be observed that, in the pitch histograms corresponding to that of the same *raga* the positions of the peaks are more or less the same. On the other hand, for adjacent segments belonging to different *raga*s, the peaks are different. Clearly, the heights of the peaks are not necessarily be the same. Further, some peaks may be missing. This primarily means that in that specific segment, the musician's improvisation was restricted to a set of notes. Similarity between adjacent segments can be calculated by warping along the cent axis using dynamic programming [8]. If the similarity measure between the adjacent segments is above threshold 'α', then adjacent segments belong to different *raga*s otherwise they belong to the same *raga*. Finding a single threshold 'α' across all the concerts is crucial. A line search is performed and a threshold is chosen empirically. For the concerts in question, 'α' value '750' is chosen.

We now summarise the overall algorithm distinguishing inter-item and intra-item applauses as follows:

1. Applauses are identified for a concert using spectral domain features.
2. GMMs are built for Vocal solo, Violin solo, Composition ensemble and ThaniAvarthanam using CFCC based cepstral coefficients.
3. Audio segments between a pair of the applauses are labeled as Vocal solo, Violin solo, Composition, and ThaniAvarthanam using the trained GMMs.
4. The composition segments are located and the pitch histograms are calculated for the composition segments.
5. Similarity measure is computed on the warped histograms of the adjacent composition segments. If the similarity measure is above the threshold 'α', then the composition segment is inter-item otherwise, intra-item.
6. Based on the inter-item locations, intra-item segments are merged into the corresponding items. A concert is thus segmented into items for archival purposes.

Figure 7 shows the overall procedure for identifying the inter and intra items in a concert. In this Figure, the first column corresponds to the audio with the applause locations marked. The second column indicates the different segments based on applause locations. The third column corresponds to the labeling of all these segments into Vocal solo, Violin solo, Composition and ThaniAvarthanam. In the last column, some of the segments are merged based on the similarity measure. Observe that in Figure 7 two composition segments 4 and 5 are merged into single item 4 based on the distance measure.

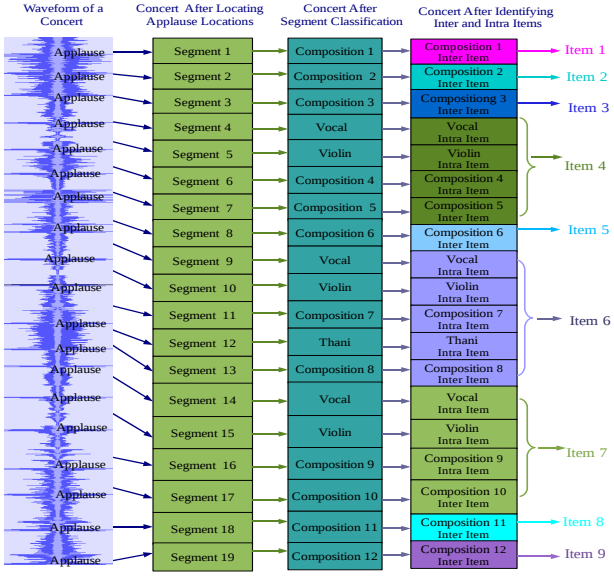

Figure 7: *Overall procedure for identifying inter and intra items in a concert.*

4. MAPPING THE COMPOSITION LIST TO ITEMS

In many performances, the meta-data is available for a concert in terms of composition lists from the audience [3]. The composition list obtained from the audience is matched to the items obtained from the algorithm discussed in Section 3. Figure 8 shows the mapping of the composition list obtained from the audience to inter-items for archival purposes. We evaluate 10 live recordings of male and female singers for mapping. The details of mapping and the mismatches (if any) are explained in Section 5.

5. EXPERIMENTAL EVALUATION

In this section we first give brief introduction to the database used for our study and then describe the experimental setup. We then provide the results obtained from various steps of the algorithm.

5.1 Database Used

For evaluation purpose, 50 live recordings of male and female singers are taken [4]. All recordings correspond to

[3] "http://www.rasikas.org"

[4] These live recordings were obtained from a personal collection of audience, musicians. These were made available for research purposes only.

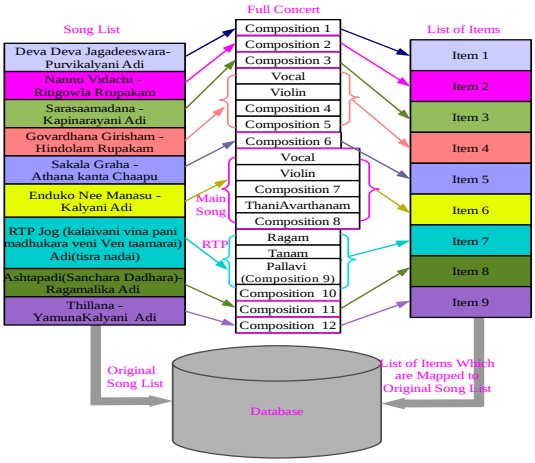

Figure 8: *Figure illustrates how the songs list is mapped to list of inter-items.*

concerts where the lead performer is a vocalist. Each concert is about 2-3 hours long. The total number of applauses across all the concerts is 990, where 394 applauses are inter-item applauses and 596 applauses are intra-item applauses. The 50 concerts consist of 394 items. All recordings are sampled at 44.1KHz sampling frequency with 16 bit resolution. For feature extraction, the analysis window chosen is 100 $msec$ and hop size is set to 10 $msec$ Table 1 gives the details of the database used. It can be observed that for a given singer, there are different possible tonic values, calculated using pitch histograms.

Singer Name	No. of Concerts	Duration (Hrs)	No. of Applause	Different Tonic
Male 1	4	12	89	158,148,146,138
Female 1	4	11	81	210, 208
Male 2	5	14	69	145, 148,150,156
Female 2	1	3	16	198
Male 3	4	12	113	145,148
Female 3	1	3	15	199
Male 4	26	71	525	140,138,145
Male 5	5	14	62	138,140

Table 1: *Database used for study, different Tonic values identified for each singer using pitch histograms.*

5.2 Experimental Setup

1. For all the 50 recordings spectral flux and spectral entropy features are extracted with a window size of 100 $msec$ with overlap of 10 $msec$. Applauses are marked for all the recordings using the Sonic-Visualizer [4]. These marked applauses are used as the ground truth for finding the applause detection accuracy.

2. To build the GMMs for the segmentation of the concert Vocal, Violin, ThaniAvarthanam, and Composition segments are manually segmented from the database. From male (female) recordings three segments are chosen for training the GMM for each class. MFCC, ChromaFCC, and CFCC features of 20 dimensions are extracted. GMMs with 32 mixtures are built for each of the four classes.

3. All the 50 recordings are segmented based on applause locations and all these segments are manually labeled as Vocal, Violin, ThaniAvarthanam, and Composition. Labeled data is used as the ground truth for finding the segmentation performance [5] . After segmenting the concerts based on the applauses, there are 990 segments in total and these segments are tested against the GMMs. All these segments are labeled using 1-best result.

4. For each concert, composition segments are highlighted using GMMs, and for these segments pitch is extracted using Yin algorithm [5]. These segments are merged into a single item if they belong to the same *raga* using the pitch histograms.

5. The performance measure used for evaluating each of the applause detection , segmentation performance and the inter-item, intra-item classification is

$$\text{Accuracy} = \frac{\text{Correctly Predicted Data}}{\text{Total Amount of Data}} \times 100\%$$

5.3 Experimental Results

The analysis of finding the end of item locations for all the concerts is done in the following three stages.

1. As explained in Section 3, first, applauses are located for using spectral domain features like spectral entropy and spectral flux. Table 2 shows the decision thresholds for applause and music discrimination and the applause detection accuracy based on the same thresholds. The applause detection accuracy is calculated at frame level.

Feature	Threshold range	Accuracy (%)
Spectral Flux (Nonorm)	0.2-1.0	85%
Spectral Flux (Peaknorm)	0.35-1.0	96%
Spectral Entropy	0.79-1.0	95%

Table 2: *Decision thresholds for applause and music discrimination and applause detection accuracy.*

2. Secondly, based on applause locations all concerts are segmented. The segments are labeled using GMMs with CFCC features. Table 3 shows the segmentation performance using MFCC, ChromaFCC, and CFCC features. Separate GMMs are built for male and female singers. Table clearly shows that CFCC features performs well in locating the composition segments in the concert. The overall segmentation accuracy for 50 concerts, including male and female recordings, is 95%.

Model	MFCC	ChromaFCC	CFCC
Male singers	78%	60%	90%
Female singers	92%	70%	97%

Table 3: *Segmentation performance using MFCC, ChromaFCC and CFCC.*

3. Finally, the composition segments in a concert are merged into a single item if they belong to the same *raga* using the pitch histograms. Table 4 shows the inter-item and intra-item classification accuracy for every singer. The overall classification accuracy for the 50 concerts is found to be 93%. We also evaluate our approach for mapping the songs list to the set of items which we obtain from the algorithm. Table 5 also shows the mapping

[5] Labeling was done by the first author and verified by a professional musician

performance. This mapping is evaluated for 10 live recordings of performances by six musicians for which songs lists were available.

Musician	Intra-item	Inter-item	Accuracy(%)
Male 1	65	24	90
Female 1	50	31	100
Male 2	37	32	100
Female 2	10	6	98
Male 3	75	38	93
Female	8	7	99
Male 4	317	227	93
Male 5	33	29	100

Table 4: *Inter and intra item classification accuracy.*

Musician	No. of concerts	Actual inter items	Predicted inter items	Mapping accuracy (%)
Male 1	1	6	7	96
Female 1	1	9	9	100
Male 2	1	7	7	100
Male 3	2	19	21	97
Male 4	2	13	14	97
Male 5	3	19	19	100

Table 5: *Performance of mapping the songs list to the set of items.*

In a concert every artist can optionally render a *ragamalika*, where a single item can be performed in different *raga*s. When a single composition is sung in *ragamalika*, the algorithm will still work, since the applause will be at the end of the item. On the other hand, for items like *viruttam* or *ragam tanam pallavi* with multiple *raga*s, owing to the extensive improvisational aspects in them, a number of applauses can occur intra-item. Such special cases are limits of the algorithm. Every concert may have one such item. For these types of items the lyrics could be same or the meter could be same. For such cases, the algorithm can be extended to include matching of the lyrics. The last item in a concert is a closure composition. This is identical in all concerts and is called the *mangalam*. When a musician continues without a pause before closure composition, the mangalam is combined with the penultimate item in the concert. This is not considered as an error in this paper.

6. CONCLUSION AND FUTURE WORK

In this work we proposed an algorithm for automatically segmenting continuous audio recordings of carnatic music into items for archival purpose. An item in a concert has a structure and each segment has specific spectral properties. These properties are exploited to label the segments. In particular, a new feature called CFCC performs well for labeling the segments when compared to MFCC and ChromaFCC features. Given meta-data in terms of names of items in a concert, the audio and the meta-data are aligned. The directions for future work will be, we need to evaluate the inter-item and intra-item classification technique for the special cases like *ragamalika*, *RTP*, *viruttam*, and a vocal solo (sung in isolation), where adjacent segments belong to different *raga*s but correspond to a single item. This requires additional analysis in terms of lyrics, meter which are quite nontrivial.

7. ACKNOWLEDGEMENTS

This research was partly funded by the European Research Council under the European Unions Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583).

8. REFERENCES

[1] Ashwin Bellur, Vignesh Ishwar, Xavier Serra, and Hema A. Murthy. A knowledge based signal processing approach to tonic identification in indian classical music. In *International CompMusic Wokshop*, Instanbul, Turkey, 2012.

[2] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*, chapter 7. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[3] B E Brodsky and B S Darkhovsky. *Non-parametric Methods in change-point problems*. Kluwer Academic Publishers, New York, 1993.

[4] C Cannam, C Landone, M Sandler, and J.P Bello. The sonic visualiser: A visualisation platform for semantic descriptors from musical signals. In *7th International Conference on Music Information Retrieval (ISMIR-06)*, Victoria, Canada, 2006.

[5] A De Cheveigne and H Kawahara. Yin, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, page 111(4):19171930, 2002.

[6] P Chordia and A Rae. Raag recognition using pitch-class and pitch-class dyad distributions. *In Proc. of ISMIR*, pages 431–436, 2007.

[7] Dan Ellis. Chroma feature analysis and synthesis. `http://www.ee.columbia.edu/~dpwe/resources/Matlab/chroma-ansyn`, 2007.

[8] Toni Giorgino. Computing and visualizing dynamic time warping alignments in r: The dtw package. *Journal of Statistical Software*, 31(7):1–24, 8 2009.

[9] T M Krishna and Vignesh Ishwar. Svaras, gamaka, motif and raga identity. In *Workshop on Computer Music*, Instanbul, Turkey, July 2012.

[10] M V N Murthy. Applause and aesthetic experience. http://compmusic.upf.edu/zh-hans/node/151, 2012.

[11] Padi Sarala, Vignesh Ishwar, Ashwin Bellur, and Hema A Murthy. Applause identification and its relevance to archival of carnatic music. In *Workshop on Computer Music*, Instanbul, Turkey, July 2012.

[12] J Serra, G K Koduri, M Miron, and X Serra. Tuning of sung indian classical music. *In Proc. of ISMIR*, pages 157–162, 2011.