APPLAUSE IDENTIFICATION AND ITS RELEVANCE TO ARCHIVAL OF CARNATIC MUSIC

Padi Sarala and Vignesh Ishwar
Dept. of
Computer Science &
Engineering
IIT Madras, India
sarala@cse.iitm.ac.in,
vigneshishwar@gmailcom

Ashwin Bellur
Dept. of
Computer Science &
Engineering
IIT Madras, India
ashwinbellur@gmail.com

Hema A Murthy
Dept. of
Computer Science &
Engineering
IIT Madras, India
hema@cse.iitm.ac.in

ABSTRACT

A Carnatic music concert is made up of a sequence of pieces, where each piece corresponds to a particular genre and rāga (melody). Unlike a western music concert, the artist may be applauded intra-performance, inter-performance. Most Carnatic music that is archived today correspond to a single audio recordings of entire concerts.

The purpose of this paper is to segment single audio recordings into a sequence of pieces using the characteristic features of applause and music. Spectral flux, spectral entropy change quite significantly from music to applause and vice-versa. The characteristics of these features for a subset of concerts was studied. A threshold based approach was used to segment the pieces into music fragments and applauses. Preliminary results on recordings 19 concerts from matched microphones show that the EER is about 17% for a resolution of 0.25 seconds. Further, a parameter called CUSUM is estimated for the applause regions. The CUSUM values determine the strength of the applause. The CUSUM is used to characterise the highlights of a concert.

1. INTRODUCTION

"*APPLAUSE (Lat. applaudere, to strike upon, clap) is primarily the expression of approval by clapping of hands*," according to the Encyclopedia Brittanica. Applause is indicative of the collective approval of a group of people for a performance. Although, initially the applause can be asynchronous, the applause becomes rhythmic within a few seconds. Identifying applauses in a football video can be used to determine the highlights of the play. Similarly, in an audio recording of a concert, identifying the location of applauses, duration of applauses can be used to archive the highlights of a concert. In Western Music a performer is applauded at the end of a concert, or at most at the end of a piece. In Classical Indian Music too, the audience applauds a performer inter-piece, and at the end of a concert. In addition to this, the musician is also applauded when there is an "aesthetic moment" within a piece. The purpose of this paper is manifold; find the location of applauses in a single continuous recording of a concert. The location and duration of the applause can then be used to characterise the approvals. Further, the concert can be segmented into individual pieces. Applauses can also occur intra-piece. These locations provide the highlights of a given concert.

The appropriate choice of audio features is crucial for classification or segmentation of an audio signal. For most audio signals, in particular music, the spectral properties change slowly with respect to time. This has led to a wide variety of *short-time* processing methods. In [1], a GMM based classifier is used to segment the singer's voice in an audio. Four different features are used, namely, short-term energy, zero cross rate, spectral flux and harmonic coefficient. In [2], a music piece is segmented into different structural components. The paper uses a combination of N-grams to model the sequential dependencies in a musical piece, and acoustic properties to segment a western music performance using the transcription and acoustic waveform. In [3] different features are compared for separating music and speech. The features include amplitude, delta-amplitude, pitch, delta-pitch, cepstra, delta-cepstra, zero-crossing rate and delta-zero crossing rate. Gaussian Mixture models are built using each of the features. In [4] has used evolutionary programming based on genetic algorithms and simulated annealing to determine the discriminative features for music and applause. A manually labeled data set is used for this purpose. The discriminative features are then used to classify the labeled segments into applauses and otherwise.

In this paper, the focus is to determine the location of an applause by processing the audio signal using appropriate features. Unlike speech and music, the characteristics of applause and music have distinct spectral characteristics. This is primarily required for processing Carnatic Music concerts.

Carnatic Music is based on the oral tradition. At a gross level, Carnatic music consists of two components, namely, *kalpita sangita* and *kalpana sangita* [5]. *kalpita sangita* in

a concert corresponds to fixed compositions, to be performed as composed or taught, while *kalpana sangita* corresponds to improvisational parts of a concert. *Kalpana sangita* is also called *manodharma sangita*, a musical aspect where the creativity of the musician plays a role. In a concert, the performer, generally illustrates the various nuances of a rāga by means of the following: *Ālāpana, tānam, kalpana svaras*, and a *kirtana* (song composed by a composer). The *Ālāpana, tānam and kalpana svaras* are the improvisational aspects of the concert, whereas the *kirtana* is the fixed composition. In a concert (*kutcheri*), applauses can be heard after each of the improvisational aspects and the *kirtana*. The audience occasionally applauds the artist in-between an improvisational piece or even a *kirtana*, when some aspect of the music appeals to them. The purpose of the work reported in this paper is to mar kthe locations of these applauses in a concert and use them as landmarks for archival purposes. The applauses can be also be used as a cue for segmenting a single concert recording into its constituent pieces. Applause is a mark of appreciation by the audience to the music. The location of the applause can be used as an index to perform search.

Earlier work on identifying applauses uses energy and zero crossing rate as criteria [6]. Although the energy during an applause is small compared to that of music, and the zero crossing rate is high for an applause, these features are very noisy. In this paper, spectral characteristics of music and applause are used to segment an individual recording into individual pieces. Liu et al [7] show that spectral flux can be used efficiently to distinguish between music, speech and environment noises. It is observed that spectral flux and spectral entropy are useful measures to distinguish between applauses and music. As these measures are quite noisy, a technique call CUSUM [8] is used to smooth out the noise and highlight the applauses.

In Section 2, we discuss briefly the different features that are used to identify music and applause. In Section 2.3, we discuss the technique based on [8] to highlight the applauses. Section 3 gives a brief detail of the database that was used in the study. In this Section, the results are tabulated in terms of misses and false-alarms. In Section 4, an analysis of CUSUM is performed to categorise applauses. A possible approach to applause classification using the CUSUM triangles is suggested. Finally in Section 5, the concluding remarks are presented.

2. FEATURE EXTRACTION

In this Section, an attempt is made to derive features that can distinguish between music and applause in a Carnatic music concert. Figures 1 and 2 show the *time-domain* characteristic of a typical sequence *music segment* and *applause segment*. Clearly, the time-domain signal corresponding to music is more structured, while that corresponding to that applause is rhythmic but not very structured. Although, music has structure, owing to its quasistationarity, the characteristics change with time, albeit slowly. On the other hand, applauses across pieces have a similar structure. Although this is not discernible in the time-domain waveform, it is evident in the spectrum. Figures 3 and 4 show the typical power spectrum of a sequence of applause segments, respectively. From Figures 3 and 4, observe that the spectra of music and applause are quite different, For applauses, it is observed that the spectra are quite flat, while for that of music, the spectra shows structure. Although, there are differences in the time domain signal too, since a rāga continuously changes, it is difficult to characterise music in the time domain. Whereas, while the spectra of music also changes, it is observed that the spectra of applause is more or less stationary. Further, the spectra dynamic range of an applause is small owing to its unpredictability. Music being predictable (based on the past) has a large dynamic range. An attempt is made in this paper to characterise the predictability of music vis-a-vis the unpredictability of applauses. An attempt is also made to quantify an applause.

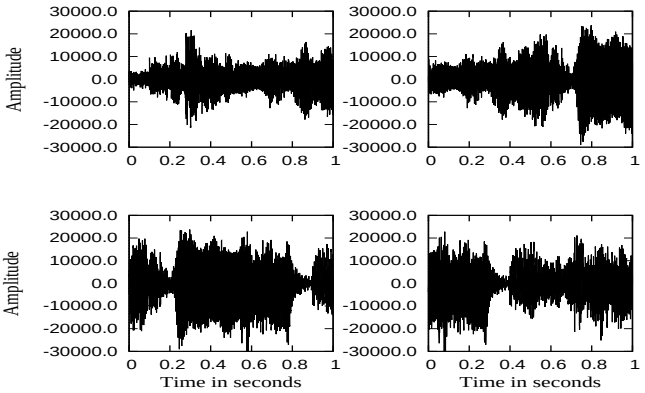

Figure 1. Typical sequence music segments (time domain)

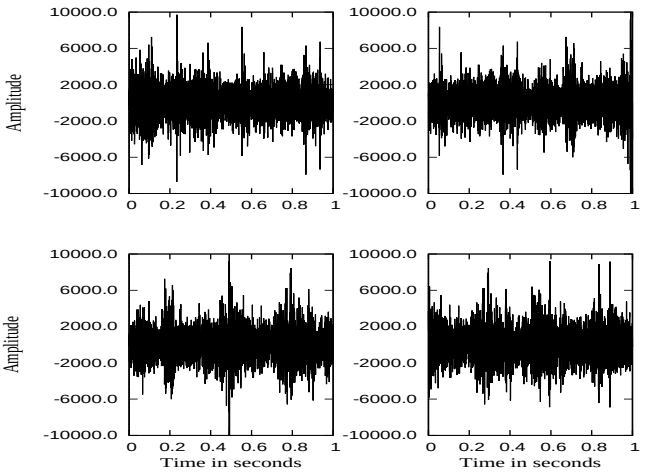

Figure 2. Typical sequence of applause segments (time domain)

Given that the spectral properties of music and applause are different, in this section, we describe two different feature extraction techniques that have been successful in detection of applauses and music. The general framework is adapted from [9]:

$$Q_{\hat{n}} = T[X_{\hat{n}}(e^{j\omega})] \quad (1)$$

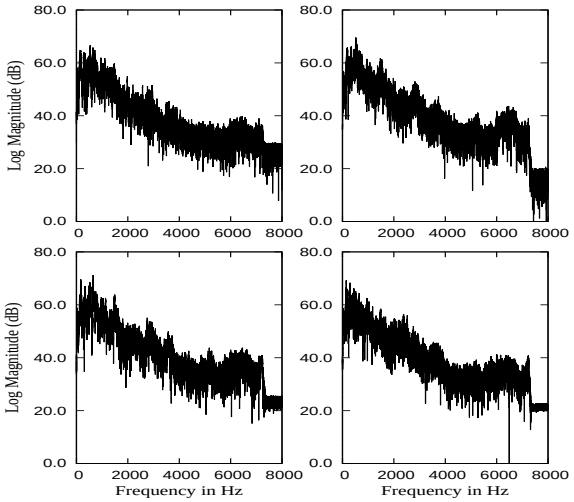

Figure 3. Typical spectra of a sequence of music segments (spectral domain)

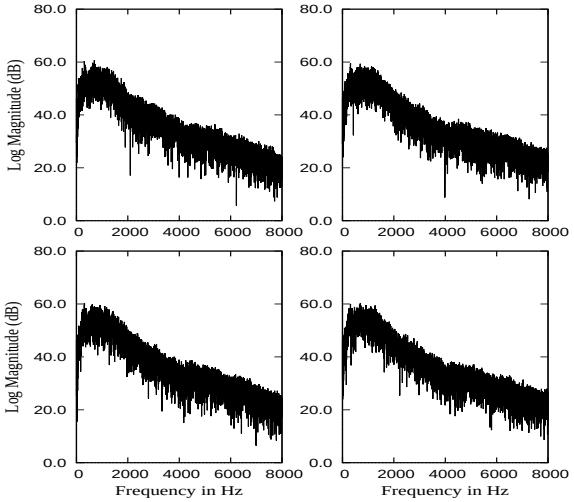

Figure 4. Typical spectra of a sequence of applause segments (spectral domain)

where $X\hat{n}(e^{j\omega})$, is given by

$$X_{\hat{n}}(e^{j\omega}) = \sum_{m=-\infty}^{\infty} w[\hat{n}-m]x[m]e^{-j\omega(\hat{n}-m)}$$

where $w[\hat{n}-m]$ is a sliding analysis window and $T[.]$ is a particular transformation. Typically the analysis windows overlap by more than 50%.

2.1 Spectral Flux (SF)

Spectral Flux SF is also called Spectral Variation. Spectral flux characterises the change in spectrum between adjacent two frames of speech. It measures how quickly the power spectrum changes. Spectral flux can be used to determine the timbre of an audio signal.

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

where $X_n(\omega)$ is the *normalised power spectrum* of the nth frame of an audio signal. Three different normalisations, were experimented with:

1. No normalisation.
2. Power spectral density normalisation: In this approach $XNorm_n(\omega)$ is defined:

$$XNorm_n(\omega) = \frac{X_n(\omega)}{\int_{\omega} X_n(\omega) d\omega} \qquad (3)$$

   This normalisation gives the relative contribution of different spectral components. Given that spectrum of applause Figure 4 is relatively flat, while that of the signal is relatively peaky, the spectral flux could be significantly different.

3. Peak normalisation: In this approach $XNorm_n(\omega)$ is defined as:

$$XNorm_n(\omega) = \frac{X_n(\omega)}{\max_{\omega} X_n(\omega)} \qquad (4)$$

In music signals, certain spectral components and their harmonics are emphasised in a melody, while others components are not present. Again, we argue that since applause has a flat spectrum, all frequency components after normalisation will be close to 1.0, while that for music will show a significant variation.

Figure 5 (a), (b) and (c) shows SF as function of time, using all of the three approaches. It can be observed that both "no normalisation" and "peak normalisation" seem to show significant change at the boundary between music and applause. But the dynamic range of "no normalisation" is very high. We therefore use peak normalisation in our analyses.

2.2 Spectral Entropy (SE)

Entropy is a measure of randomness of a system. Shannons 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

$$\mathrm{H(X)} = -\sum_{i=1}^{N} \mathrm{p}(x_i) log_2 \left[\mathrm{p}(x_i)\right] \qquad (5)$$

Given that power spectrum can be thought of a power spectral density, the power spectral can be thought of as a probability density function. The entropy of the power spectral density function is then computed:

$$\begin{aligned} PSD_n(\omega) &= \frac{\mid X_n(\omega) \mid^2}{\int_{\omega} \mid X_n(\omega) \mid^2 d\omega} \\ SE[n] &= -\int_{\omega} PSD_n(\omega) \log PSD_n(\omega) d\omega \end{aligned} \qquad (6)$$

The continuous frequency ω becomes discrete, as the signal is sampled and the discrete short-time Fourier transform is computed for every frame:

$$\begin{aligned} PSD_n[k] &= \frac{\mid X_n[k] \mid^2}{\sum_k \mid X_n[k] \mid^2} \\ SE[n] &= -\sum_{k} PSD_n[k] \log PSD_n[k] \end{aligned} \qquad (7)$$

where k is the index of a DFT bin.

Figure 5(d) shows SE as a function of time. Observe that the spectral entropy is also distinctly different for applause and music.

SE and SF are complementary features, in that SF is measured across frames, while SE is measured within every frame. Both SF and SE are not smooth functions of time, and therefore a simple threshold based approach will not detect boundaries accurately. The functions $SE[n]$ and $SF[n]$ are first smoothed using a *moving average filter*. Then, to get the exact boundaries of applause, a technique called Cumulative Sum in the next Section, which can be used to emphasise and characterise applauses.

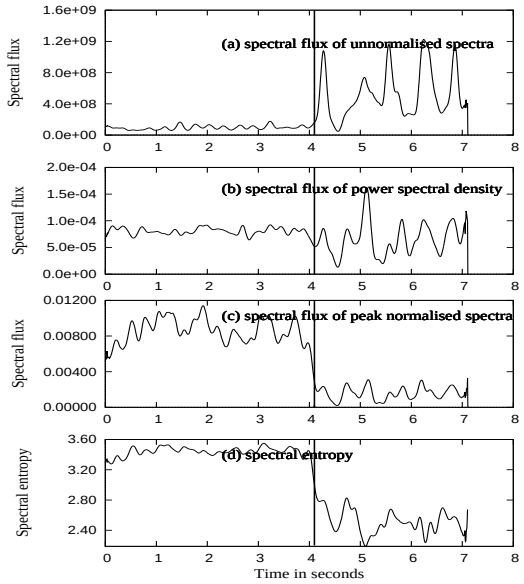

Figure 5. Different measures of spectral analyses for determining applause positions. The solid line corresponds to a boundary. The signal before the solid line corresponds to an applause and the signal after the solid line corresponds to that of music.

2.3 Cumulative Sum(CUSUM)

From Figures 5 it is clear that both parameters do show significant change at a boundary. But a simple threshold may not be sufficient to determine the duration and strength of the applauses. 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 *inhomongeneous.* A non-parametric approach discussed in [8], can be used to identify the statistical inhomogeneity. This is achieved by sequentially estimating a Cumulative Sum (CUSUM) on the time series of the feature in question. CUSUM is estimated 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 has been successful in detecting certain kinds of anomalies in network data [10, 11].

3. EXPERIMENTAL ANALYSIS

In this section, we first describe that the database that was used in the study. Next, we evaluate the performance of features extracted in the previous section, at the frame level. The manually marked applauses by a musician at the frame level are used as the ground truth.

3.1 Database used in the study

Nineteen concerts were taken for study. All the concerts are live recordings of complete concerts. All were recorded using a Sony PCM-D50 recorder, The recorder was placed in the audience, and the recordings include environmental noise, conversations between people in the audience, etc. All the concerts are vocal concerts, in that the lead musician is a singer. Each concert has about 15-20 applauses resulting in a total of 343 applauses.

3.2 Performance Evaluation

The features $SF[n]$ and $SE[n]$ are smoothed using a rectangular moving average filter. The moving average filter of length 15 is applied three times. This sort of approximates a Gaussian window. As $SF[n]$ and $SE[n]$ are unidimensional features, a simple threshold is employed to determine whether a given frame corresponds to that of an applause or music. Figure 6 shows the Detection Error Tradeoff curves [12] obtained for different thresholds on the raw $SE[n]$ and $SF[n]$. As onsets of any event in music [1], are at least 0.5 seconds long, a leeway of 0.25 seconds is permitted in the detection of the applause. The Equal error rates (EER) are given in Table 1

Table 1. *EER for applause detection*

Method	EER
Spectral Flux (no norm)	44.55 %
Spectral Flux (peak norm)	23.33%
Spectral Entropy	17.33%

From the Table, we observe that both spectral flux (peak norm) and spectral entropy are quite effective in detecting the applauses, while spectral flux (no norm) is not very effective. Although in Figure 5, there is a significant change in spectral flux, when the signal changes from an applause to music, clearly a threshold based method is inadequate.

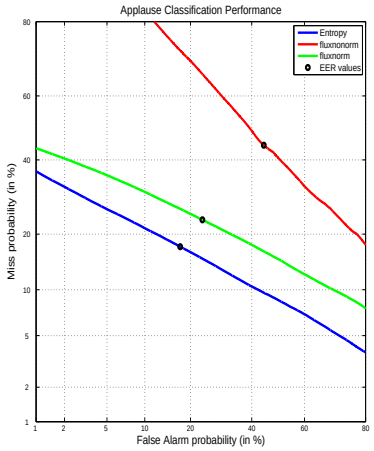

Figure 6. Detection Error Tradeoff curves for applause detection using different methods

4. CHARACTERISING APPLAUSES USING CUSUM

The CUSUM was computed for both spectral flux (peak norm) and spectral entropy. Figure 7 shows the spectral flux (peak norm) and entropy at the specific boundary between music and applause. Clearly the location of the applause that is marked by the solid black line in $SE[n]$, and $SF[n]$ of Figure 5 are clearly visible in $Cusum[n]$ of SF and $Cusum[n]$ of SE.

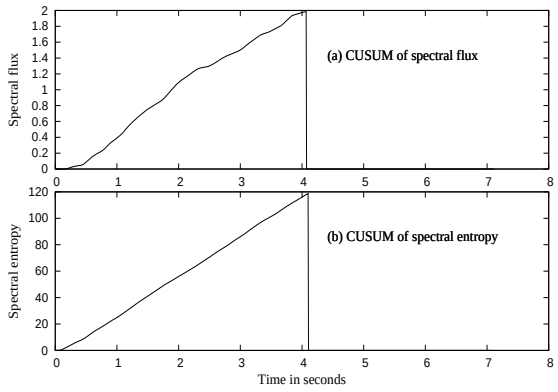

Figure 7. Cumulative sum on spectral flux an spectral entropy

4.1 Applause detection

As mentioned in the previous Section, the CUSUM is computed on the $SE[n]$ and $SF[n]$. The average of the CUSUM values were computed for the entire concert. After experimenting with a number of concerts, the parameter a was chosen as $1.5 \times$ *the average* $SF[n](SE[n])$. Θ is chosen as 0. In Figure 7, 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 the an applause. The height of the triangle is a measure of the strength of an applause. The CUSUM measure of applauses can be used to characterise applauses.

Figure 8 consists of a sequence of CUSUM triangles (for a carefully chosen value of a) for the entire piece [1] Eight of the applauses in this piece are indicated by the location of peaks. There are nine applauses in this concert. The height and base of the triangle more or less characterise the applause [2].

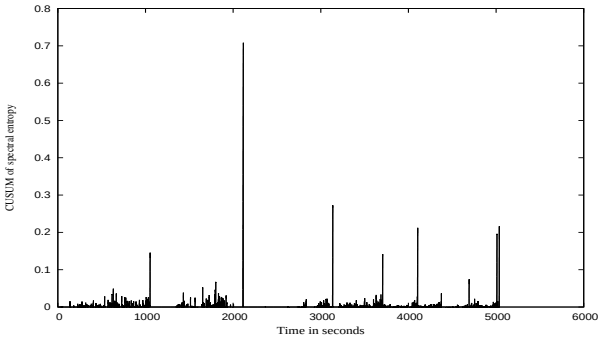

Figure 8. CUSUM for a rāgam, tānam, pallavi

Figure 9 shows the applauses and CUSUM triangles at $\approx$ 1045 *seconds* and $\approx$ 2095 *seconds* of Figure 8. 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. Notice that the scale on both X-axis and Y-axis are different for CUSUM of the applauses. 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.

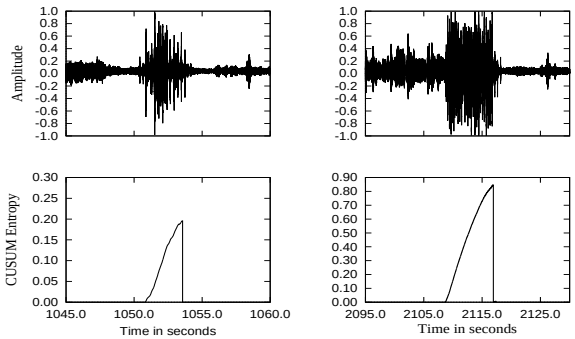

Figure 9. Types of applauses and the corresponding CUSUM of Entropy

The CUSUM can be used further to categorise applauses by the size and shape of the triangle. Although this technique has to be verified on a large database,

[1] A rāgam, tānam, pallavi is a particular kind of piece in Carnatic music concert that is replete with improvisation.

[2] Since the duration is about 6000 seconds, the triangles appear as peaks.

Artist Name	highlight 1	highlight 2	highlight 3
Abhishek Raghuram & Jayatirth Mevundi	tānam	RTP	jugalbandi
Bombay Jayashree	vocal Ālāpana	violin Ālāpana	RTP/tānam
Sanjay Subramanian	mridangam+violin	kriti	tānam

Table 2. Highlights of concerts using CUSUM

preliminary analysis do show that the CUSUM characteristics seldom change across concerts. The major drawback of the CUSUM based approach is the choice of a. The choice of this parameter depends up the stationarity of the signal. As music signals are quasistationary an adaptive choice of threshold would be appropriate. The appropriate technique for automatic choice of threshold needs to be explored.

Alternatively, it is observed that thresholds are easy to determine for the features of interest. Therefore, one could use the thresholds on spectral flux and entropy to determine the location of the applauses. The CUSUM can be computed for these regions alone. The area of the triangle corresponds to the strength of applause both in terms of duration and entropy/spectral flux. Longer the duration and larger the entropy/flux, the more effective the applause. Table 2 shows the location of the top three highlights for a sample of about three concerts. From the table we observe that the highlights are quite accurately captured. 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 the CUSUM values from different features that are mentioned in Section 2. The events are named using the ground truth obtained by manually listening to the pieces. There is one misidentification in Sanjay Subramanian concert that corresponds to the mridanga with the violin in the background playing a single note. It is also worth noting that these were indeed the highlights of the specific concerts as verified by a listener. Especially, the *kriti* in the Sanjay Subramanian concert was rendered very well and thus received a significant applause.

5. CONCLUSION

In this paper, we discussed a technique for applause identification in musical performances. The spectral characteristics of music and applause are significantly different. Two different techniques based on processing the spectra are explored. Spectral flux of peak normalised spectra and spectral entropy are used to detect applauses. Spectral entropy is shown to perform better than spectral flux in detecting applauses.

Applause identification is very important for Indian Music as most music performances are single recordings. Further, many of the old recordings from long playing records and cassettes that have been digitally mastered correspond to single recording of multiple pieces. CUSUM is a parameter that is used to highlight the applause. Larger the value of CUSUM louder and longer the applause. The highlights of the concert are determined using CUSUM. The location of the highlights in a concert are then archived.

6. 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).

7. REFERENCES

[1] T. Zhang, "Automatic singer identification," in *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, vol. 1, july 2003, pp. I – 33–6 vol.1.

[2] J. Paulus, "Improving markov model based music piece structure labelling with acoustic information," in *International Society for Music Information Retrieval Conference*, August 2010, pp. 303–308.

[3] M. J. Carey, E. S. Parris, and H. LLoyd-Thomas, "A comparisons of features for speech, music discrimination," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 1, march 1999, pp. 149 – 152.

[4] J. O. Roman Jarina, "A discriminative feature selection for applause sounds detection," in *in Proc. 8th Int. Workshop on Image Analysis for Multimedia Interactive Service*, 2007.

[5] T. M. Krishna, *Kalpita sangita, Kalpana sangita and Manodharma*. Private Communication, 2007-2011.

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

[7] L. Lu, H. Jiang, and H. Zhang, "A robust audio classification and segmentation method," in *International ACM Multimedia Conference*, Canada, September 2001, pp. 203–211.

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

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

[10] H. Wang, D. Zhang, and K. Shin, "Syn-dog: Sniffing syn flooding sources," in *ICDCS*, Bangalore, India, July 2002, pp. 421 – 428.

[11] H. Liu and M. S. Kim, "Real-time detection of stealthy ddos attacks using time-series decomposition," in *ICC*, Bangalore, India, July 2010, pp. 1 – 6.

[12] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," in *EUROSPEECH'97*, 1997, pp. 1895–1898.