Sādhanā (2018) 43:154
https://doi.org/10.1007/s12046-018-0922-y

Segmentation of continuous audio recordings of Carnatic music concerts into items for archival

SARALA PADI* and HEMA A MURTHY

Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India
e-mail: padi.sarala@gmail.com; hema@cse.iitm.ac.in

MS received 9 June 2016; revised 30 June 2017; accepted 8 March 2018; published online 10 August 2018

Abstract. Concert recordings of Carnatic music are often continuous and unsegmented. At present, these recordings are manually segmented into items for making CDs. The objective of this paper is to develop algorithms that segment continuous concert recordings into items using applause as a cue. Owing to the 'here and now' nature of applauses, the number of applauses exceeds the number of items in the concert. This results in a concert being fragmented into different segments. In the first part of the paper, applause locations are identified using time, and spectral domain features, namely, short-time energy, zero-crossing rate, spectral flux and spectral entropy. In the second part, inter-applause segments are merged if they belong to the same item. The main component of every item in a concert is a composition. A composition is characterised by an ensemble of *vocal (or main instrument), violin (optional) and percussion*. Inter-applause segments are classified into three segments, namely, *vocal solo*, *violin solo*, *composition* and *thaniavarthanam* using tonic normalised cent filter-bank cepstral coefficients. Adjacent composition segments are merged into a single item, if they belong to the same melody. Meta-data corresponding to the concert in terms of items, available from listeners, are matched to the segmented audio. The applauses are further classified based on strength using Cumulative Sum. The location of the top three highlights of every concert is documented. The performance of the proposed approaches to applause identification, inter-applause classification and mapping of items is evaluated on 50 live recordings of Carnatic music concerts. The applause identification accuracy is 99%, and the inter- and intra-item classification is 93%, while the mapping accuracy is 95%.

Keyword. Cent filter-bank cepstral coefficients; segmentation of concerts; applause detection; classification of music segments.

1. Introduction

Applause is an indication of the appreciation given by the audience to the performer. Applause locations in a football video enable the determination of the location of the highlights of the game. In many genres of music (Jazz, Indian music) applause is a medium of communication between the audience and the performer. In a Western music concert, the audience applauds the artist only at the end of a concert or at most at the end of an item. In Indian classical music, 'applauses are more or less here and now', in that the audience applauds the artist not only at the end of an item but also intra-item or within-item [1].

Indian classical music comes in two different forms, namely, Carnatic music and Hindustani music. Carnatic music is popular in the Southern part of India whereas Hindustani music is more popular in northern Karnataka, Maharastra and north of Vindhyas. Carnatic music is made up of two components, namely, *kalpita sangita* and *kalpana sangita* [2]. *Kalpita sangita* in a concert corresponds to fixed *compositions*, to be performed as composed or taught, while *kalpana sangita* corresponds to improvisation. *Kalpana sangita* is often referred to as *manodharma sangita*, 'a musical aspect where the creativity of the musician plays a significant role' [2]. In a concert, the performer generally illustrates the various nuances of a *rāga* by means of the following: *Ālāpana, tānam, kalpana svaras* and a *kīrtana* (composition composed by a composer). The *Ālāpana, tānam* and *kalpana svaras* are the improvisational aspects of the concert, whereas the *kīrtana* is the fixed *composition*[1]. In a concert (*kutcheri*), applauses can be heard after each of the improvisational aspects, and at the end of *kīrtana*. The audience occasionally applauds the artist in-between an improvisational piece or even during the rendition

The work presented in this paper is an extension and includes extensive analysis on Carnatic music.

*For correspondence

[1] In this paper we refer to the *kīrtana* as a *composition*.

of a *kīrtana*, especially when the musician aesthetically blends the lyrics of the *kīrtana* with that of the melody.

In the literature, different features and techniques are used to detect applauses along with other events like speech, music, cheer and laughter for content-based retrieval. To detect these events in an audio, various time and spectral domain features are used in tandem with machine learning methods [3–6]. These events play a significant role in golf, baseball and soccer games (or for that matter any sport) for extracting the highlights [3–6]. Applauses are also used to segment speech meetings for archival and content-based retrieval purposes [7–9].

Event-based analysis is popular in audio and video for identifying landmarks. In this paper too, an attempt is made to automatically segment the continuous audio recordings of Carnatic concerts into items using applauses as a cue. Applauses locations are first obtained using low-level time- and spectral-domain features. The number of applauses is much larger than the number of the items performed in a concert. This leads to the fragmentation of a item. The various fragments that make an item must be merged. Another cue is used to merge the fragments that make up the same item. In every item in a concert, a composition is rendered. The composition is rendered with accompaniment (violin and percussion). Inter-applause segments are therefore classified as *vocal solo*, *violin solo*, *composition* and *thaniavarthanam* segments. Adjacent composition segments are merged if they are rendered in the same *rāga*. Reviews of concerts are often available in social media (http://www.rasikas.org). In particular, the first phrase of the composition, composer and tala information is available. The list provided by listeners is matched to the segmented concert. Additionally, applauses are further characterised by strength. Strong applauses correspond to the highlights of a concert.

The rest of the paper is organised as follows. In section 2, to give perspective to the work, we review some of the important concepts in Carnatic music that are relevant for this paper. Section 3 discusses the features and methods used for detection of applauses in Carnatic music. As *compositions* in Carnatic music are tuned to a particular *rāga* and can be performed in a key that is chosen by the performer, a novel feature called cent filter-bank-based cepstral coefficients (CFCC) is developed in section 4 to differentiate various segments in a concert. Algorithms based on matching pitch histograms are developed in section 5 to merge adjacent composition segments that belong to the same item. Section 6 discusses the semantic inferences that can be drawn from the foregoing study on applauses.

2. Carnatic music: a brief introduction

Sruti (tonic), rāga, gamakā and *tala* are the basic elements of Carnatic music. *Tonic* refers to the pitch chosen by the performer for a concert. *Rāga* is the melodic component, while *tāla* corresponds to the rhythmic component. *Rāga* is equivalent to 'mode' or 'scale' in Western classical music, whereas *tāla* is equivalent to 'rhythm.' Unlike Western music, a *rāga* in Carnatic music is not made up of a scale of fixed pitches, but is made up of *svaras*, which are *notes* that are embellished by inflections (or *gamakās*) [10]. The *rāga* is defined relative to the *tonic*. A *composition or kīrtana* in Carnatic music is based on a specific *rāga* or *rāgas* (as in a *rāgamālika*) and is set to a specific rhythm (*tāla).*

A *rāga* in Carnatic music corresponds to the sequence of *svaras*. Although the written *svaras* of two *rāgas* may be identical, the phraseology used is what differentiates them. The scale used in any *rāga* need not be linear [11].

A *tāla* corresponds to the rhythm that is associated with a specific item. *Tāla* also comes in many forms. For example, *chatursra jati adi tāla* corresponds to an eight-beat cycle. In addition, *nadai* is also another variant. Depending on the *nadai*, the number of aksharas that make up a beat can vary. The standard adi *tāla* is made up of 4 aksharas per beat, a *tisra nadai adi tāla* will consist of 3 aksharas per beat.

2.1 *A concert in Carnatic music*

The duration of a concert is about 2–3 hours long, and the artist renders between 8 and 10 items in a concert. The lead performer chooses a specific tonic and all accompanying instruments also tune to the same tonic. Figure 1 shows a finite state diagram that takes an applause-centric view (the focus of this paper). Although the most complex item can consist of a *vocal alapana*, a *violin alapana*, a *composition*, a *niraval*, a *svaraprasthara* and a *thaniavarthanam*, applauses can occur also within each of these segments, thus dividing the segments further. In this figure, we assume that a *composition* segment corresponds to not only the *composition* but also any improvisation that includes the ensemble of vocal/violin and percussion. It can be seen from the figure that a *composition* segment is mandatory in an item. A single *composition* would correspond to the rendition of the *kirtana* without any improvisation. A concert is replete with applauses. As most Indian music is improvisational, the audience applauds the artist

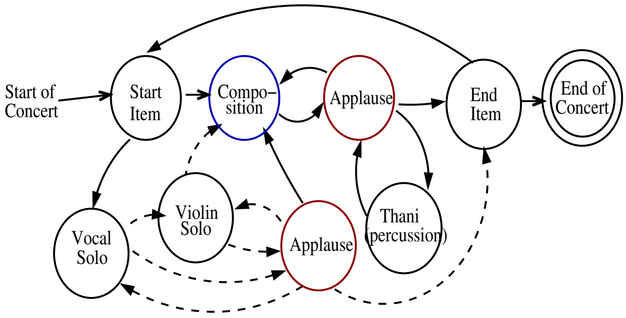


Figure 1. General structure of Carnatic music concert. Dotted lines indicate that certain segments are optional and solid lines indicate mandatory segments.

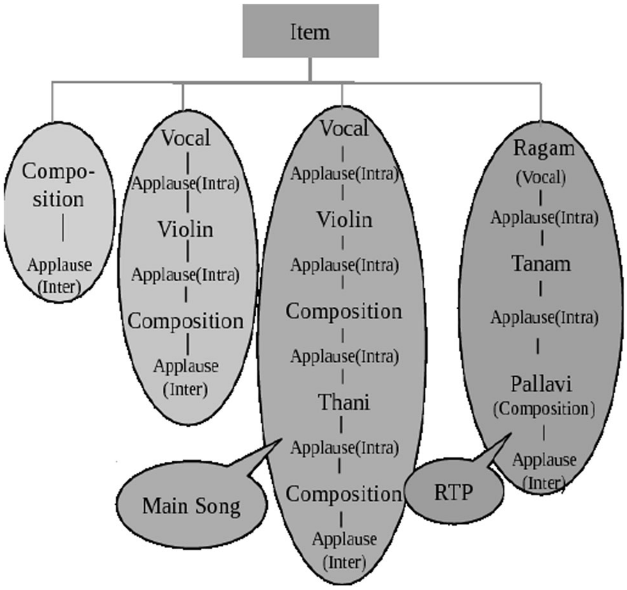


Figure 2. General structure of an item in Carnatic music concert.

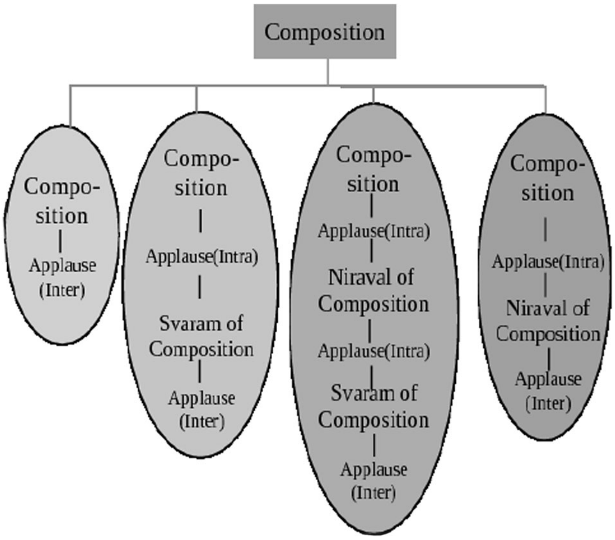


Figure 3. Structure of composition in an item in a concert.

spontaneously. 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 [1]. Occasionally, applauses are missing after some segments. This is indicated by the dotted lines in the figure.

There are many segments in an item with respect to Carnatic music that lend themselves to improvisation. *Ālāpana* is a segment where the vocalist elaborates the *rāga* with his/her imagination [12] or *manodharma*. *Violin solo* is similar to the *vocal solo*, where the violinist explores the *rāga*. A good violinist generally attempts to give a sketch that is more or less similar to that of the vocalist. This is generally about half the duration of the *ālāpana* by the vocalist (or main performer).

The mridangam, ghatam are pitched percussion instruments and add colour to the segment that is rendered. The ability of a percussion vidwan to emphasise the nuances of the vocalist through a rendering that blends with the lyrics of the *composition* or the *svaras* of the *niraval* adds aesthetics to a concert. In addition to this, the percussion artist also performs independently during a segment call *thaniavarthanam*. Mridangam is the main accompanying instrument, while the others are considered subordinate to the mridangam. The mridangam, ghatam, kanjira, morsing artists explore various aspects of the rhythm during this segment. This segment also involves significant improvisation.

Composition is an important segment in a concert and it can be thought of as an ensemble where the *vocal solo*, the *violin solo* and the *percussion* are present simultaneously[2]. An item in a vocal Carnatic music concert can have different structures and it generally ends with the *composition* segment. In fact, an item can occasionally have no *composition* segment as in a *viruttam or slokam*. Although this is indicated in figure 1, in this paper we assume that a *composition* segment is present in every item. The composition is rendered by the artist set to the *tāla* and *rāga* as designed by the composer albeit in the tonic of the artist. There are sections in the *composition* that can also be improvised, namely, the *niraval* and *svaraprasthara*, and therefore can lead to applauses.

Figure 2 shows the general structure of an item in Carnatic music. As indicated in the figure, within an item a *composition* may be further fragmented owing to the applauses. This is especially true for the *MainSong* in a concert, where different parts of the *composition* lend themselves to significant improvisation. Each fragment is referred to as a *composition* segment, hereafter. Since *composition* segments are mandatory in every item, the beginning and end of an item can be found in the vicinity of *composition* segments.

Figure 3 shows the structure of *composition* segment in a concert. From figures 2 and 3, observe that applauses can occur after every item, every segment and also within a segment. Therefore, applauses are classified into 3 types, namely, inter-item, intra-item (within item) and within segment.

2.2 *Tonic in Carnatic music*

Tonic is the basic pitch chosen by the performer. The lead performer chooses a particular tonic and accompanying artists also tune their instruments to this tonic. A *tambura*[3] is

[2] Hereafter we refer to this as a (the main) song.

[3] An instrument that maintains the tonic throughout the concert.

Table 1. Comparison of frequencies of the 12 notes for Indian and Western classical music traditions.

	Note	Natural (Indian)	Frequency (Hz-C-4)	Ratio Natural scale	Ratio scale	Harmonic (Western) *ratio* = $2^{(1/12)}$
1	S	1	261.63	1	1	261.63
2	R1	16/15	279.07	1.067	1.059	277.19
3	R2/G1	9/8	294.33	1.125	1.122	293.69
4	R3/G2	6/5	313.96	1.200	1.189	311.16
5	G3	5/4	327.04	1.250	1.260	329.68
6	M1	4/3	348.84	1.333	1.335	349.19
7	M2	17/12	370.64	1.417	1.414	370.08
8	P	3/2	392.45	1.500	1.499	392.00
9	D1	8/5	418.61	1.600	1.588	415.32
10	D2/N1	5/3	436.05	1.667	1.682	440.00
11	D3/N2	9/5	470.93	1.800	1.782	466.22
12	N3	15/8	490.56	1.875	1.888	493.96
13	(S)	2	523.26	2	2.000	523.35

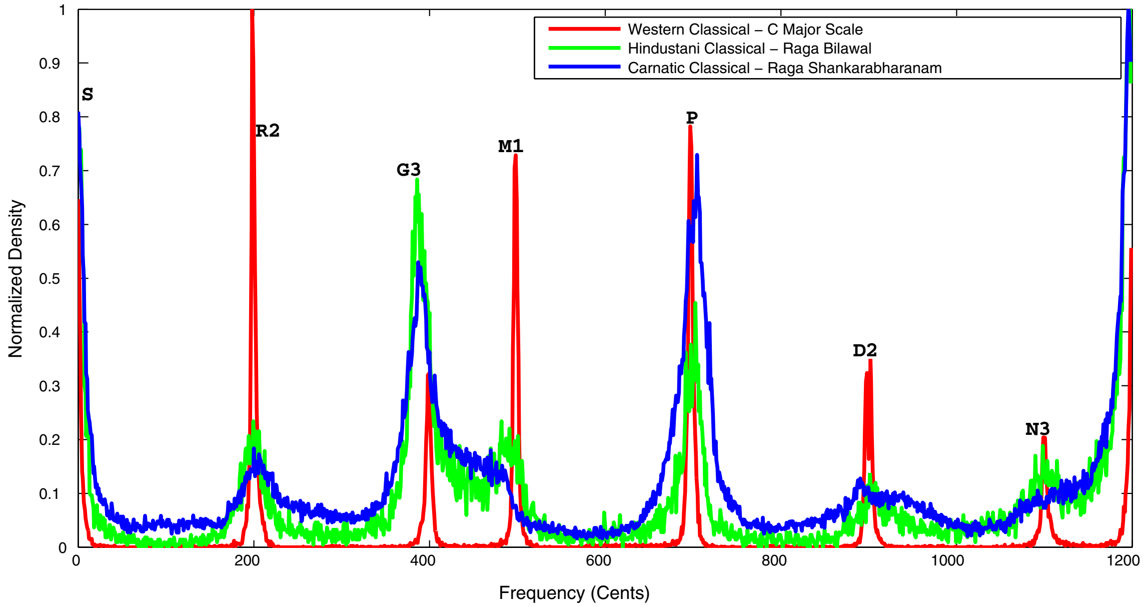


Figure 4. Comparison of C-Major (Western), bilawal (Hindustani) and sankarabaranam (Carnatic) pitch histograms (reproduced from [14] with permission.

tuned and is the first instrument that is strummed and heard in a concert. The strumming of the *tambura* provides a common point of reference to both the performer and the listener. Once the tonic is defined, the scale is defined. Indian music is based on the natural scale, while Western music is based on the equitemperament scale, where adjacent notes have the same frequency ratio [13]. Melodic analysis of a Western classical music is normally based on ‘a quantised representation of pitches and durations within a well-defined framework of possible relationships’ [13]. Table 1[4] compares Indian and Western music scales. From the scale, it can be seen that the semitone locations are quite different in Indian classical music and Western classical music scales. In addition to this, *gamakās* in Carnatic music add colour to the rendition of a *composition* and thus the frequency associated with a *svara* can span even two adjacent *svaras*. Figure 4 shows the tonic normalised pitch histograms of melodies that use the same scale in Western, Hindustani and Carnatic music (C-Major, bilawal and sankarabaranam, respectively). Observe that the pitches are more or less discrete in Western classical music, and also observe that the pitch histogram corresponding to Carnatic music shows significantly more inflection when compared with Hindustani music.

[4] This table is Courtesy: M V N Murthy, Professor, IMSc, Chennai, India.

Table 2. The 12-semitone scale for four different musicians.

Sl no.	Carnatic music svara	Label	Frequency ratio	Carnatic music scale for different tonic values (tonic in Hz)			
				138	156	198	210
1	*Shadja (Tonic)*	*S*	1.0	*138*	*156*	*198*	*210*
2	Shuddha rishaba	R1	(16/15)	147.20	166.40	211.20	224
3	Chatushruthi rishaba	R2	(9/8)	155.250	175.50	222.75	236.25
4	Shatshruthi rishaba	R3	(6/5)	165.250	187.20	237.60	252
3	Shuddha gAndhara	G1	(9/8)	155.250	175.50	222.75	236.25
4	ShAdhArana gAndhara	G2	(6/5)	165.60	187.20	237.60	252
5	Anthara gAndhara	G3	(5/4)	172.50	195.0	247.5	262.5
6	Shuddha madhyama	M1	(4/3)	184.0	208.0	264.0	280
7	Prati madhyama	M2	(17/12)	195.50	221.0	280.5	297.5
8	Panchama	P	(3/2)	207.00	234.0	297.0	315
9	Shuddha daivatha	D1	(8/5)	220.80	249.60	316.8	336
10	Chatushruthi daivatha	D2	(5/3)	230.00	260.0	330.0	350
11	Shatshruthi daivatha	D3	(9/5)	248.40	280.80	356.4	378
10	Shuddha nishAdha	N1	(5/3)	230.0	260.0	330.0	350
11	Kaisika nishAdha	N2	(9/5)	248.40	280.80	356.4	378
12	KAkali nishAdha	N3	(15/8)	258.75	292.50	371.25	393.75

As mentioned earlier, the tonic chosen for a concert is maintained throughout the concert using an instrument called the 'tambura' [15]. To analyse Carnatic music, it is important that the notes are tonic normalised; otherwise, the frequencies occupied by the same note can vary with tonic as indicated in table 2. Table 2 gives the frequencies of the 12 semitones used in Carnatic music for four different tonics (138, 156 Hz corresponding to male artists and 198, 210 Hz corresponding to that of female artists). As explained in the paper [16], pitch histograms of the segments of the same *rāga* look similar on the cent scale. Therefore, any melodic analysis of Indian music requires a normalisation with respect to the tonic.

3. Applause detection

As explained in section 1, owing to the audience participation in a concert, applauses can be used to automatically segment a concert into items. In this section, we detail the experiments performed to distinguish between music and applause segments. As explained in [5], both time-domain and spectral-domain features can be used to discriminate between music and applause. In [17] applause detection was performed using spectral-domain features using an empirical rule-based approach. In this paper, both spectral-domain and time-domain features are used for detecting applause locations using both empirical rules and supervised machine learning (Gaussian mixture models (GMMs) and Support Vector Machines (SVMs)).

Figure 5 corresponds to that of the waveform and corresponding spectrogram of four segments, namely, *composition*, *vocal solo*, *violin solo* and *thaniavarthanam* followed by an applause. From figure 5, it can be observed that applause and music both are temporally and spectrally distinct. It can also be inferred that applause segments have similar structure across concerts, and music segments significantly vary based on singer voice, melody and accompaniment. In the case of music segments, vertical bars correspond to either voiced components or onsets due to percussion. In the case of applause segments, the vertical bars correspond to synchronisation in clapping and have no voiced components. Short-time energy (STE), zero-crossing rate (ZCR), spectral entropy (SE) and spectral flux (SF) features are used to distinguish between music and applause segments. An overview of time- and spectral-domain features that enable distinction between applause and music is now presented. The general framework for feature definition is adapted from [18].

- STE: The time-dependent STE of an audio signal is defined as

$$S_n = \sum_{m=--\infty}^{\infty} (x[m])w[n-m])^2 \tag{1}$$

  where $w[n]$ is a Hamming window function. used. Figure 6 shows the waveform of music and applause segments and their corresponding STE values. In figure 6, the amplitude gradually increases for applause, due to synchrony in clapping, reaches a peak and tapers off again. In the case of music segments, the amplitude continuously fluctuates between low and high energy values. As a result, when average energy is calculated for a window of size 100 ms, it is high for an applause segment and low for music segments.
- ZCR: ZCR is defined as the rate at which zero crossings occur. It is a simple measure of the frequency content of a signal. The *ZCR* is defined as

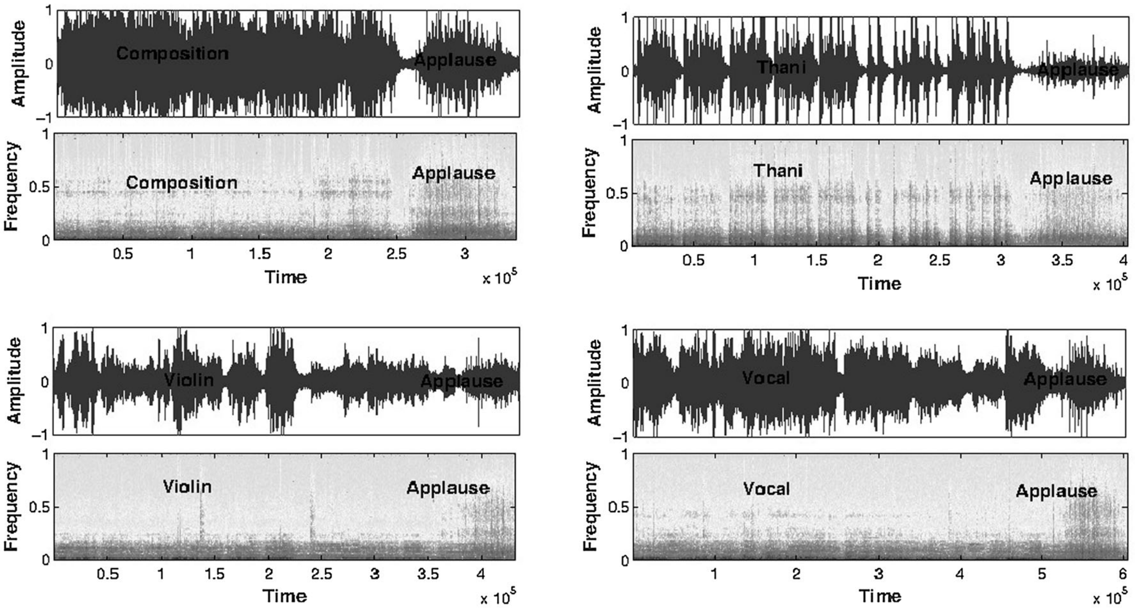


Figure 5. Waveform and spectrogram representations of applause following vocal solo, violin solo, composition and thaniavarthanam segments.

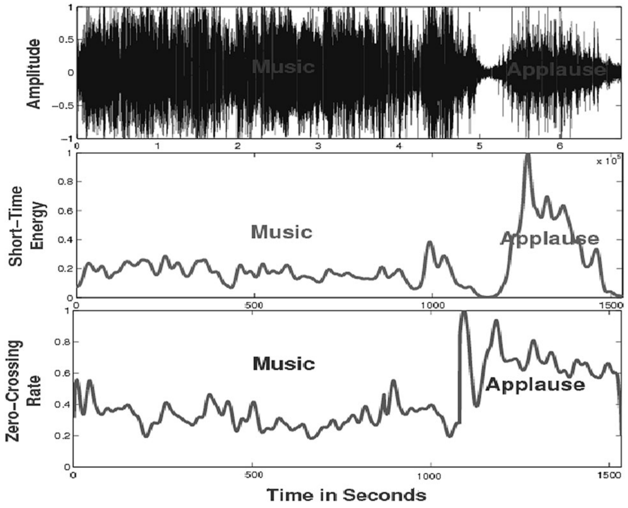


Figure 6. Time domain measures for determining applause positions.

$$Z_n = \sum_{m=-\infty}^{\infty} \mid sgn(x[m]) - sgn(x[m-1]) \mid w[n-m]$$

$$\text{where } sgn(x[n]) = \begin{cases} 1, & \text{x[n]} \geq 0 \\ -1, & \text{x[n]} < 0 \end{cases} \tag{2}$$

$x[n]$ is a discrete time audio signal and $w[n]$ is a window function. As applause segments are relatively unstructured compared with that of music segments, the ZCR is much higher for applause in comparison with that of voiced segments (figure 6) [17].

- SF: SF determines the change in spectra between adjacent frames. It is defined as the 2-norm of two adjacent frames:

$$SF[n] = \int_{\omega} (\mid X_n(e^{j\omega}) \mid - \mid X_{n+1}(e^{j\omega}) \mid)^2 d\omega \tag{3}$$

where $X_n(e^{j\omega})$ is the Fourier Transform of the input signal and it is defined as

$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} w[n-m]x[m]e^{-j\omega m}. \tag{4}$$

$X_n(e^{j\omega})$ is the Fourier Transform of the input signal and $|XNorm(e^{j\omega})|^2$ is $|X_n(e^{j\omega})|^2$ after peak normalisation. Reference [17] discusses spectral domain features and different normalisations of spectral flux feature for discriminating applause and music segments. As explained in [17], SF with peak normalisation was found to be the best. It is defined as follows:

$$|XNorm_n(e^{j\omega})|^2 = \frac{|X_n(e^{j\omega})|^2}{\max_{\omega}(|X_n(e^{j\omega})|^2)}. \tag{5}$$

As shown in figure 8, the SF values for applause segments are very high compared with music segments. This effect is further accentuated in figure 7 after normalisation. This is because music has redundancy while applause is not predictable, which leads to a relatively flat applause spectrum [17]. The loudness of the applause is characterised by the amplitude of the spectral components. Normalisation by the spectral peak makes all applause segments similar.

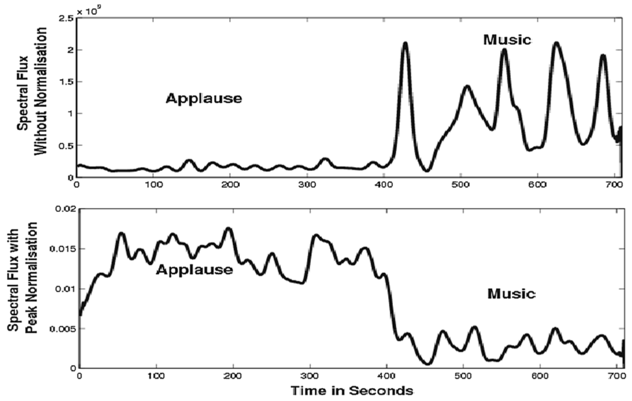


Figure 7. Spectral flux without normalisation and with peak normalisation measures for applause and music segments.

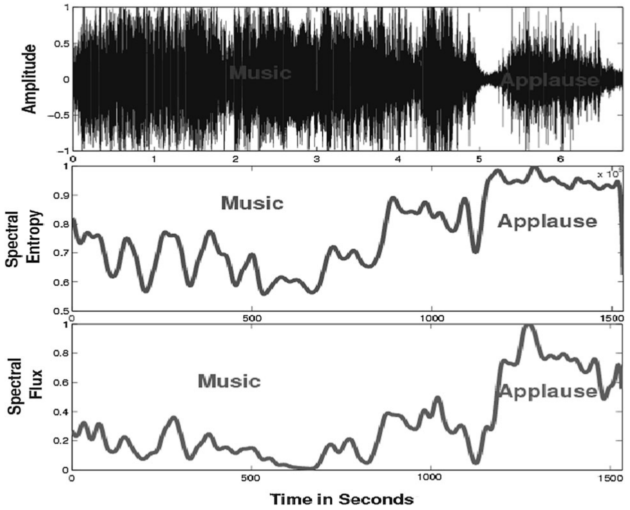


Figure 8. Spectral domain measures for determining applause positions.

- SE: SE is a measure of randomness of a system. Shannon's entropy of a discrete stochastic variable $X = \{X_1, X_2, ..., X_N\}$ with probability mass function $p(X) = \{p_1, p_2, ..., p_N\}$ is given by

$$H(X) = -\sum_{i=1}^{N} p(X_i)\log_2[p(X_i)]. \tag{6}$$

The power spectrum can be thought of as power spectral density. The entropy of the power spectral density function is given by

$$\begin{aligned} PSD_n(\omega) &= \frac{|X_n(e^{j\omega})|^2}{\int_\omega |X_n(e^{j\omega})|^2 d\omega} \\ SE[n] &= -\int_\omega PSD_n(\omega)\log PSD_n(\omega)d\omega \end{aligned} \tag{7}$$

where $X_n(\omega)$ is the Fourier Transform of the input signal and is defined in Eq. (5). As shown in figure 8, the SE value is high for applause compared with that of music segments.

Table 3. Decision thresholds for applause and music discrimination.

Feature	Alpha (α)	Threshold range
Short-time energy	1.9	0.29–1.0
Zero-crossing rate	2.5	0.49–1.0
Spectral flux (peak norm)	2.5	0.35–1.0
Spectral entropy	1.25	0.79–1.0

Table 4. Parameters used for techniques based on machine leaning.

Method	Parameters
SVM	Gaussian kernel width $\sigma = 10$
GMM	Num. mixtures for applause = 2 Num. mixtures for music = 32

3.1 *Classification of applause and music*

This section explains the methods and techniques used for classifying music and applause in Carnatic music concerts using empirical rules or rules based on statistical learning.

- Applause detection using thresholds: As STE, ZCR, SE and SF features are one-dimensional features, a single parameter α is chosen for each feature. The threshold is obtained as follows:

$$\text{threshold} = \alpha \times \text{mean of each recording } (\mu).$$

Table 3 shows the 'α' values for every feature and the approximate threshold ranges for every feature; 'α' is fixed for every feature, while the threshold varies based on the value μ for every concert.

- Applause detection using machine learning: Threshold-based methods are simple to implement, but choosing α is crucial. To address this issue, a machine learning framework is used for classifying music and applause. During training, data from different environments are chosen. Machine learning techniques learn well on the average and this property is exploited here. Machine learning techniques like GMMs and SVMs are used to discriminate applause and music segments using time- and spectral-domain features (STE, ZCR, SF and SE). Carnatic music consists of different segments like *vocal solo*, *violin solo*, *composition* and *thaniavarthanam*. Since applauses across concerts have similar structure, two mixtures are adequate to represent applause, while owing to variations across music segments, it is best represented by 32 mixtures. GMM is a generative model and SVM is a discriminative model. Since the given problem consists of only two classes, SVMs are also

Table 5. Amount of train data used for applause and music.

Class	No. of segments	Duration of segments
Applause	40 segments	4–5 s each (8 min)
Music	vocal solo, violin solo composition, thaniavarthanam (ghatam, kanjira, morsing)	2 segments each with 30 s (8 min)

explored. Parameters and amount of data chosen for GMM and SVM are given in tables 4 and 5.

3.2 *Applause detection results*

The database consists of the following: (a) 50 personal live recordings of male and female singers[5], (b) published music collections like *Charsur*[6], (c) *ARKAY* recordings[7] and (d) some on-line collections like *Sangeethapriya*[8]. All recordings correspond to 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 1131 and all recordings are sampled at 44.1 kHz sampling frequency with 16-bit resolution.

Tables 6 and 7 give details of the databases used for applause detection. The *Charsur* database is already segmented into items. Hence many of the recordings do not contain applauses. Thus only some recordings that have applauses are taken for experimentation. Similarly selected recordings that have applauses are taken from *Sangeethapriya* and *ARKAY* music collections. Tables 6 and 7 show the number of applauses for each database. When this work was started only the personal collections were available, so most of the experiments were made on data obtained from table 6. Most of the analysis, namely, training and evaluation of machine learning models and determination of thresholds, was performed on the database given in table 6. The trained models (GMM and SVM) and the thresholds for time- and spectral-domain features are used to segment the new recordings given in table 7 to ensure scalability across recordings.

Initially, for 70 recordings, applause locations were marked manually using the Sonic-Visualiser [19] tool. It was observed that applause duration varies from 3 to 36 s. Time-domain and spectral- domain features (STE, ZCR, SE and SF) are extracted for 70 recordings with a window size of 100 ms (4410 samples), with an overlap of 10 ms (441 samples). The overlap of 10 ms is chosen for better resolution in detecting the applauses. For training the GMM and SVM classifiers, applauses and music are manually segmented from some of the live recordings tabulated in table 6. Table 5 shows the amount of data used for applause and music. The time-domain and spectral domain features (STE, ZCR, SE and SF) features are extracted for applause and music and smoothed using a rectangular window of size 15 for training the GMM and SVM classifiers. The experiments for SVM classifiers are performed using the libsvm tool [20]. The performance of applause and music discrimination is calculated using the relation

$$\text{accuracy} = \frac{\text{correctly identified frames}}{\text{total number of frames}} \times 100\%.$$

All the features are smoothed and normalised between 0 and 1. Table 8 shows the classification accuracy for applause and music using rule-based and machine-learning-based methods. From table 8, it can concluded that machine-learning-based approaches are more consistent. From table 8, we can observe that the SVM-based method with ZCR, STE, SE and SF features are not only consistent but also robust across different databases.

Table 6. Database used for study and different tonic values identified for each singer using pitch histograms.

Singer name	No. of concerts	Duration (h)	No. of applauses	Different tonics for each musician (Hz)
Male 1	4	12	89	158, 148, 146, 138
Female 1	4	11	81	210, 208
Male 2	5	14	68	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	546	140, 138, 145
Male 5	5	14	62	138, 140

Table 7. Other databases used for applause detection.

Database name	No. of recordings	Duration (h)	No. of applauses
Charsur	14	8	60
Sangeethapriya	4	9	50
ARKAY recordings	2	5	31

4. Segmentation of Carnatic music concert

A typical Carnatic music concert can be made up of about 10–12 items, while it can be segmented into 30 fragments due to applauses. These fragments have to be first labelled

[5] These live recordings were obtained from personal collections of listeners and musicians. Recordings were made available for research purposes only.

[6] http://www.charsur.org. Live concerts have been licensed from Charsur for research purposes.

[7] About 40 concerts are available at the time of this writing.

[8] http://www.sangeethapriya.org

Table 8. Applause and music classification accuracy on all the databases.

Feature name	Personal recordings			Charsur			Sangeethapriya			ARKAY		
	Threshold	GMM	SVM	Threshold	GMM	SVM	Threshold	GMM	SVM	Threshold	GMM	SVM
STE	85.00	98.34	97.65	84.95	99.01	99.14	83.58	99.25	99.31	82.27	99.48	99.48
ZCR	98.00	98.43	98.68	99.14	95.45	99.14	95.20	93.56	99.31	96.12	95.55	99.48
SE	95.00	90.40	96.00	98.46	98.65	99.14	93.72	94.41	99.31	97.67	98.70	99.48
SF	96.00	98.63	98.68	95.36	99.14	99.14	92.52	99.37	99.31	93.76	99.48	99.48

before fragments that belong to the same item are merged. Using ground-truth data, MFCC features were used to build GMM models. GMM models are trained for each type of segment. When MFCCs were used as features, it was observed that a small amount of training data from each concert and each type of segment was required. This works well for concerts seen during training but performs poorly for unseen data. This is primarily because the tonic varies across concerts. A common frequency range is required for all the concerts irrespective of tonic. To address this issue, chroma filter-banks are used, especially for Western classical music [21, 22]. In chroma filter banks, the *key* is used to normalise the frequency range and the frequency range is folded to a single octave (0–1200 cents):

$$chroma\,frequency = 1200\log\left(\frac{f}{key}\right)\%1200$$

where % corresponds to the modulo operation[9]. Chroma filter banks discussed in [22] are primarily designed for Western classical music. To account for the continuous pitch histograms observed in Carnatic music, chroma filter banks that are non-overlapping are replaced by a set of overlapping filters. Cepstral coefficients derived using the Chroma filter banks are referred to as Chroma filter-bank-based cepstral coefficients (ChromaFCC). Even though chroma filter banks are normalised with respect to tonic, it is not adequate for Indian music. The folding over to a single octave is not appropriate, as the characteristic traits of the musician/melody may be observed in a particular octave or across octaves. The poor performance on segmentation using MFCC features suggests that cent filter banks, which are normalised with respect to the tonic, are required to capture the note positions. This paper proposes a feature called CFCC that seems to show promise for various tasks in Carnatic music.

4.1 *Cent filter-bank feature extraction*

Cent filter-bank feature extraction is similar to MFCC extraction. To estimate MFCC, the frequencies are mapped to the mel scale; in the case of CFCC, the frequencies of the music signal are transformed to the cent scale using the tonic (Eq. (8)). A flow chart of the algorithm for estimating CFCC is given in figure 9. The CFCC are then used to build the GMM models.

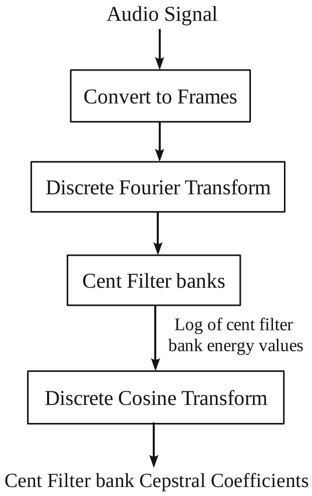

Audio Signal
Convert to Frames
Discrete Fourier Transform
Cent Filter banks
Log of cent filter bank energy values
Discrete Cosine Transform
Cent Filter bank Cepstral Coefficients

Figure 9. Cent filter-bank energy feature extraction.

$$\mathrm{centscale} = 1200\log_2\left(\frac{\mathrm{f}}{\mathrm{tonic}}\right). \tag{8}$$

4.1a *Tonic identification*: Tonic identification Extraction of CFCC features requires accurate estimation of tonic. Significant work has been done on tonic identification in Carnatic music using pitch histograms, group-delay method, multi-pitch method and Non-negative Matrix Factorisation (NMF) techniques [15, 16, 23]. The variants are primarily based on the amount of data available for tonic estimation. In the proposed task, as the entire concert is available, pitch histograms are adequate for estimating tonic [16]. A concert is divided into six parts and pitch histograms are computed for each part. Reference [24] is used to extract pitch framewise. Although [24] works better for monophonic music, in polyphonic music, the estimated pitch corresponds to that of the prominent voice in that frame. The pitch histograms of each of the six parts are multiplied and a single histogram is obtained. The melody across the six parts varies significantly; the *svara* common to all the segments corresponds to the tonic. The

[9] Key corresponds to tonic in the context of Indian classical music.

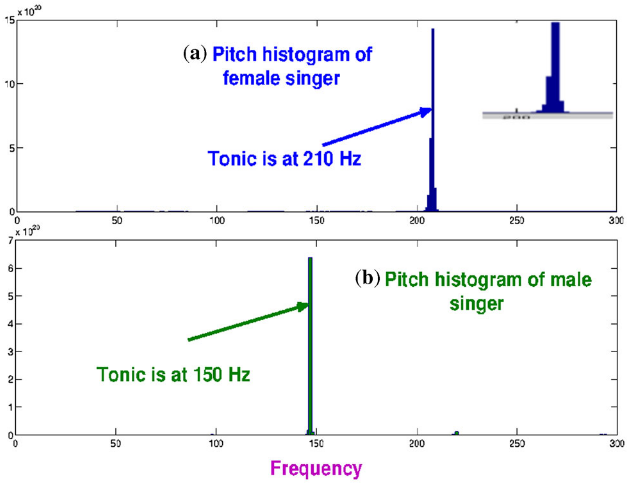


Figure 10. Composite pitch histogram for a female and a male singer.

tonic frequency is the location of the maximum peak. Figure 10 shows the pitch histogram for a female and male singer. The tonic identification accuracy using pitch histograms is 100% for the database given in table 6. Typical tonic values for each of the different musician in the database are given in table 6.

4.1b *Importance of CFCC feature for Carnatic music*: The significance of cent filter banks for Carnatic music is explained with an example in figures 11 and 9. Figure 11 shows the filter-bank coefficients and figure 9 shows the filter-bank energies of the same melodic motif sung by 7 different musicians (male and female singers)[10]. In figure 11, the filter banks are placed based on the mel scale as defined in Eq. (9). In figure 9, the filter banks are placed based on the cent scale as defined in Eq. (10). As can be seen, in the case of the mel scale, the harmonics of a motif sung by different musicians are emphasised based on the tonic of the composition segment. For example, in the case of DKP, the first harmonic is at the 60th filter, second harmonic is at the 120th filter and so on. In the case of MSS, the first harmonic is at the 75th filter, second harmonic is at the 150th filter and so one. In the case of *KVN*, the first harmonic is at the 50th filter, second harmonic is at the 100th filter and so on. The locations of the harmonics are highlighted in the melodic motifs of DKP, MSS and KVN.

$$\mathrm{Melscale} = 2595 \log_{10}\left(1 + \frac{\mathrm{f}}{700}\right), \tag{9}$$

$$\mathrm{centscale} = 1200 \log_2\left(\frac{\mathrm{f}}{\mathrm{tonic}}\right). \tag{10}$$

Figure 12 shows the cent filter-bank energies for thes same melodic motif sung by 7 musicians (male and female singers)[11]. The filter banks are placed based on the cent scale. After normalisation, across DKP, MSS and KVN, the first harmonic is at the 25th filter, while the second harmonic is at the 50th filter and the third harmonic is at the 70th filter. This makes it possible to compare melodic motifs across artists. Here MSS and DKP are female artists, while KVN is a male artist. Yet another variant for music is the Constant-Q transform (CQT) [25]. In this transform, the filters are placed based on the $\log\frac{f}{2}$ scale, to account for the harmonicity that exists in Western classical music. In the context of Carnatic music, this is equivalent to using a fixed tonic of 2 Hz and is therefore not considered here. The location of the harmonics are highlighted in the melodic motifs of DKP, MSS and KVN.

Figure 13 shows the mel filter bank, cent filter-bank energies and chroma filter-bank energies for short segments of Carnatic music, corresponding to that of *vocal solo*, *violin solo* and *composition*, respectively. From figure 13, it can be seen that while in the chroma filter bank the note is emphasised, in the mel filter bank, the prominent pitch is completely masked. On the other hand, in the cent filter bank, not only is the prominent pitch emphasised but the timbrel characteristics of the segments are also emphasised in the vicinity of the prominent pitch. These filter-bank energies are then converted to cepstral coefficients using the discrete cosine transform. To maintain consistency across the three approaches, the range of frequencies was set to 8000 Hz for mel filter bank, and six octaves were chosen for both chroma and cent filter banks The choice of six octaves is because the range of human voice is at most three octaves. The chroma filter bank was also normalised with the tonic of the performe; 8000 Hz for the mel filters accommodates six octaves across all musicians with varying tonic.

4.2 *Identifying the composition segment locations in a concert*

The fragments obtained using applauses must be merged based on the characteristics of items. The assumption that we make is that an item is rendered in a single *rāga* or alternatively, there are no applauses in a *rāgamālika* item. In order to distinguish between inter- and intra-item applauses, the location of the change in the *rāga* must be determined. As explained in section 2, the end of an item is always after the *composition* segment. Therefore, *composition* segment locations probably indicate the end of an item. GMMs are built for four classes, namely a *vocal solo*, a *violin solo*, a *composition* and

[10] DKP – D K Pattamal and MSS – M S Subbalakshmi are female singers and ALB – Alathur Brothers, GNB – G N Balasubramaniam, KVN – K V Narayanaswamy, SS – Sanjay Subramaniam and TMK – T M Krishna are male singers.

[11] DKP - DK Pattamal and MSS - MS Subbalakshmi are female singers and ALB – Alathur Brothers, GNB – GN Balasubramaniam, KVN – KV Narayanaswamy, SS – Sanjay Subramaniam and TMK – TM Krishna are male singers.

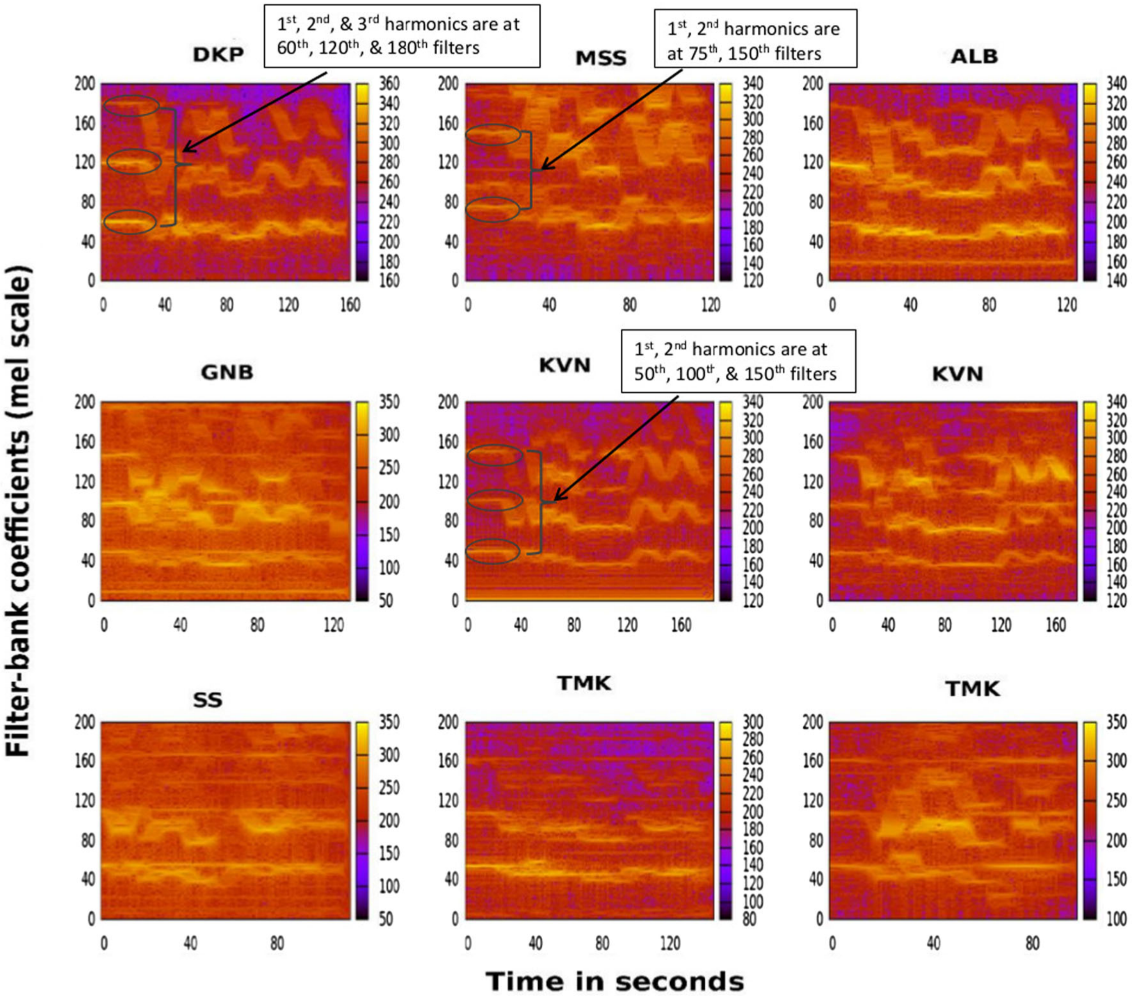


Figure 11. Filter-bank coefficients (filters are placed based on the mel scale) of melodic motifs sung by different musicians.

thaniavarthanam using CFCC features. Figure 14 shows the framework for segmenting the concert into these 4 segments using GMMs. The database used is the same as that used in section 3.2. Table 6 gives the details of the database used. It can be observed that for a given singer, there are different possible tonic values. Nevertheless, the tonic is preserved across all items in a concert.

To build the GMMs for the segmentation of the concert, *vocal solo*, *violin solo*, *composition* and *thaniavarthanam* segments are manually segmented from the database tabulated in table 6. From male/female recordings, three manually marked 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 4 classes. All 50 recordings are segmented based on applause locations and all these segments are manually labelled as *vocal solo*, *violin solo*, *composition* and *thaniavarthanam*. Labelled data are used as the ground truth for evaluating the segmentation performance[12]. After segmenting the concerts based on the applauses, 990 segments in total are obtained. These segments are tested against the trained GMMs. All these segments are labelled using 1-best result. The GMM training and testing details are as follows.

4.2a *Building the models*: During training, the following steps are followed.

1) Models are built separately for male and female musicians to account for timbrel differences. From male/female recordings, three segments are randomly chosen for each class.
2) Twenty-dimensional MFCC, ChromaFCC and CFCC features are extracted with a frame size of 100 ms and hop size of 10 ms. CFCC and ChromaFCC are extracted after tonic normalisation. Thirty-six filters are used.

[12] Labelling was done by the first author and verified by a professional musician.

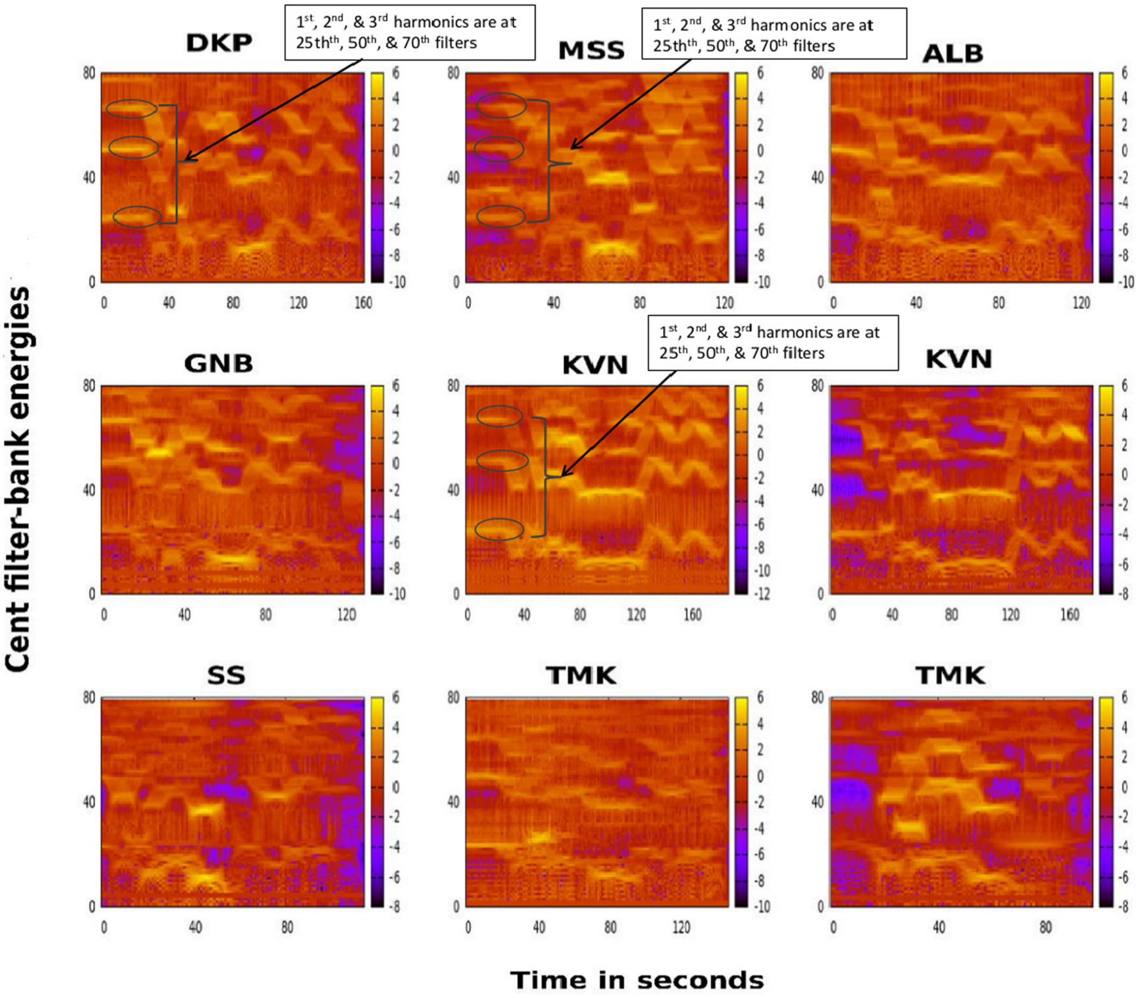


Figure 12. Filter-bank energies (filters are placed based on the cent scale) of melodic motifs sung by different musicians.

3) Thirty-two-mixture GMM models are built for the 4 classes, namely *vocal solo*, *violin solo*, *composition* and *thaniavarthanam*.

4.2b *Testing the models*: During testing, 50 live recordings are taken and for each concert the following steps are performed.

1) Tonic is identified for each concert using the pitch histograms.
2) Applause locations are identified for each concert using the approach discussed in section 3.1.
3) Once the applause locations are identified, segments between a pair of applauses are used for testing. For 50 concerts, there are 990 applauses.
4) the segments between pair of applauses are tested against the trained GMM models.
5) The segments are labelled using the 1-best result and the locations of composition segments are extracted.
6) The recordings of male and female singers are tested against the GMM models built for male and female singers separately.
7) The segmentation performance is calculated using the relation

$$\text{accuracy} = \frac{\text{correctly identified segments}}{\text{total number of segments}} \times 100\%.$$

Table 9 shows the segmentation performance using MFCC, ChromaFCC and CFCC features. Table 9 clearly shows that CFCC features perform well in locating the *composition* segments in the concert. The overall segmentation accuracy for 50 concerts, including male and female recordings, is 95%. The segmentation performance is based on the assumption that there is an applause after every segment.

5. Inter- and intra-item classification

As shown in figures 2 and 3, the segments between the pair of applauses can be the end of an item (inter-item) or within an item (intra-item). Hence, applauses in a concert can be either inter-item or intra-item applauses. Finding these

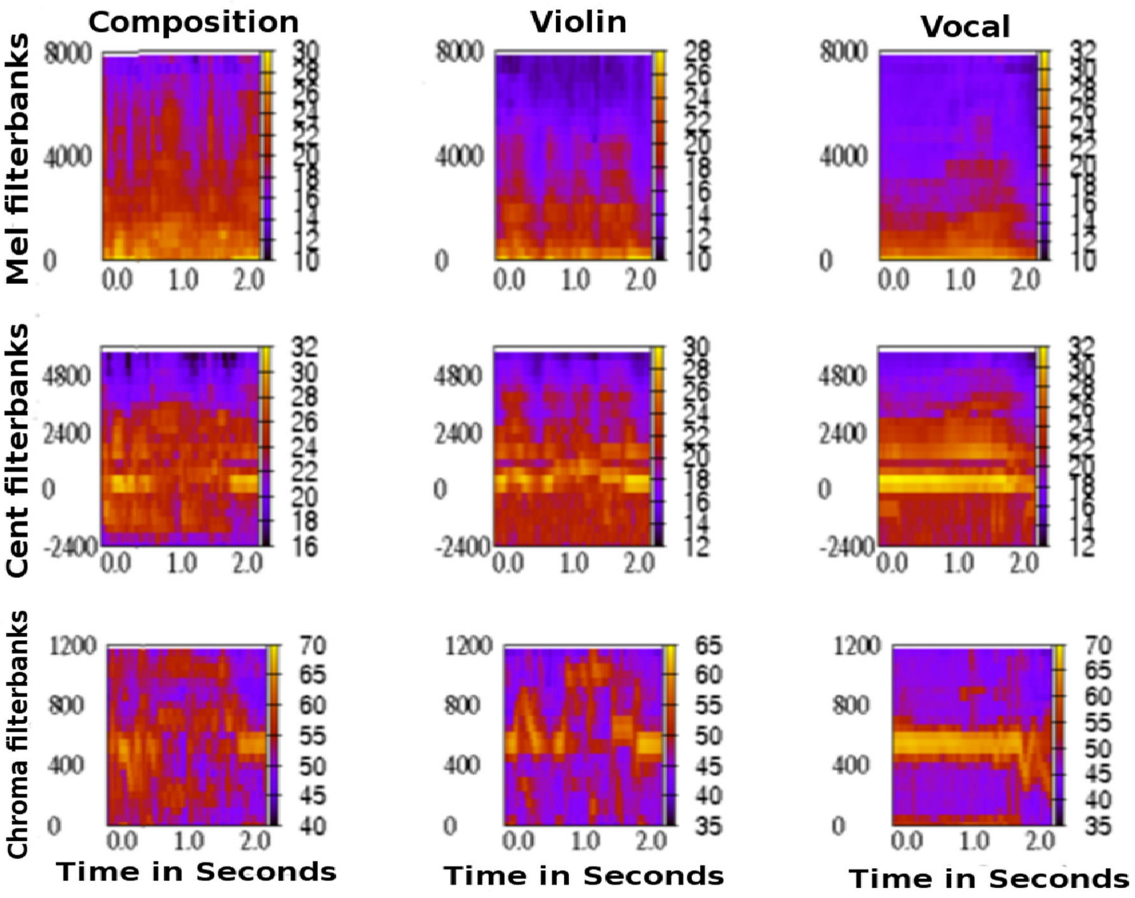


Figure 13. Time–frequency representations of composition, vocal solo and violin solo using MFCC, CFCC and ChromaFCC.

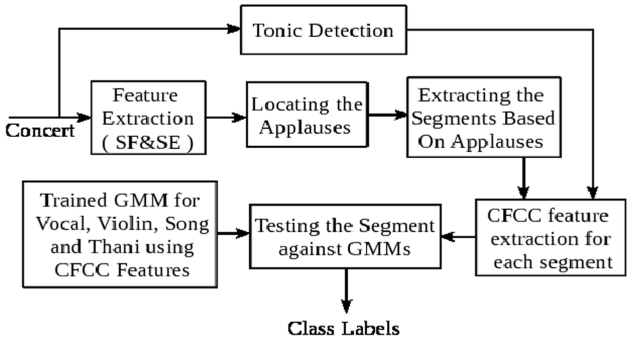

Tonic Detection
Concert
Feature Extraction (SF&SE)
Locating the Applauses
Extracting the Segments Based On Applauses
Trained GMM for Vocal, Violin, Song and Thani using CFCC Features
Testing the Segment against GMMs
CFCC feature extraction for each segment
Class Labels

Figure 14. A GMM-based system for segmenting a concert.

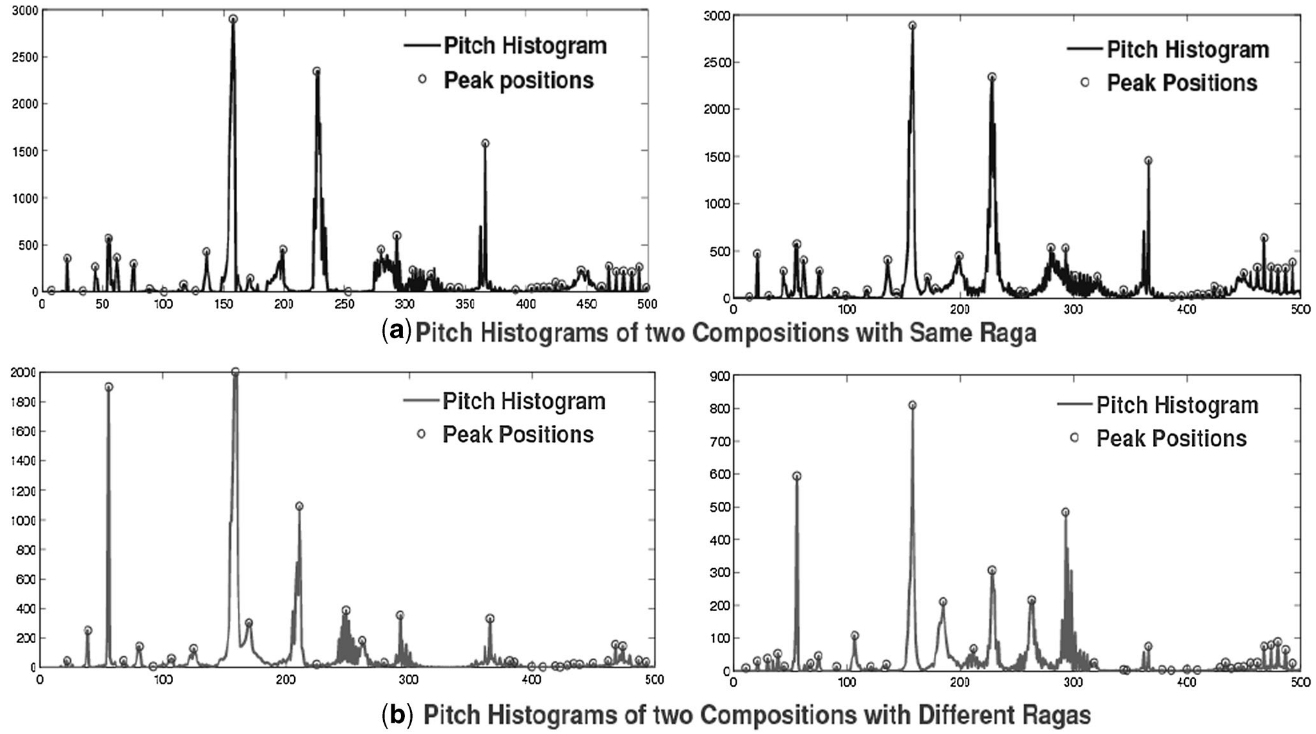
Table 9. Segmentation performance using MFCC, ChromaFCC and CFCC.

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

inter-item applause locations in a concert indicates the end of item locations in a concert. Using inter-item applauses, not only the number of items in a concert can be determined but also a concert can be segmented into items for archival.

As explained in section 4, the inter-item locations are always after *composition* segments. Hence, finding the *composition* segments aids in finding the inter-item locations. As shown in Figure 3, a *composition* can be fragmented into a number of segments owing to applauses. As a result, an item can be made up of more than one *composition* segment. If an item has more than one *composition* segment but belongs to the same item, then these segments must be merged. Segments can be merged based on *rāga* information.

Rāga uses a sequence of *svara*s upon which a melody is constructed [11]. *Rāga* recognition is a complex task because different *rāgas* can use the same scale with distinctly different phraseologies. Therefore, *rāga* identification cannot be performed with pitch histograms. In the context of segmentation, *rāga* identification is not required and only *rāga* change detection is required. An important property of a Carnatic music concert is that adjacent items in a concert are generally rendered in *distinctly* different



(a) Pitch Histograms of two Compositions with Same Raga

(b) Pitch Histograms of two Compositions with Different Ragas

Figure 15. Pitch histograms for same *rāga* and different *rāga* segments.

*rāga*s. For example, one would not find a concert where two adjacent pieces are sung in say *harikamboji* and *kamboji*, or even *khamaas*. Pitch histograms are therefore adequate to detect *rāga* change [26].

Pitch histograms for a segment show the relative notes sung by a musician [26]. Based on the pitch histograms, adjacent segments can be compared. Figure 15a shows the pitch histogram for the adjacent segments belonging to the same *rāga*, while figure 15b shows the pitch histogram for the adjacent segments that belong to different *rāga*s. It can be observed that in the pitch histograms corresponding to that of the same *rāga*, the positions of the peaks are more or less the same. On the other hand, for adjacent segments belonging to different *rāga*s, the locations of the peaks are different. Clearly, the heights of the peaks need not 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 *svara*s. A similarity measure between the two normalised pitch histograms is used as a measure to determine *rāga* change. Similarity between adjacent segments is calculated using Euclidean distance.

We now summarise the overall algorithm for finding number of items in a concert 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 labelled 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 histograms of the adjacent *composition* segments. If the similarity measure is above a threshold 'α', then the *composition* segment is inter-item, otherwise it is intra-item[13].
6) Based on the inter-item locations, intra-item fragments are merged into the corresponding items.
7) Inter- and intra-item classification accuracy is calculated using the relation

$$\text{accuracy} = \frac{\text{correctly identified items}}{\text{total number of items}} \times 100\%.$$

Figure 16 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 labelling of all these segments into *vocal solo*, *violin solo*, *composition* and *thaniavarthanam*. In the last column, some of the segments are

[13] The threshold was determined by performing a line search.

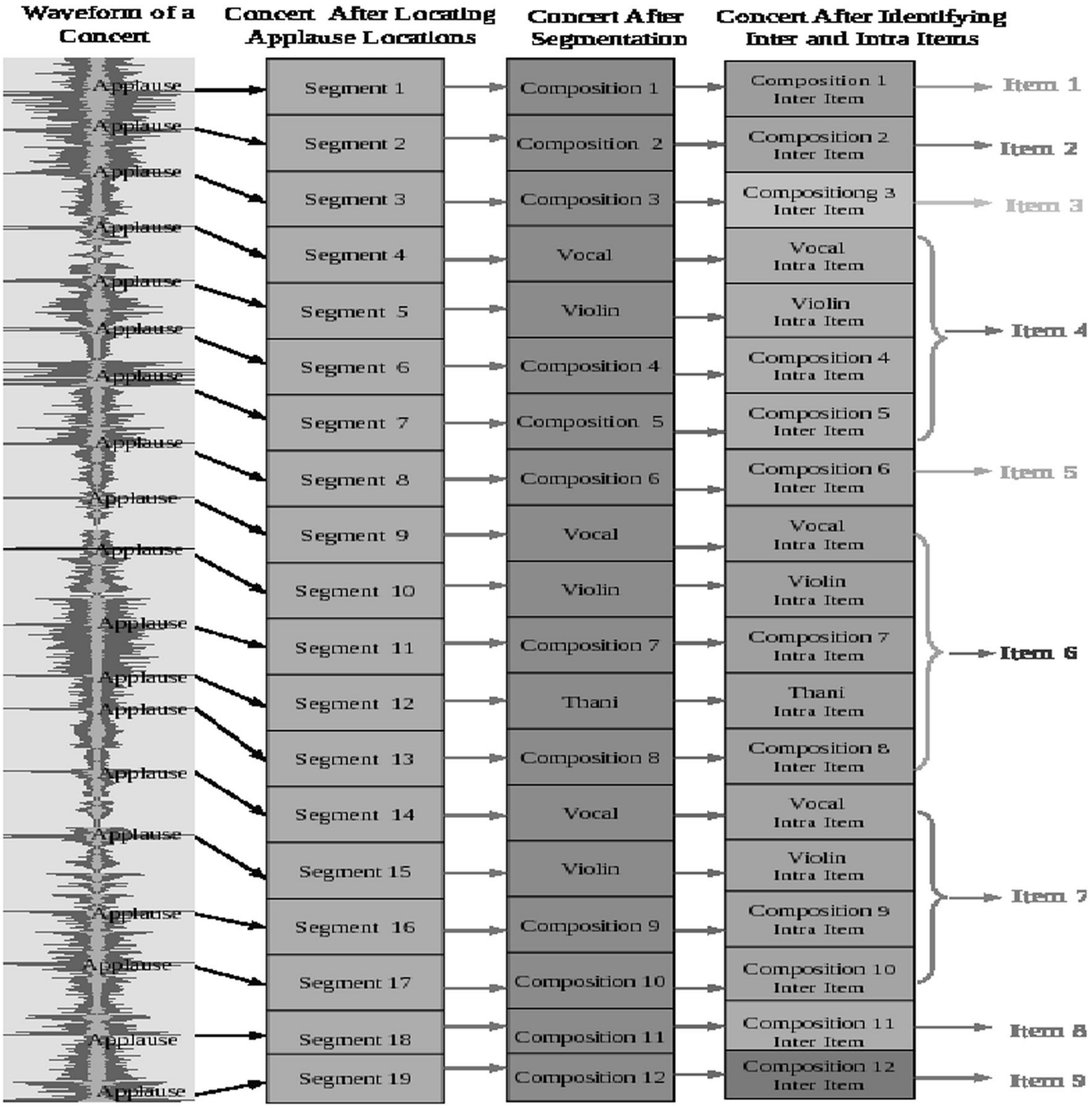


Figure 16. Overall procedures for identifying inter- and intra-items in a concert.

merged based on the similarity measure. From figure 16, observe that two *composition* segments 4 and 5 are merged to obtain item 4.

5.1 *Inter- and intra-item classification accuracy*

The total number of applauses for the database in table 6 is 990, where 394 applauses are inter-item applauses and 596 applauses are intra-item applauses. Table 10 shows the inter-item and intra-item classification accuracy for every musician. The overall classification accuracy for the 50 concerts is found to be 93%.

Table 10. Inter- and intra-item classification accuracy.

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

Inter- and intra-item classification purely depends on *composition* segments in a concert. As a result, the algorithm for finding inter-item locations works perfectly even if some applauses are missing after segments like *vocal solo*, *violin solo* and *thaniavathanam* segments. The algorithm reported in this paper makes mistakes when applauses are absent after

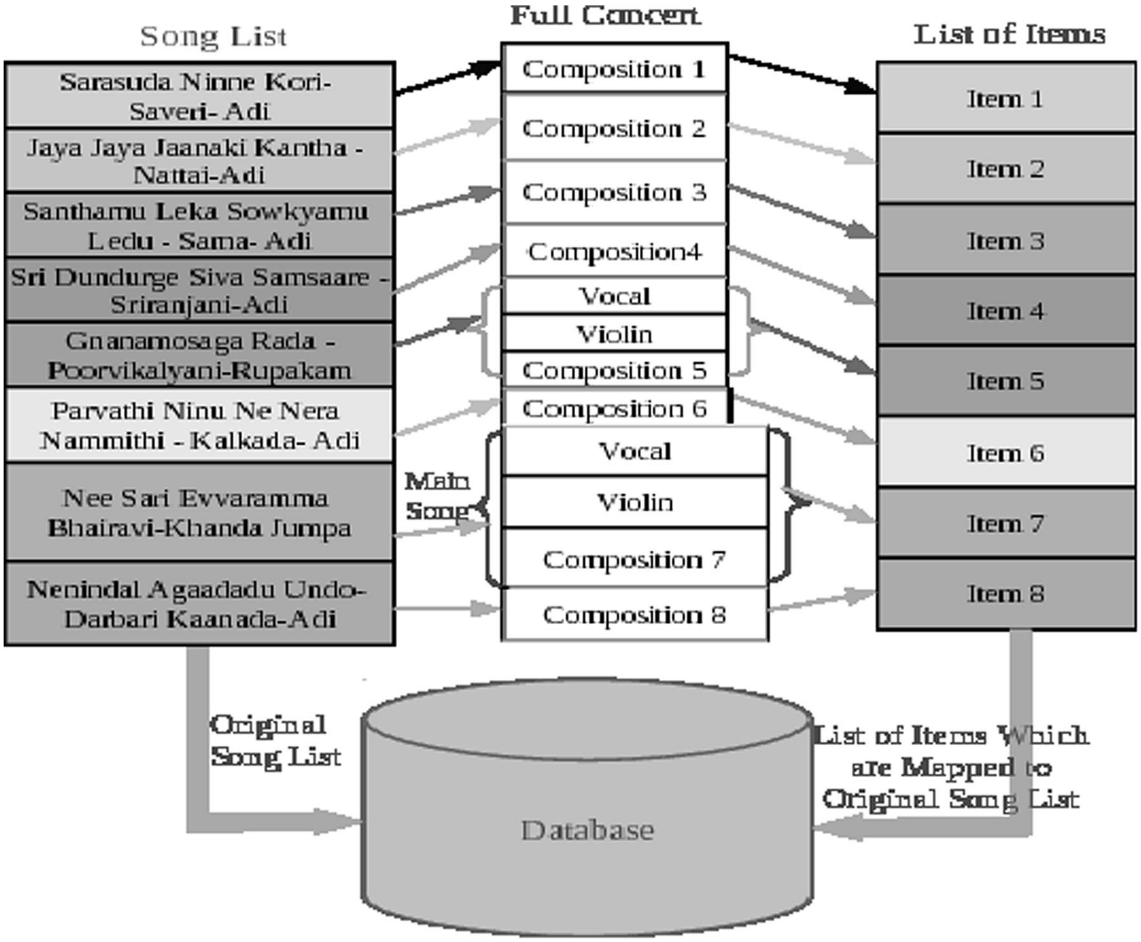

Song List
Full Concert
List of Items
Sarasuda Ninne Kori-Saveri- Adi
Jaya Jaya Jaanaki Kantha - Nattai-Adi
Santhamu Leka Sowkyamu Ledu - Sama- Adi
Sri Dundurge Siva Samsaare - Sriranjani-Adi
Gnanamosaga Rada - Poorvikalyani-Rupakam
Parvathi Ninu Ne Nera Nammithi - Kalkada- Adi
Nee Sari Evvaramma Bhairavi-Khanda Jumpa
Nenindal Agaadadu Undo- Darbari Kaanada-Adi
Composition 1
Composition 2
Composition 3
Composition4
Vocal
Violin
Composition 5
Composition 6
Vocal
Violin
Composition 7
Composition 8
Main Song
Item 1
Item 2
Item 3
Item 4
Item 5
Item 6
Item 7
Item 8
Original Song List
Database
List of Items Which are Mapped to Original Song List

Figure 17. Figure illustrates how the compositions list is mapped to list of inter-items.

composition segments (which is unlikely). In the database of table 6, in most of the recordings, the composition is fragmented due to improvisation.

5.2 *Mapping the audio and list to items*

In many performances, the meta-data are available for a concert in terms of *item* lists from the audience[14]. The *item* list obtained from the audience/artist is matched to the items obtained from the algorithm discussed in section 5. Figure 17 shows the mapping of the *composition* list to inter-items for archival. A typical segmented concert (using the proposed approach) is available on the web[15]

Ten live recordings of male and female musicians are used for mapping (based on meta-data available from social media). Table 11 shows the mapping accuracy for 10 live recordings of male and female musicians. The details of mapping and the mismatches (if any) are explained as follows.

Table 11. Performance of mapping the compositions list to the set of items.

Musician	No. of concerts	Actual inter-items	No. of 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

In a concert, every artist can optionally render a *ragamālika*. When a single *composition* is sung in *ragamālika*, 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 *rāga*s, owing to the extensive improvisational aspects in them, a number of

[14] http://www.rasikas.org

[15] http://www.iitm.ac.in/donlab/music/mapping_songslist/index.php. This concert is by Srividya Janakiraman accompanied by Gyandev on violin and Sriram on mridangam. This concert was performed at the ARKAY Convention Center, Mylapore, Chennai.

applauses can occur intra-item. Such special cases are limitations of the algorithm. Every concert may have one such item. 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 the closure *composition*, the *mangalam* is combined with the penultimate item in the concert. This is not considered as an error in this paper.

6. Characterising the applauses

In order to extract the semantic inference for content-based retrieval, we have used a technique called Cumulative Sum (CUSUM), which characterises the applause locations for the purpose of highlights' extraction. This section gives an overview of CUSUM technique and describes how CUSUM can be used to find the highlights in a concert.

6.1 *CUSUM*

From figures 6 and 8, it is clear that both time and spectral domain features do show significant change at music–applause boundaries. However, a simple threshold may not be sufficient to determine the duration and strength of an applause. The non-parametric approach discussed in [27] can be used to identify the change points where the time series become statistically non-homogeneous. This is achieved by sequentially estimating CUSUM on the time series of the feature under study. The estimation of CUSUM is as follows:

Let $X[n]$ be the value of time series at time n,

$$Y[n] = X[n] - a$$
$$Cusum[n] = \begin{cases} Cusum[n-1] + Y[n], & Y[n] > 0 \\ 0 & \text{otherwise} \end{cases}$$

If $Cusum[n] > \Theta$, then it suggests that there is a significant structural shift in the series. The values of a and Θ have to be estimated empirically and may vary across different data sets. The method works on the assumption that the underlying process is stationary, and is successful in detecting certain kinds of anomalies in network data [28, 29]. To compute CUSUM, $SE[n]$ and $SF[n]$ are thought of as time series. At the boundary between music and applause, the time series becomes statistically *inhomogeneous.*

Figure 18 shows the SE feature and its corresponding CUSUM values for the entire concert. From figure 18, most peaks correspond to applause segments, but some spurious peaks are also present.

In figure 19, the region where the CUSUM crosses the zero axis corresponds to the beginning of an applause. The CUSUM continuously increases and drops back to zero. This location marks the end of the applause. CUSUM is quite effective in estimating the duration of the applause. CUSUM can also be used to determine the strength of an applause. The height of the triangle is a measure of the strength of an applause. The CUSUM measures can be used to characterise the applauses. Figure 19 shows the applauses and CUSUM triangles when spectral entropy is used as a feature for different audio segments. The first applause is short and not loud, while the second is long and loud. From the figure it is clear that the *CUSUM triangle* captures both duration and strength of an applause. From figure19, we can observe that the CUSUM for the second applause is about 3 times that of the first applause. The first applause corresponds to an aesthetic moment at the beginning of the *ālāpana*, while the second applause corresponds to that of the end of the *tānam*.

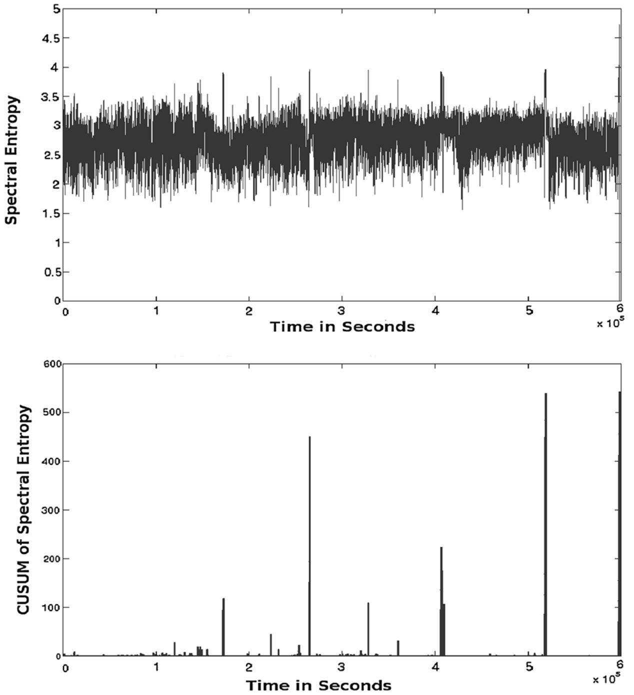


Figure 18. Entropy feature and its corresponding CUSUM values for entire concert.

Figure 20 shows applause locations (peaks) in a 3-hour concert using CUSUM of SE as the feature. This concert contains 25 applauses and 9 items. Figure 20 consists of a sequence of CUSUM triangles for the entire concert. The height and base of the triangle more or less characterise the applause. In this figure, the triangles are not visible because CUSUM values are plotted for a 3-hour concert. The CUSUM triangles thus become delta functions. Figure 21 shows the CUSUM values for a small part of a concert where the triangles are clearly visible.

6.2 *Evaluation of CUSUM technique*

For each concert, top 3 applauses that head the list of CUSUM value are considered as important or highlights of

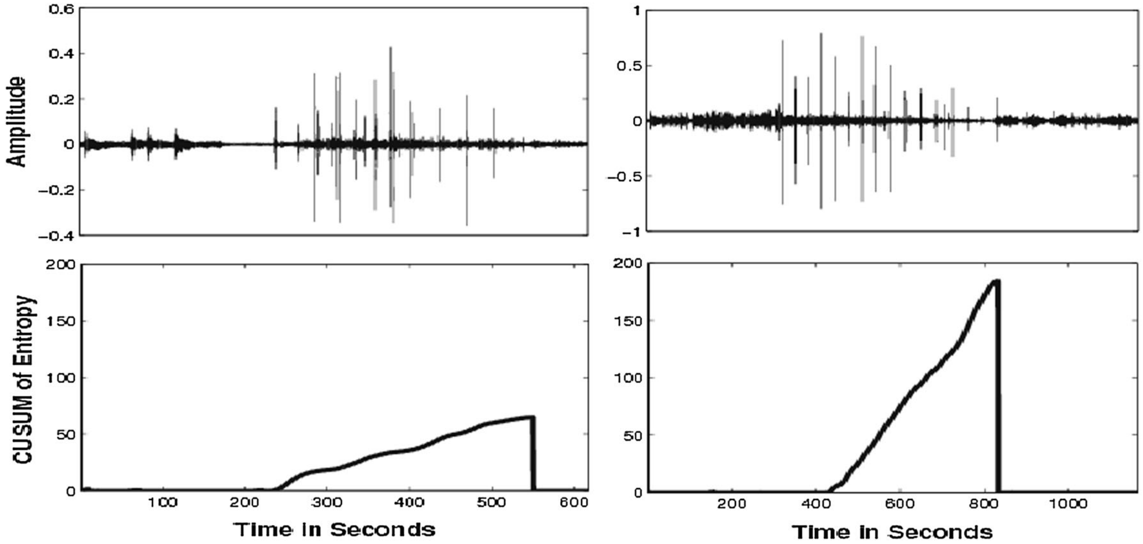


Figure 19. Types of applauses and the corresponding CUSUM of entropy feature.

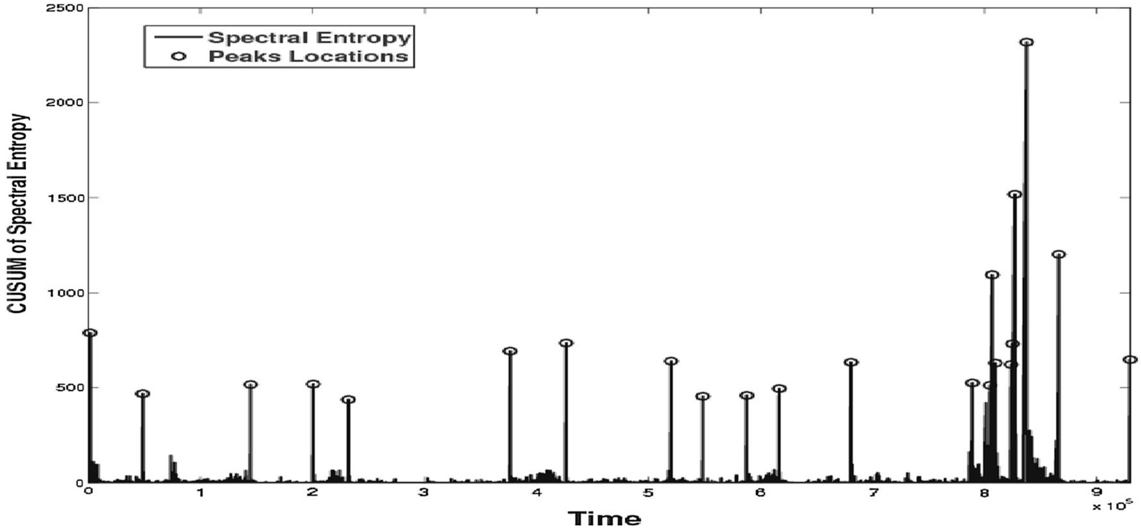


Figure 20. Applause locations for a concert using spectral entropy.

the concert. Table 12 shows the highlights of all concerts given in table 6 using this technique. From the table, we observe that the highlights are captured quite accurately. The highlights correspond to the end of a particular fragment of music. In the table, we have taken a union of the most important events in the concert based on CUSUM values for features that are mentioned in section 3. The events are named using the ground truth obtained by manually listening to the pieces. It is also worth noting that they were indeed the highlights of the specific concerts as verified by a professional musician.

7. Conclusions

A concert in Carnatic music is replete with applauses, and different time- and spectral-domain methods are explored to discriminate between music and applause segments. It is shown that machine learning methods based on GMM/SVM are preferred.

Tonic normalised CFCC features are used to classify inter-applause segments. The novel contribution is that the end of an item is signalled by a *composition* segment. The fragments that make up a composition are merged using *rāga*

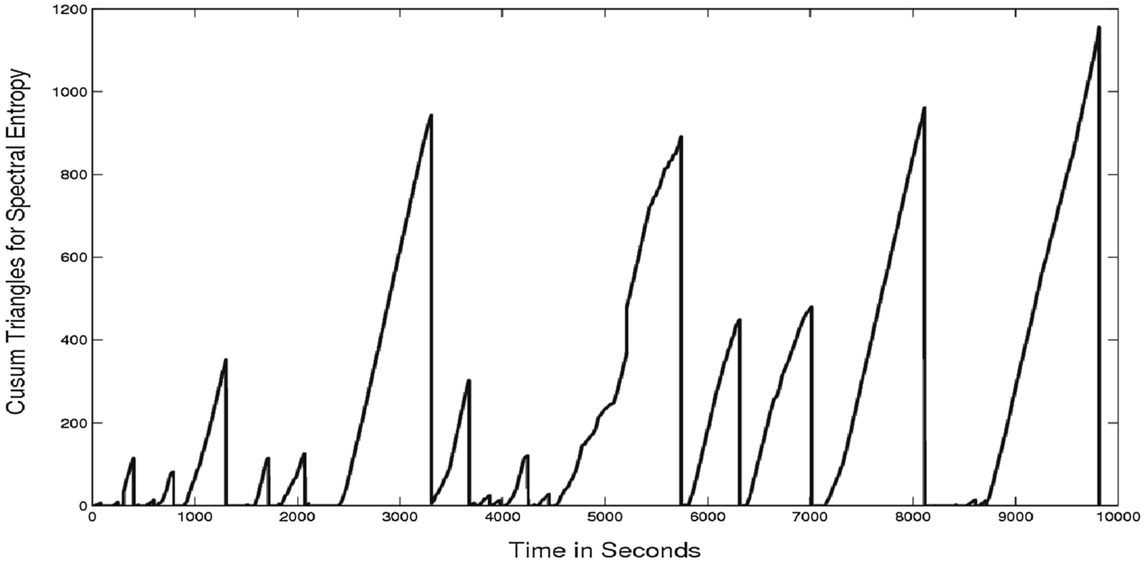


Figure 21. CUSUM triangles for a concert using spectral entropy.

Table 12. Different highlights captured based on CUSUM values.

S. No.	Highlights
1	Main Song
2	Varnam
3	Tanam
4	Aalapana of RTP
5	Thaniavarthanam
6	Svaram of Main Song
7	Niraval
8	Raga alapana of Main Song
9	Raga alapana of vocal solo
10	Raga alapana of violin solo
11	Within Thaniavarthanam
12	Pallavi

information. The paper conclusively shows that continuous unsegmented recordings can be gainfully segmented using applauses as a cue into items. The items can be tagged using meta-data available from reviews for archival.

Acknowledgements

This research was partly funded by the European Research Council under the European Union's Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583). We would like to thank Mr R K Ramakrishnan for arranging these concerts by Srividya Janakiraman and also seeking permission from the artists for the CompMusic project. We also thank Vidwan T M Krishna for giving permission to use his 26 personal recordings for applause analysis.

References

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

[2] Krishna T M 2013 *A southern music: the Karnatik story*. India: Harpercollins India

[3] Jarina R and Olajec J 2007 Discriminative feature selection for applause sounds detection. In: *Proceedings of the Eighth International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS'07*, IEEE, pp. 13–13

[4] Olajec J, Jarina R and Kuba M 2006 Ga-based feature extraction for clapping sound detection. In: *Proceedings of the Eighth Seminar on Neural Network Applications in Electrical Engineering, NEUREL 2006*, IEEE, pp. 21–25

[5] Carey M J, Parris E S and Lloyd-Thomas H 1999 A comparison of features for speech, music discrimination. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, IEEE, vol. 1, pp. 149–152

[6] Shi Z, Han J and Zheng T 2011 Heterogeneous mixture models using sparse representation features for applause and laugh detection. In: *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, September, pp. 1–5

[7] Li Y X, He Q H, Kwong S, Li T and Yang J C 2009 Characteristics-based effective applause detection for meeting speech. *Signal Process.* 89(8): 1625–1633

[8] Li Y X, He Q H, Li W and Wang Z F 2010 Two-level approach for detecting non-lexical audio events in spontaneous speech. In: *Proceedings of the International Conference on Audio Language and Image Processing (ICALIP)*, IEEE, pp. 771–777

[9] Manoj C, Magesh S, Sankaran M S and Manikandan M S 2011 A novel approach for detecting applause in continuous meeting. In: *Proceedings of the IEEE International Conference on Electronics and Computer Technology*, India, April, pp. 182–186

[10] Koduri G K, Ishwar V, Serrà J and Serra X 2014 Intonation analysis of rāgas in carnatic music. *J. N. Music Res. (Special Issue on Computational Approaches to the Art Music Traditions of India and Turkey)* 43: 72–93

[11] Krishna T M and Ishwar V 2012 Svaras, gamaka, motif and raga identity. In: *Proceedings of the Workshop on Computer Music*, July, pp. 12–18

[12] Pesch L 2009 *The Oxford illustrated companion to south Indian classical music*. Oxford: Oxford University Press

[13] Serra J, Koduri G K, Miron M and Serra X 2011 Assessing the tuning of sung Indian classical music. In: *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, pp. 157–162

[14] Dutta S 2016 *Analysis of motifs in Carnatic music: a computational perspective*. Master's Thesis, Indian Institute of Technology Madras

[15] Bellur A and Murthy H A 2013 A novel application of group delay function for identifying tonic in Carnatic music. In: *Proceedings of the 21st European Signal Processing Conference (EUSIPCO)*, IEEE, pp. 1–5

[16] Bellur A, Ishwar V, Serra X and Murthy H A 2012 A knowledge based signal processing approach to tonic identification in Indian classical music. In: Serra X, Rao P, Murthy H and Bozkurt B (Eds.) *Proceedings of the 2nd CompMusic Workshop*, July 12–13, Istanbul, Turkey. Barcelona: Universitat Pompeu Fabra, pp. 113–118

[17] Sarala P, Ishwar V, Bellur A and Murthy H A 2012 Applause identification and its relevance to archival of carnatic music. In: Serra X, Rao P, Murthy H and Bozkurt B (Eds.) *Proceedings of the 2nd CompMusic Workshop*, July 12–13, Istanbul, Turkey. Barcelona: Universitat Pompeu Fabra

[18] Rabiner L R and Schafer R W 2011 *Theory and applications of digital speech processing*. Upper Saddle River, NJ: Pearson International

[19] Cannam C, Landone C, Sandler M B and Bello J P 2006 The sonic visualiser: a visualisation platform for semantic descriptors from musical signals. In: *Proceedings of the ISMIR Conference*, pp. 324–327

[20] Chang C C and Lin C J 2011 LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2(3): 27

[21] Müller M, Kurth F and Clausen M 2005 Audio matching via chroma-based statistical features. In: *Proceedings of the ISMIR Conference*, vol. 2005, p. 6

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

[23] Salamon J, Gulati S and Serra X 2012 A two-stage approach for tonic identification in Indian art music. In: *Proceedings of the Workshop on Computer Music*, July, pp. 119–127

[24] De Cheveigné A and Kawahara H 2002 Yin, a fundamental frequency estimator for speech and music. *J. Acoust. Soci. Am.* 111(4): 1917–1930

[25] Brown J C and Puckette M S 1992 An efficient algorithm for the calculation of a constant q transform. *J. Acoust. Soc. Am.* 92(5): 2698–2701

[26] Chordia P and Rae A 2007 Raag recognition using pitch-class and pitch-class dyad distributions. In: *Proceedings of the ISMIR Conference*, pp. 431–436

[27] Brodsky E and Darkhovsky B S 1993 *Nonparametric methods in change point problems*. New York: Springer Science & Business Media

[28] Wang H, Zhang D and Shin K G 2002 Syn-dog: sniffing syn flooding sources. In: *Proceedings of the ICDCS*, July, pp. 421–428

[29] Liu H and Kim M S 2010 Real-time detection of stealthy ddos attacks using time-series decomposition. In: *Proceedings of the IEEE International Conference on Communications (ICC)*, IEEE, pp. 1–6