Segmentation of Carnatic Music Items using KL2, GMM and CFB Energy Feature

Krishnaraj Sekhar PV
Dept. of Computer Science & Engineering
Indian Institute of Technology, Madras
pvkrajpv@gmail.com

Sridharan Sankaran
Dept. of Computer Science & Engineering
Indian Institute of Technology, Madras
sridharan.sankaran@gmail.com

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

***Abstract*—Every Carnatic music concert is made up of many musical items. Every musical item has a lyrical composition (*kriti*) which can be optionally preceded by an *ālāpanā* segment. The duration of the *ālāpanā* along with the *rāgā* in which the *ālāpanā* has been rendered is a strong indication of an artist's creativity and musical knowledge. Hence automatic segmentation of an item to extract the *ālāpanā* segment is of great value in qualitative assessment of a concert. Segmenting a musical item into *ālāpanā* and *kriti* has applications in musical retrieval. To find the boundary between *ālāpanā* and *kriti*, KL2 distance on Cent Filterbank Energy feature is used that locates change in timbre property. A GMM is used to verify the boundary. To further improve the accuracy of segmentation, rules based on musical domain knowledge are automatically applied. Using this approach a frame-level accuracy of 91.34% was obtained.**

I. Introduction

Structural segmentation of musical items directly from audio is a well-researched problem in the literature. With the ever increasing volume of available digital music, efficient storage, indexing and retrieval has become an issue. Segmentation of a musical item into its structural components has several applications. The segments can be used to index the audio for music summarisation, searching, browsing the audio (especially when an item is very long) and recommendation. With content distribution through online portals such as iTunes.com, there is a definite need to allow users to listen to samples of various parts of the song before the song is purchased.

Most approaches to segmentation have relied upon statistical methods. In [1] segmentation is proposed based on significant change in statistical properties. Sheh [2] proposes to use EM-based HMM for chord-based segmentation. Rhodes [3] incorporates the expected segment duration as an explicit prior probability distribution in a Bayesian framework for audio segmentation.

Non machine learning approaches have primarily used time frequency features to identify segment boundaries. In [4], 12-dimensional chroma vectors are extracted at the frame-level, a similarity measure is performed between the segments and then the segments are agglomerated to determine the chorus segments. Serra [5] uses 12-dimensional enhanced-chroma features. Goto [4], extracts 12-dimensional chroma vectors at the frame-level, performs a similarity measure between the segments and then agglomerates the segments to determine the chorus segments. While these approaches have been attempted to segment Western music compositions, the task of segmenting an item into *ālāpanā* and *kriti* in Carnatic music involves differentiating between the textures of the music during *ālāpanā* and *kriti*. While the *kriti* segment involves both melody and rhythm and hence includes the participation of percussion instruments, the *ālāpanā* segment involves only melody contributed by lead performer and the accompanying violinist.

In [6] segmentation of a full length concert or an individual item using applause as a boundary is attempted. While this approach is suitable when applauses are present and audible, alternate algorithms have to be explored.

The objective of this paper is to come up with an approach to identify the ālapana and *kriti* segments using Cent filterbank based features, KL2 distance metric and GMM.

The rest of this paper is organised as below. Section II explains the significance of *ālāpanā* and *kriti* segments in Carnatic music. Section III briefly outlines why MFCC is not suitable for our purpose and why CFB based feature was chosen. Section IV describes the various steps involved in extracting Cent Filterbank (CFB) energy feature. Section V describes the KL2 measure and its use in segmentation . Section VI describes in detail our segmentation approach. Section VII tabulates and discusses the results.

II. Characteristics of segments in Carnatic Music

Carnatic music is a classical music tradition widely performed in the southern part of India. *Rāgās* (melodic modes), *tālas* (repeating rhythmic cycle) and lyrics form the three pillars on which Carnatic music rests. A typical Carnatic music concert varies in duration from 90 mins to 3 hours and is made up of a succession of musical items. These items are standard lyrical compositions (*kritis*) with melodies set to specific *rāgās* and rhythm structure set to specific *tālas*.The *kritis* can be optionally preceded by *ālāpanā*.

Ālāpanā is a way of rendition to explore the features and beauty of a *rāgā*. *Ālāpanā* in Sanskrit means a dialog. Since *ālāpanā* is purely melodic with no lyrical and rhythmic components, it is best suited to bring out the various facets of a *rāgā*. The performer brings out the beauty of a *rāgā* using creativity and internalised knowledge about the grammar of

the *rāgā*. During *ālāpanā*, the performer improvises each note or a set of notes gradually gliding across octaves, emphasising important notes thereby evoking the mood of the *rāgā*. After the main artist finishes the *ālāpanā*, optionally the accompanying violinist may perform an *ālāpanā* in the same *rāgā*.

The *kriti*s are central to any Carnatic music concert. Every *kriti* is a confluence of 3 aspects —lyrics, melody and rhythm. Every musical item in a concert will have the mandatory *kriti* segment and optionally *ālāpanā* segment. The syllables of a lyrics of the *kriti* go hand in hand with the melody of the *rāgā* thereby enriching the listening experience. The lyrics are also important in Carnatic music. While the *rāgā* evokes certain emotional feelings, the lyrics further accentuate it, adding to the aesthetics and listening experience.

III. The Choice of Feature

It has been well established in [6] and [7] that MFCC features are not suitable for modelling music analysis tasks where there is a dependency on the tonic. When MFCCs are used to model music, a common frequency range is used for all musicians, which does not give the best results when variation in tonic is factored in. With machine learning techniques, when MFCC features are used, training and testing datasets should have the same tonic. This creates problems when music is compared across tonics. To address the issue of tonic dependency, a new feature called cent filterbank (CFB) energies was introduced in [6]. Hence, modelling of Carnatic music using cent filter-bank based features that are normalised with respect to the tonic of the performance is the preferred approach for this paper.

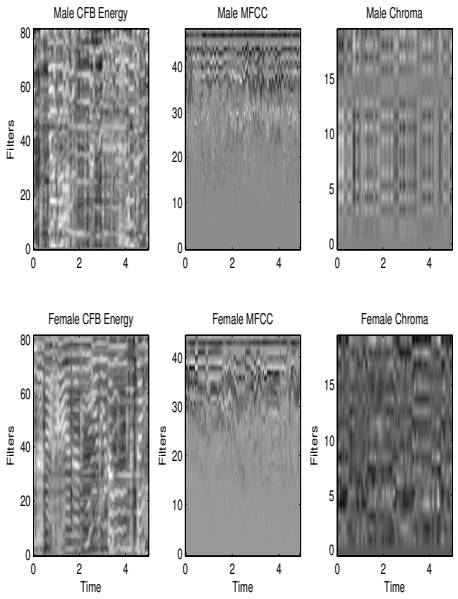

Fig. 1. Time-frequency template of *kriti* segment using various features

IV. CFB Energy Feature Extraction

As mentioned earlier, notes that make up a melody in Carnatic music are defined with respect to the tonic. The lead performer chooses the tonic and the accompanying instruments are tuned to the same tonic . Even for the same musician, tonic can vary across concerts. Nevertheless, the tonic chosen for a concert is maintained throughout the concert using an instrument called the *tanpura* (drone). The analysis of a concert therefore should depend on the tonic. The tonic ranges from 160 Hz to 250 Hz for female and 100 Hz to 175 Hz for male singers. Tonic normalisation in CFB removes the spectral variations as evident in Fig. 1. CFB energy feature extraction is carried out as below:

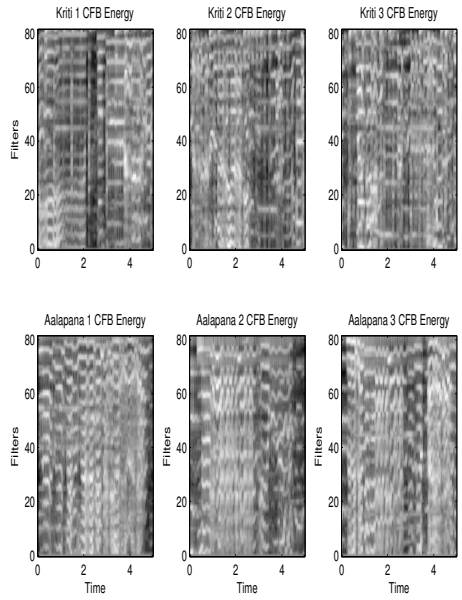

Fig. 2. Time-frequency template of *kriti* Vs. *ālāpanā* segment using CFB Energy feature

1) The audio signal is divided into frames of 100ms.
2) The short-time DFT is computed for each frame.
3) The frequency scale is normalised by the tonic. The cent scale is defined as: Cent = 1200 . log_2 (f / tonic)
4) Six octaves corresponding to [–1200 : 6000] cents are chosen for every musician, considering the rich harmonics involved in musical instruments.
5) The cent normalised power spectrum is then multiplied by a bank of 80 filters that are spaced uniformly in the linear scale to account for the harmonic of pitch. The choice of 80 filters is based on experimentations in [6].
6) The filterbank energies are computed for every frame and used as a feature after removing the bias.

CFB energy features were extracted for every 100 ms of the musical item, with a shift of 10 ms. Thus, a 80 dimensional feature is obtained for every 10 ms of the item, resulting in N feature vectors for the entire item. Fig. 2 shows the CFB Energy extracted for 3 segments of *ālāpanā* and *kriti* of 5 seconds each.

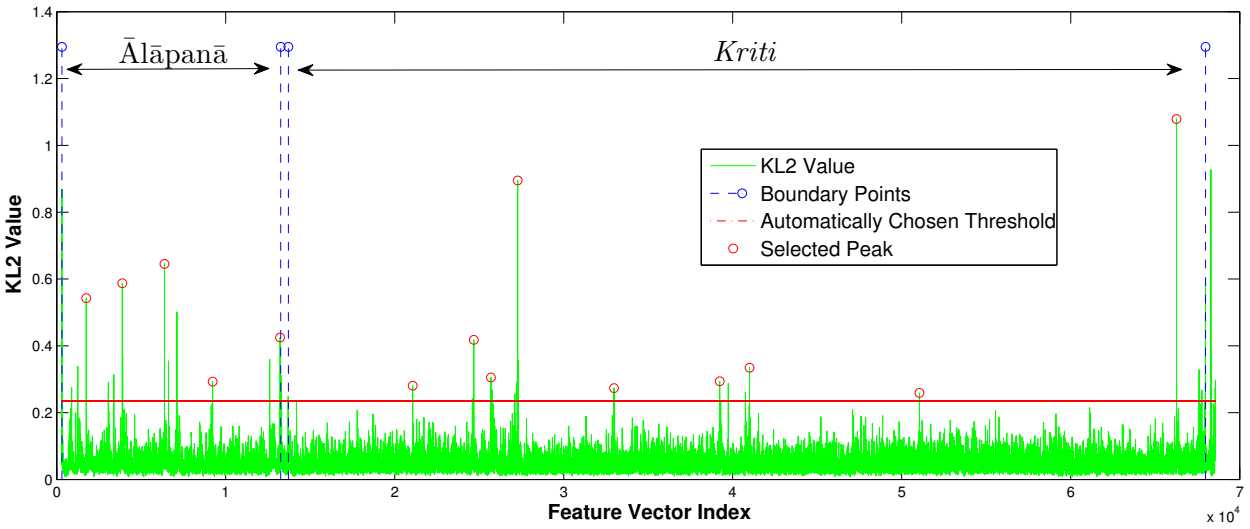

Fig. 3. KL2 Values and possible segment boundaries.

V. KL2 Distance Measure

Metric based segmentation is a popular technique for segmentation. It relies on the computation of a distance between two acoustic segments to determine whether they have similar timbre or not. Change in timbre is an indicator of change in acoustic characteristics such as speaker, musical instrument, background ambience etc. There are broadly 2 kinds of distance measures that can be used. One is statistics based distance and the other is likelihood based.

KL Divergence is an information theoretic non-symmetric measure that gives the difference between two probability distributions P and Q. The larger this value, the greater the difference between these PDFs. It is given given by:

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i)\,\log\frac{P(i)}{Q(i)}. \tag{1}$$

As mentioned in [8], since $D_{\mathrm{KL}}(P\|Q)$ measure is not symmetric, it can not be used as a distance metric. Hence its variation, KL2 metric is used here for distance computation. It is defined as follows

$$KL2 = D_{\mathrm{KL}}(P\|Q) + D_{\mathrm{KL}}(Q\|P) \tag{2}$$

Since there is a considerable change in timbre from *ālāpanā* to *kriti* due to presence of percussion instrument(s) in the *kriti* segment, we can model *ālāpanā* and *kriti* as two probability density functions and use KL2 distance metric to check if there is a change in acoustic properties between adjacent segments. The filterbank energies computed in every frame are normalised using the following formula:

$$FBE[i] = FBE[i]/\sum_{1}^{80} FBE[i]. \tag{3}$$

This behaves like a probability density function. KL2 distance is computed between adjacent frames to determine the divergence between two adjacent spectra.

VI. Segmentation approach

A. Boundary Detection

In order to detect the boundary separating *ālāpanā* and *kriti*, individual feature vector needs to be labelled. One naive approach to find the boundary would be to label each and every feature vector. Since each feature vector corresponds to 10 ms, there is a need to label too many feature vectors for the entire musical item. Moreover, there would be small intervals of time during the *kriti*, when percussion content would be absent either due to inter-stroke silence or due to aesthetic pauses deliberately introduced by the percussionist.

So, a better approach would be to extract a segment of feature vectors from the item and try to label the segment as a whole. Hence, finding the boundary between *ālāpanā* and *kriti* would involve:

- Iterate over the N feature vectors, one at a time.
- Consider a segment of specified length to the left and right of the current feature vector.
- Use a machine learning technique to label these two segment as a whole. This reduces the resolution of the segmentation process to the segment length.
- Use music domain knowledge to correct and agglomerate the labels to find the boundary between *ālāpanā* and *kriti*.

This approach is computationally intense. To further improve the efficiency of this process, we have to reduce the search space for the boundary. The following approach using KL2 was used:

- Iterate over the N feature vectors, one at a time.

- Consider a window of length 500 feature vectors (5 seconds), W_n, where n denotes the starting position. $n = 1, 2..N - 500$
- Average the density function obtained earlier for the entire window length.
- Calculate KL2 distance between 2 successive windows, W_n and W_{n+1}.
- Larger values of KL2 distance denote large change in distribution.
- A threshold was automatically chosen such that, there is 3 seconds spacing between adjacent peaks of KL2 values. This is to prevent the algorithm from generating too many change points. The choice of 3 seconds was empirically arrived at, as a trade-off between accuracy and efficiency.
- The peaks extracted will correspond to a array of K possible boundaries $B = [b_1, b_2, .., b_K]$ between *ālāpanā* and *kriti*.

Fig. 3 shows the output of the algorithm described above.

B. Boundary verification using GMM

From the *K* possible boundary values, the actual boundary between *kriti* and *ālāpanā* needs to be identified. In order to verify the boundaries GMMs were used. GMMs were trained using CFB energy features after applying DCT for compression. GMMs were trained for both the classes *kriti* and *ālāpanā* using a training dataset with 32 mixtures per class. The approach is as follows:

- A window of length 1000 feature vectors (10 Seconds) was extracted to the left and right of the possible boundary points, B.
- Labels for left segments, LSL and the right segments, RSL were computed using GMMs.

C. Label smoothing using Domain Knowledge

Now, using the set of possible boundaries, B and their left and right segment labels, LSL and RSL, we need to assign the label for individual feature vector, L. The following approach was used to find L.

$$L[n] = \begin{cases} LSL[1], \text{if } 1 \leq n \leq B[1] \\ RSL[k], \text{if } B[k] < n < (B[k] + B[k+1])/2 \\ \qquad\qquad (k = 1..K-1) \\ LSL[k], \text{if } (B[k-1] + B[k])/2 \leq n \leq B[k] \\ \qquad\qquad (k = 2..K) \\ RSL[K], \text{if } B[K] < n \leq N \end{cases}$$

Domain information was used to improve the results. To agglomerate the labels, a smoothing algorithm was used as described below.

- An item can have utmost 2 segments-*ālāpanā* and *kriti*.
- If present, *ālāpanā* must be atleast 30 seconds long.
- *Kriti* may be preceded by *ālāpanā*, and not vice versa.
- If a smaller segment of a particular label (*ālāpanā* or *kriti*), was identified in between two larger segments of different label, then the smaller segment is relabelled and merged with the adjacent larger segments.

VII. Experimental Results

A. Dataset Used

Experiments were conducted on 40 concerts. The details of the dataset used is given in Table I. Durations are given in approximate hours (h).

TABLE I
DIVISION OF DATASET.

	Male	Female	Total
No. of Artists	11	11	22
No. of Concerts	26	14	40
No. of items with *ālāpanā*	95	59	154
No. of Items without *ālāpanā*	104	63	167
Total no. of items	199	122	321
Total duration of *kriti*	30 h	18 h	48 h
Total duration of *ālāpanā*	12 h	7 h	19 h
Total duration of Concerts	43 h	27 h	70 h

B. Results

Experiments were performed using both MFCC and CFB based features. Two metrics were used to calculate the accuracy of segmentation —frame-level accuracy and item classification accuracy. As mentioned in Section II, any musical item in a concert can be, a *kriti* preceeded by an *ālāpanā* or without *ālāpanā*. Depending on whether the *ālāpanā- kriti* boundary was properly detected, the item classification results were arrived at.

1) With CFB Energy Feature: Table II shows the confusion matrix for the frame-level classification using CFB based feature. Table III shows the performance for the frame-level classification.

TABLE II
CONFUSION MATRIX: FRAME-LEVEL LABELLING

	kriti	*ālāpanā*
kriti	1,64,11,759	7,77,804
ālāpanā	13,16,925	56,87,274

TABLE III
PERFORMANCE: FRAME-LEVEL LABELLING

	kriti	*ālāpanā*
Precision	0.9257	0.8797
Recall	0.9548	0.8120
F-measure	0.9400	0.8445
Accuracy	0.9134	

Table IV shows the confusion matrix for the item classification using CFB based feature. Table V shows the corresponding performance for the item classification.

TABLE IV
CONFUSION MATRIX: ITEM CLASSIFICATION

	Without *ālāpanā*	With *ālāpanā*
Without *ālāpanā*	155	12
With *ālāpanā*	26	128

TABLE V
PERFORMANCE: ITEM CLASSIFICATION

	Without *ālāpanā*	With *ālāpanā*
Precision	0.8564	0.9143
Recall	0.9281	0.8312
F-measure	0.8908	0.8707
Accuracy	0.8816	

2) With MFCC Feature: Table VI shows the confusion matrix for the frame-level classification using MFCC feature. Table VII shows the performance for the frame-level classification.

TABLE VI
CONFUSION MATRIX: FRAME-LEVEL LABELLING

	kriti	*ālāpanā*
kriti	1,39,58,342	32,31,221
ālāpanā	52,60,214	17,43,985

TABLE VII
PERFORMANCE: FRAME-LEVEL LABELLING

	kriti	*ālāpanā*
Precision	0.7263	0.3505
Recall	0.8120	0.2490
F-measure	0.7668	0.2912
Accuracy	0.6490	

Table VIII shows the confusion matrix for the item classification using MFCC feature. Table IX shows the corresponding performance for the item classification.

TABLE VIII
CONFUSION MATRIX: ITEM CLASSIFICATION

	Without *ālāpanā*	With *ālāpanā*
Without *ālāpanā*	145	22
With *ālāpanā*	115	39

TABLE IX
PERFORMANCE: ITEM CLASSIFICATION

	Without *ālāpanā*	With *ālāpanā*
Precision	0.5577	0.6393
Recall	0.8683	0.2532
F-measure	0.6792	0.3628
Accuracy	0.5732	

It can be observed that, using this approach, frame level labelling accuracy of 91.34% and item classification accuracy of 88.16% has been achieved using CFB Energy feature. Where as using the MFCC feature, frame level labelling accuracy of 64.90% and item classification accuracy of 57.32 has been achieved.

VIII. CONCLUSION

CFB based features are most suited to analyse Indian Classical music that depends on tonic which varies across musicians. Detecting timbre changes using KL2 distance measure and validating that with a machine learning approach form a robust combination for segmentation based on changes in timbre properties. Such a segmentation into *ālapana* and *kriti* can feed as the input to further finer segmentation activities such as segmenting the *ālapanā* into vocal *ālapanā* and violin *ālapanā* , segmenting the *kriti* into *pallavi, anupallavi, caranam* [7], retreiving the *tani āvartanam* (solo percussion) segment from a concert etc.

ACKNOWLEDGMENT

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).

REFERENCES

[1] A. L. Berenzweig and D. W. Ellis, "Locating singing voice segments within music signals," in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*. IEEE, 2001, pp. 119–122.

[2] A. Sheh and D. P. Ellis, "Chord segmentation and recognition using em-trained hidden markov models," *ISMIR 2003*, pp. 185–191, 2003.

[3] C. Rhodes, M. A. Casey, S. Abdallah, M. Sandler *et al.*, "A markov-chain monte-carlo approach to musical audio segmentation," 2006.

[4] M. Goto, "A chorus-section detecting method for musical audio signals," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 5. IEEE, 2003, pp. V–437.

[5] J. Serra, M. Muller, P. Grosche, and J. L. Arcos, "Unsupervised music structure annotation by time series structure features and segment similarity," *Multimedia, IEEE Transactions on*, vol. 16, no. 5, pp. 1229–1240, 2014.

[6] P. Sarala and H. A. Murthy, "Inter and intra item segmentation of continuous audio recordings of carnatic music for archival." in *ISMIR*, 2013, pp. 487–492.

[7] S. Sankaran, K. Sekhar, and H. A. Murthy, "Automatic segmentation of composition in carnatic music using time-frequency cfcc templates," in *Proceedings of 11th International Symposium on Computer Music Multidisciplinary Research (CMMR), Plymouth, UK*, 2015.

[8] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA speech recognition workshop*, vol. 1997, 1997.