Automatic Segmentation of Composition in Carnatic Music Using Time-Frequency CFCC Templates

Sridharan Sankaran, Krishnaraj Sekhar PV, and Hema A Murthy *
Indian Institute of Technology, Madras

Abstract. The Pallavi in a composition in Carnatic music is akin to the chorus or refrain in Western music albeit with a key difference; the Pallavi (or part of it) can be rendered with a number of variations in melody, without any change in the lyrics, and is repeated after each segment of a composition.

The objective of this paper is to automatically segment a composition using a partial pallavi query as the template. Cent filterbank [7] based energies and slope features were extracted for both the query and the entire composition. A sliding window of the query duration was used, and the correlation between the mean normalised t-f template and the composition segment within the sliding window was computed. The locations of maximum correlations correspond to Pallavi repetitions. We were able to segment 66% of the compositions accurately. Failures were primarily due to large melodic variations in the repetitions.

1 Introduction

Structural segmentation of compositions directly from audio is a well researched problem in the literature. Segmentation of a composition into its structural components has several applications. The segments can be used to index the audio for music summarisation and browsing the audio (especially when an item is very long). Most approaches to segmentation have primarily relied on machine learning techniques [1,2,4–6], including techniques that use the hierarchical structures in music [4]. Frames are first classified based on their audio properties, and then agglomerated to find the homogeneous segments. This is followed by beat tracking, and merging of adjacent frames that might belong to the same segment. Logan [5] uses clustering and HMMs to categorize short segments based on their acoustic features, and the most frequent category is identified to be the chorus. Rhodes [6] incorporates the expected segment duration as an explicit prior probability distribution in a Bayesian framework for audio segmentation. In [2], using HMMs, the regions in music audio that exhibit ‘texture’ with steady statistical properties are identified as chorus segments.

* 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).
Non-machine learning approaches [3,8] use chroma based time-frequency features to identify the chorus segments in an audio. In [3], 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. In [8], 12-dimensional enhanced-chroma features are used. Although this technique is capable of detecting modulated repetitions, the fundamental assumption is that these repetitions are restricted to the specific pitch class. 12 different types of modulated repetitions are therefore analysed for repetitions.

While these techniques have been attempted for Western music where the repetitions have more or less static time-frequency melodic content, finding repetitions in improvisational music is a difficult task. In Indian music, the melody content of the repetitions varies significantly during repetitions within the same composition due to the improvisations performed by the musician. A musician’s rendering of a composition is considered rich, if (s)he is able improvise and produce a large number of melodic variants of the line while preserving the grammar and identity of the composition and the rāga. Further, the same composition when rendered by different musicians can be sung in different tonics. Hence matching a repeating pattern of a composition across recordings of various musicians requires a tonic-independent approach.

The objective of this paper is primarily to segment a composition into its constituent parts using Pallavi (or a part of Pallavi) of the composition as the query template (see Section 2 for a description of the structure of compositions in Carnatic music).

In our initial attempts, the query was first manually extracted from 75 popular Carnatic music compositions. In 65 of these compositions, the lead artist was a vocalist accompanied by a violin and one or more percussion instruments while in the remaining 10 compositions, an instrumentalist was the lead artist accompanied by one or more percussion instruments. We then used time-frequency motifs to locate the repetitions of this query in the composition. Cent-filterbank based features [7] were used to obtain tonic normalised features. Although the Pallavi line of a composition can be improvised in a number of different ways with variations in melody, the timbral characteristics and some parts of the melodic characteristics of the Pallavi query do have a match across repetitions. The composition is set to a specific tala (rhythmic cycle), and lines of a Pallavi must preserve the beat structure. With these as the cues, given the Pallavi or a part of it as the query, an attempt was made to segment the composition. The time-frequency motif was represented as a matrix of mean normalised cent filterbank based features. Cent filterbank based energies and slope features were extracted for the query and the entire composition. The correlation co-efficients between the query and the composition were obtained while sliding the query window across the composition. The locations of the peaks of correlation indicate the locations of the Pallavi. We also attempted to extract the query automatically for all the compositions using the approach described in Section 4.2, and cross-checked the query length with the manual approach.
The rest of the paper is organised as follows. Section 2 discusses the characteristics of Pallavis in Carnatic music. Section 3 discusses the representation of the query as a time series or as a trajectory in the time-frequency plane. It also discusses cent filterbank based features that are relevant for Carnatic music. Section 4 discusses the approach and results obtained. The results presented in this paper are promising, in that it can indeed enable segmentation of not only Carnatic compositions but also music of other genres such as Western and Hindustani music.

2 Characteristics of Pallavis in Carnatic Music

Carnatic music is a classical music tradition widely performed in the southern part of India. In Carnatic music, a song or composition typically comprises of 3 segments—Pallavi, Anupallavi and Caranam—although in some cases there can be more segments due to multiple Caranam segments. While many artists render only 1 Caranam segment (even if the composition has multiple Caranam segments), some artists do render multiple Caranam or all the Caranams.

Segmentation of compositions is important both from the lyrical and melody aspects. In the Pallavi, a musical theme is initiated from a certain angle, developed a little further in the Anupallavi and further enlarged in the Caranam, maintaining a balanced sequence - one built upon the other. Pallavi, being the first segment, also plays a major role in presenting a gist of the rāga, which gets further elaborated in Anupallavi and Caranam. Similar stage-by-stage development from lyrical aspect can also be observed. An idea takes form initially in the Pallavi, which is the central lyrical theme, further emphasised in the Anupallavi and substantiated in the Caranam.

The Pallavi or a part of Pallavi is repeated multiple times with improvisation for the following reasons: 1) The central lyrical theme that gets expressed in the Pallavi is highlighted by repeating it multiple times, 2) the melodic aspects of the rāga and the creative aspects of the artist (or the music school) jointly get expressed by repetitions of Pallavi. These set of improvisations in a given composition also stand out as signatures to identify an artist or the music school. Since Pallavi serves as a delimiter or separator between the various segments, locating the Pallavi repetitions also leads to knowledge of the number of segments in a composition (≥3) as rendered by a certain performer.

A commonly observed characteristic of improvisation of Pallavi (or a part of it) is that for a given composition, a portion (typically half) of the repeating segment will remain more or less constant in melodic content through out the composition while the other portion varies from one repetition to another. For instance, if the first half of the repeating segment remains constant in melody, the second half varies during repetitions and vice-versa.
3 Time Frequency Templates

The spectrogram is a popular time-frequency representation. The repeated line of a Pallavi is a trajectory in the time-frequency plane. Fig. 1 shows spectrograms of the query and the matched and unmatched time-frequency segments of the same length in a composition using linear filterbank energies. The frequency range is set appropriately to occupy about 5 octaves for any musician. Although the spectrogram does show some structure, the motifs corresponding to that of the query are not evident. This is primarily because the motif is sung to a specific tonic. Therefore the analysis of a concert also crucially depends on the tonic.

![Fig. 1. Time-frequency template of music segments using FFT spectrum (X axis: Time in frames, Y axis: Frequency in Hz)](image)

3.1 The Choice of Features

MFCC features are seldom used for music modeling tasks, and chroma features are preferred (Ellis, 2007; Müller et al., 2005). The basic issue in using MFCCs for music processing is that the rendering of a given item depends on the key. When MFCCs are used to model music, a common frequency range is used for all musicians, which does not give the best results.

To address this issue, chroma filter-banks were suggested and are extensively used in music, especially in Western classical music. Chroma is also based on the octave. In chroma, the key of the composition is used to normalise the music w.r.t. the tonic. All octaves are folded to a single octave. The chroma filter-banks discussed in (Ellis, 2007), are primarily for Western classical music; where the
scale is equi-temperament and is characterised by a unique set of 12 semitones, subsets of which are used in performances. As indicated in (Serra et al., 2011), Indian music pitches follow a just intonation rather than an equi-temperament intonation. In (Krishna and Ishwar, 2012), it is shown that just intonation is also not adequate because the pitch histograms across all rāgas of Carnatic music appear to be more or less continuous. To account for this, the chroma filter-banks include a set of overlapping filters in the implementation. 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 the tonic, it is not enough for Indian music as the folding over to a single octave is not appropriate. This is due to the fact that the characteristic traits of the musician/melody may be observed while traversing across octaves. Furthermore, the filters span less than a semitone to account for the fact that in Indian music, two adjacent svaras (notes) need not be separated by a semitone.

Constant Q-Transform (CQT) based filter-banks were proposed (Brown and Puckette, 1992) for music instrument identification. In CQT, the filter-banks are placed on the log₂ (f/2) frequency scale as defined in the equation:

\[ \text{CQT} = 1200 \times \log_2 \left( \frac{f}{2} \right) \]

The issues with tonic variant (MFCC and CQT) and tonic invariant (Chroma and pitch) features for Carnatic music processing tasks are summarised as follows:

MFCC: A common frequency range is used for all musicians, which is not appropriate as melody characteristics change with a corresponding change in the tonic.

CQT: In CQT based features, the frequency range needs to be carefully chosen and it is difficult to choose the appropriate frequency for any specific task.

Chroma: Even though these features are normalised with respect to the tonic, the folding over into a single octave is not appropriate as the music characteristics can be across octaves.

Pitch: As Indian music is polyphonic in nature, the singing voice is always accompanied by instruments, and thus, the predominant pitch extraction is a tough task. This can lead to octave errors and errors in pitch values due to noise or due to instruments. In addition, melody characteristics depend not only on the pitch but also on timbre.

To address the issue of tonic dependency, a new feature called cent filterbank energies was introduced in [7]. The fundamental difference between MFCC and Cent filter bank Cepstral Coefficients (CFCC) is that in CFCC, the frequency spectrum of the signal is mapped to the cent scale (after normalising with tonic) as opposed to the Mel scale.

Hence, modeling 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.
3.2 Cent Filterbank Energy

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 tambura (drone). The analysis of a concert therefore depends on the tonic. The tonic ranges from 160 Hz to 250 Hz and 100 Hz to 175 Hz for female and male singers, respectively. CFCC extraction is carried out as below:

1. The audio signal is divided into frames.
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:
   \[ \text{cent} = 1200 \cdot \log_2 \left( \frac{f}{\text{tonic}} \right) \]  
   (1)
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 filters that are spaced uniformly in the linear scale to account for the harmonic of pitch.
6. The filterbank energies are computed for every frame and used as a feature after removing the bias.
7. DCT is applied on filterbank energies and the required co-efficients are retained.

The filterbank energies were computed for both the query and the composition. The time-dependent filterbank energies were then used as a query. Fig. 2 shows a time-frequency template of the query and some matched and unmatched examples from the composition. A sliding window approach was used to determine the locations of the query in the composition. The locations at which the correlation is maximum corresponds to matches with the query. Fig. 4 shows a plot of the correlation as a function of time. The location of the peaks in the correlation, as verified by a musician, correspond to the locations of the repeating query.

3.3 Cent Filterbank Slope

In Fig. 2, it can be seen that the presence of the strokes due to the mridangam (a percussion instrument that accompanies the main artist) destroy the motif. To address this issue, cent filterbank based slope was computed along frequency. Let the vector of log filter bank energy values be represented as \( F_i = (f_{1,i}, f_{2,i}, \ldots, f_{n_f,i})^T \), where \( n_f \) is the number of filters. Mean subtraction on the sequence \( F_i \), where \( i = 1, 2, \ldots, n \) is applied as before. Here, \( n \) is the number of feature vectors in the query. To remove the effect of percussion, slope
values across consecutive values in each vector $F_i$ are calculated. Linear regression over 5 consecutive filterbank energies is performed. A vector of slope values $s = (s_{1,i}, s_{2,i}, ..., s_{F-1,i})^t$ for each frame of music is obtained as a result.

Fig. 3 shows a plot of the time-dependent query based on filter bank slope and corresponding matched and unmatched segments in the composition. One can observe that the motifs are significantly emphasised, while the effect of percussion is almost absent.

4 Experimental Results

The experiments were performed primarily on Carnatic music, though limited experiments were done on other musical genres - Hindustani (north Indian classical music) and Western music.

For Carnatic music, a database of 75 compositions by various artists was used. The database comprised of compositions rendered by a lead vocalist or lead instrumentalists, the instruments being flute, violin and veena. The tonic information was determined for each composition. Cent filterbank based energies and cent filterbank based slope features were extracted for each of these compositions and used for the experiments. For every 100 millisecond frame of the composition, 80 filters were uniformly placed across 6 octaves (the choice of number of filters was experimentally arrived at to achieve the required resolution). The correlation between the query and the moving windows of the composition was computed as indicated in Section 3.

The algorithms used are described below:
4.1 Finding Match with a Given Query

The query for each composition was extracted manually and the CFCC features were computed. Then the Algorithm 1 was used for both cent filterbank based cepstral and slope features.

As we can see in Fig. 4 and Fig. 5, the identified repeating patterns clearly stand out among the peaks due to higher correlation. The spectrogram of the
Algorithm 1 Composition-Query Comparison

1. Extract CFCC Energy feature for the composition and for the query.
2. Using a sliding window approach, move across the composition in one frame steps.
3. Extract composition segments of length same as query at each step.
4. Compute the correlation between extracted composition segments and the query segment.
5. Locate the positions which give high correlation, which are the matches.
6. Repeat the above steps for CFCC slope feature also.

The experiments were repeated with MFCC and chroma features. The comparative performance of these three features is tabulated in Section 4.3.
4.2 Automatic Query Detection

It is possible to find the query automatically if query is found at the beginning of the composition. This is indeed true for Carnatic music as the composition rendering starts with the *Pallavi*. The approach mentioned in Algorithm 2 was used for the automatic query detection.

As mentioned in the algorithm, the correlations of the composition with progressively increasing query lengths can be seen in Fig. 7 and Fig. 8. The product of these correlation values is computed at each time instance within the composition and result is plotted in Fig. 9. As we can see, the unwanted spurious peaks have all been smoothed out resulting in clear identification of the actual query length.

![Figure 7](image_url)

**Fig. 7.** Intermediate output (I) of the automatic query detection algorithm using slope feature

---

**Algorithm 2 Automatic Query Detection**

1: Extract CFCC Energy feature for the composition.
2: Extract segments of varying lengths of 0.5 to 3 seconds (50 to 300 frames) in steps of 0.5 seconds (50 frames).
3: For each of these segments, considering this as the query, calculate the correlation as in Algorithm 1.
4: Multiply the above computed correlations corresponding to each frame.
5: Look for the first significant peak, which corresponds to the query length.
6: Repeat the above steps for CFCC Slope feature also.
It was observed that for all the 75 compositions, the durations of their queries calculated using the automatic method matched closely with the actual query length, thereby producing similar segmentation results. A subset of the results is tabulated in Table 2.

**Table 1.** Comparison between various features

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Total Songs</th>
<th>Successfully Segmented</th>
<th>% Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFCC</td>
<td>75</td>
<td>35</td>
<td>46.47</td>
</tr>
<tr>
<td>MFCC</td>
<td>75</td>
<td>23</td>
<td>30.67</td>
</tr>
<tr>
<td>Chroma with overlap</td>
<td>75</td>
<td>17</td>
<td>22.67</td>
</tr>
<tr>
<td>Chroma without overlap</td>
<td>75</td>
<td>8</td>
<td>10.67</td>
</tr>
</tbody>
</table>

### 4.3 Performance of Various Features

We compared the performance of CFCC features with that of MFCC and Chroma Filterbank based Cepstral Coefficients. MFCC features were extracted using the HTK toolkit. Features were computed for 100 ms frames with a shift of 10 ms. 80 filters were used between 0 Hz to 22 KHz. Chroma and Chroma with overlap features were computed for 100ms frame with a shift of 10 ms. 24 filters were used corresponding to the 12 semi-tones of 100 cents each.
Fig. 9. Final output of the automatic query detection algorithm using slope feature.

Table 2. Manual vs automatic query extraction (CFB Energy: Cent filter bank cepstrum, CFB Slope: Cent filterbank energy slope). Time is given in second.

<table>
<thead>
<tr>
<th>Composition</th>
<th>Manual</th>
<th>CFB Energy</th>
<th>CFB Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>4.90</td>
<td>5.09</td>
<td>5.08</td>
</tr>
<tr>
<td>S2</td>
<td>12.07</td>
<td>12.52</td>
<td>12.52</td>
</tr>
<tr>
<td>S3</td>
<td>6.03</td>
<td>6.10</td>
<td>6.11</td>
</tr>
<tr>
<td>S4</td>
<td>11.84</td>
<td>9.75</td>
<td>11.73</td>
</tr>
<tr>
<td>S5</td>
<td>8.71</td>
<td>8.79</td>
<td>8.76</td>
</tr>
<tr>
<td>S6</td>
<td>5.58</td>
<td>5.75</td>
<td>5.8</td>
</tr>
<tr>
<td>S7</td>
<td>8.50</td>
<td>8.59</td>
<td>8.59</td>
</tr>
<tr>
<td>S8</td>
<td>4.65</td>
<td>4.86</td>
<td>4.85</td>
</tr>
<tr>
<td>S9</td>
<td>11.79</td>
<td>11.73</td>
<td>8.48</td>
</tr>
<tr>
<td>S10</td>
<td>12.84</td>
<td>12.92</td>
<td>12.85</td>
</tr>
</tbody>
</table>
The comparative performances are shown in the Table 1. It was found that CFCC was performing better than the other features. Hence we used this feature along with domain knowledge to improve the overall accuracy of segmentation.

4.4 Detailed Findings Using CFCC Features

Out of 75 compositions, 35 compositions were correctly segmented into Pallavi Anupallavi and Caranam using full query length when compared with the ground truth marked by a musician. The segment lengths were incorrectly detected in the remaining compositions primarily due to the following reasons:

1. False negatives: It is quite possible that while repeating the Pallavi (query) with melodic variations, the artist may repeat only a part of the query some times. In such cases, correlation of the partial match will be low. Also often, the melodic content may vary drastically during repetitions leading to low correlation with the reference query.

2. False positives: Some portions of a composition (such as Anupallavi / Caranam) may have similar melodic content to that of the Pallavi query, though the lyrics of these portions can be entirely different. These portions result in higher correlation due to melodic similarity.

In order to address the false negative results, we further experimented with half the length of those queries. In Carnatic music, though the Pallavi is repeated with various melodic variations, usually either the 1st half or the 2nd half of the query remains static in melodic content. In other words, if the 1st half of the query undergoes melodic variation during repetitions, the 2nd half remains melodically invariant and vice versa. Taking this as a cue, we experimented by considering the 1st half and 2nd half of the original query as the new queries.

The results showed that using one of the two half length queries, better correlation of matched segments was obtained thereby increasing our segmentation success from 35 to 48 compositions resulting in 64% success rate. Fig. 10 shows the correlation plot of a composition with the full query and the corresponding half query.

To address the false positives, we used the rhythm cycle. If the query length is L seconds, the repetitions should ideally occur at nL (n = 1,2,3..N), with margins for human errors in maintaining the rhythm. Any instances of elevated correlation that are not around nL are not likely to be repetition of the query and hence can be discarded. Fig. 11 shows the correlation plot using this approach. As we can see, only the false positive peaks have been discarded, resulting in only the true positive peaks. Using this approach our segmentation success increased from 48 to 50 compositions resulting in 66.66% as the overall success rate. However, this approach will be effective only when the artist maintains the rhythm cycle more or less accurately.

We repeated the experiment on a small sample of Hindustani and Western music compositions - 5 compositions in each genre. Since these genres of music do
Fig. 10. Correlation for full query Vs. half query

Fig. 11. False positive elimination using rhythmic cycle information
not have segments similar to Pallavi, Anupallavi and Caranam; we restricted our experiments to locating the repetitions of a query segment. In the case of Western music, we were able to locate all the repetitions corresponding to the query. This is because of little melodic variation between repetitions in Western music. In the case of Hindustani music, we were able to identify all the repetitions with a false positive rate of 33%. Fig. 12 shows the correlation plot of a Hindustani composition and a Western music composition.

**Fig. 12.** Repeating pattern recognition in other Genres

For those Carnatic music compositions where segmentation results are accurate, it is possible to automatically match the audio segments with the Pallavi, Anupallavi and Caranam lyrics of the composition by looking up a lyrics database. This can enhance the listening experience by displaying the lyrics as captions when the composition is played. We were able to automatically generate captions as SRT (SubRip Text) files for each composition using automatic segmentation and lyrics database lookup.

5 Conclusions

In this paper, we have presented a novel approach to composition segmentation using cent filterbank based features and correlation. This approach is particularly suited for Carnatic music as compared to other fingerprinting algorithms. In Indian music, the same composition can be sung in different tonics. Further,
a number of different variants of the Pallavi can be sung. This can vary from musician to musician and the position of the composition in the entire concert. A large number of variations is an indication that the musician has chosen the particular composition for a more detailed exploration. The segmentation of the composition into Pallavi, Anupallavi and Caranam can be performed using the repeating Pallavi line. The location of these segments will also enable the listener to identify the locations of niraval and svaraprasthara-improvisation segments of a concert that show the prowess of the musician in discovering and presenting the melodic flow of a rāga. Although the results presented in this paper are of a preliminary nature, it can have far reaching applications in terms of indexing and archival of Indian music. Though we have largely confined ourselves to segmenting Carnatic compositions, this approach should equally work well for other musical genres such as Hindustani music (north Indian classical music), Indian film music and Western music.

References