Structural Segmentation of Irish Traditional Music using Chroma at Set Accented Tone Locations

Cillian Kelly
Technological University Dublin, cillian.kelly@tudublin.ie

Mikel Gainza
Technological University Dublin, Mikel.Gainza@tudublin.ie

David Dorran
Technological University Dublin, david.dorran@tudublin.ie

See next page for additional authors

Authors
Cillian Kelly, Mikel Gainza, David Dorran, and Eugene Coyle

Audio Engineering Society

Convention Paper

Presented at the 127th Convention
2009 October 9–12 New York NY, USA

Structural Segmentation of Irish Traditional Music using Chroma at Set Accented Tone Locations

Cillian Kelly[1], Mikel Gainza[1], David Dorran[1], and Eugene Coyle[1]

[1] *Dublin Institute of Technology, Kevin Street, Dublin 2, Ireland*

Correspondence should be addressed to Cillian Kelly (`cillian.kelly@dit.ie`)

ABSTRACT

An approach is presented which provides a structural segmentation of Irish Traditional Music. Chroma information is extracted at certain locations within the music. The resulting chroma vectors are compared to determine similar structural segments. Chroma is only calculated at 'set accented tone' locations within the music. 'Set accented tones' are considered to be impervious to melodic variation and are entirely representative of an Irish Traditional tune. Results show that comparing 'set accented tones' represented by chroma significantly increases the structural segmentation accuracy than when 'set accented tones' are represented by pitch values.

1. INTRODUCTION

1.1. Overview

Structural segmentation of music aims to provide information on the segment boundaries, musical form and semantic labelling of a piece of music. The Segment boundaries are points of significant change within the music. These boundaries are the points at which one structural segment ends and another begins. The musical structure is discovered by determining which structural segments are repeated throughout the music. Semantic labels are meaningful labels attributed to each of the structural segments. For example, the semantic labels used for popular music include verse, chorus etc. The structural segmentation approach proposed in this paper deals specifically with Irish Traditional Music. This music type, by nature is repetitive. There are obvious repeating patterns audible throughout this music type. The difficulties in structurally segmenting this music type are due to melodic variation present between different renditions of repeating structural segments.

The information provided by structurally segmenting music can be used for audio thumbnailing. An audio thumbnail is the most representative segment of a piece of music [1]. Using popular music as an example, the chorus is considered the most representative segment of this genre. Audio thumbnails can be used as short previews of a song provided to users by online music stores. Audio browsing also utilises information provided by an audio structural segmentation. The location of each structural segment within the audio is provided. Using this information, a user may browse directly to the segment of interest within the audio. A further use of structural segmentation information is audio looping. The locations of each structural segment can be used as loop points so that each end of the loop rhythmically correspond.

This paper details a technique for structurally segmenting music into its constituent structural segments. To achieve a structural segmentation, specific notes of the music known as 'set accented tones' are extracted from the music. 'Set accented tones' are considered to remain constant despite the presence of melodic variation [2]. Consequently, using these 'set accented tones' reduces the audio to a set of invariants which are entirely representative of the audio. Chroma information is then extracted at each of the 'set accented tone' locations. Chroma information encodes the prominence of each note of a musical scale at discrete locations within the music. Structural segments are determined based on similarities between the resulting chroma vectors.

Traditional approaches to structurally segment music extract features (e.g. pitch, loudness) from the musical piece and then search for repeating patterns of these features. The technique presented here extends the approach detailed in [3]. The approach is justified and described in detail in the remainder of this paper. This paper is structured as follows: Section 2 provides a literature review of existing approaches towards structurally segmenting music. Certain details of these approaches which make them unsuitable for use with Irish Traditional Music are highlighted. In Section 3 an overview of the musical theory of Irish Traditional Music is provided to increase understanding of the proposed approach. Section 4 provides the implementation details of the proposed approach, including determining the 'set accented tone' locations, calculating chroma vectors at each location and the subsequent matching technique used to determine equivalent segments. The results of this approach are presented in Section 5. Finally, in Section 6, conclusions are outlined and details of future work are provided.

2. LITERATURE REVIEW

Traditional approaches towards structurally segmenting music firstly divide the audio waveform into short segments called frames. Audio features are then extracted for each of these frames. Following this, the frames are grouped together based on the extracted feature values. The resulting groups of frames indicate structural segments within the audio signal. Finally, semantic labels for the resulting structural segments are computed.

In [4], low level audio features, such as zero crossings, spectral flux and spectral centroid are used to describe individual frames of the audio signal. Sequences of audio frames which display similar feature values throughout the signal are grouped together. The resulting groups of frames are considered to be the structural segments of the music. The approach detailed in [4] relies on large changes between sections, when there are very musically smooth transitions between sections the approach performs poorly [4].

In [5] an attempt is made to structurally segment audio using a pitch contour as the extracted audio feature. Firstly, the audio is divided into overlapping frames. Following this, a pitch value is extracted for each of the audio frames. Pitch detection, in [5] is computed via autocorrelation. Following the pitch detection, sections of the audio that have a stable pitch are labeled as notes, resulting in a pitch contour. Once the pitch contour has been determined, Dannenberg attempts to discover melodic phrases that repeat within the audio. Phrases are considered similar if they have approximately the same duration and pitch values. Clustering is utilised in [5] to group similar phrases together. The resulting clusters correspond to a structural segmentation of the music.

A method to identify a chorus of a popular song by identifying repeated segments of an audio signal using chroma as an audio feature is proposed in [1]. Firstly, in [1] the audio signal is divided into frames aligned to the beat positions of the music. A beat-tracking system [6] is employed to dynamically provide a frame segmentation which is specific to each song. The resulting frames are typically between one-quarter and one-half of a second in length. Following this, chroma information is extracted for each frame. Only frequencies in the range of 20 *Hz* to 2 *kHz* are considered in [1]. It is said this range provides a sufficiently rich description of the melodic content. A self-similarity matrix is then computed between frames by correlating the chroma vectors of each frame. To discern similarities among extended regions of the similarity matrix, correlation filtering is applied along the diagonals of the matrix.

Other approaches such as [7] attempt to structurally segment music using Mel Frequency Cepstral Coefficients (*MFCCs*) to describe the audio signal. *MFCCs* are the result of a cosine transform of the logarithm of the short-term energy spectrum expressed on a mel-frequency scale [8]. They are a compact representation of an audio spectrum that takes into account the non-linear human perception of audio. *MFCCs* give an indication of the timbre of the music, and changes in timbre can be detected by using this feature. In [7] distance measures are calculated between the *MFCC* values for each adjacent audio frame. Audio frames which result in a large distance value indicate a change in timbre in the music and therefore provide candidates for segment boundaries.

An approach is proposed in [9] that uses 'beat-synchronous chroma vectors' to determine cover versions of a given song. Chroma information is extracted on each beat location of the music, and entire songs are compared according the resulting chroma vectors.

An approach is presented in [3] which creates a selective pitch contour only using the notes at 'set accented tone' locations. The 'set accented tones' are located on certain beats of the music. Melodic patterns are searched for amongst this reduced representation pitch contour to determine structural segments.

3. IRISH TRADITIONAL MUSIC AND JUSTIFICATION OF APPROACH

The approach presented in this paper deals specifically with the structural segmentation of Irish Traditional Music. For this reason, a certain amount of knowledge of the structure of this music type is required to better understand the proposed method. Structural segments in Irish Traditional Music are referred to as 'parts' and are referred to using upper case letters. Irish traditional tunes are comprised of two or more parts and parts are typically repeated. The tune itself is also usually repeated a number of times in its entirety. A typical structure of an Irish Traditional tune with an 'A' part and a 'B' part would be 'AABBAABB'. However, it is in the nature of this music type that any rendition of a tune is open to interpretation. So the structure of a particular tune could be different depending on the musician who is playing it. It is common that a considerable amount of melodic variation is present between two renditions of the same part within the same performance. Again, the nature of this variation is unique to each musician. Consequently, the problem arises of how similar must two parts be before they are considered to be equivalent.

It is important to note (as will be seen in Section 4.3) that there are a finite number of time signatures present within Irish Traditional Music. These time signatures along with the associated number of 'set accented tones' per part are outlined in Table 1.

The method proposed in this paper uses a set of notes referred to as 'set accented tones' These notes were identified in [2] and are considered to be impervious to melodic variation. For Irish Traditional Music these 'set accented tones' are located on certain beats of the music as illustrated in Figure 1.

Irish Traditional Music is primarily focused on the melody of the music. Indeed, very often the music in question contains *only* melody, there may be no accompanying instruments or rhythm section of any kind. For this reason, harmonic features such as chroma are appropriate for this genre. Using aggregated low level features towards structurally segmenting music relies on the presence of musically different segments throughout the music. This is not ideal for Irish Traditional Music, where the loudness, instrumentation, and rhythm can remain constant

Table 1: *An overview of the time signatures present within Irish Traditional Music and the bars per part and 'set accented tones' per part associated with each time signature.*

Tune Type	Time Signature	Bars Per Part	Set Accented Tones Per Part
Reel Hornpipe March	4/4	8	16
Jig	6/8	8	16
Polka Barn Dance	2/4	8	16
Slip Jig	9/8	4	12
Slide	12/8	4	12
Waltz Mazurka	3/4	8	24

throughout, indeed, it is often the goal of the musician to achieve musically smooth transitions between segments. Utilising timbre as a feature via *MFCCs* would also not be suited to Irish Traditional Music because there may be no timbre changes present within a piece of music of this genre. A pitch contour has also been considered unsuitable for Irish Traditional Music, because a pitch contour encodes all notes of the melody. Consequently, all melodic variation would also be encoded in the pitch contour increasing the difficulty of discerning which segments are equivalent.

The approach proposed in [3] relies on the preprocessing task of pitch detection. Any errors generated by the pitch detector reduces the performance of the subsequent structural segmentation algorithm. Using chroma at 'set accented tone' locations instead of pitch will eradicate the need to determine a single note value. It is proposed in this paper, that chroma information provides a more accurate and rich description of a 'set accented tone' location and will therefore provide more accurate structural segmentation results. Audio frames with similar harmonic content will produce similar chroma information. Thus, the similarity of two audio frames can be determined by computing a distance measure between the chroma vectors representing those frames.

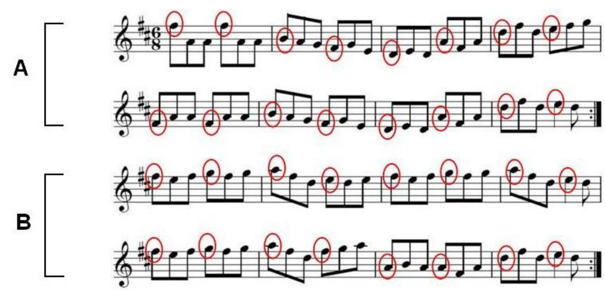

Fig. 1: *An Irish Traditional tune in 6/8 time with an 'A' part and a 'B' part with the 'set accented tones' highlighted.*

4. PROPOSED APPROACH

The structural segmentation method proposed in this paper focuses particularly on Irish Traditional Music and furthers the approach presented in [3]. The following paragraph gives an overview of the proposed approach, a detailed explanation of the approach is provided in the subsequent sections.

The approach presented here structurally segments Irish Traditional Music using chroma information only at 'set accented tone' locations. The locations of the 'set accented tones' are identified using a beat tracker. The 'set accented tone' locations are encapsulated with a customised window, designed to be the exact length of each 'set accented tone'. Subsequently, chroma information is computed at the 'set accented tone' locations. Chroma represents the prominence of each of the twelve musical notes of the scale for a given audio frame. The results of calculating chroma are therefore contained in a 12-element chroma vector. The resulting chroma vectors are compared to determine which parts are similar. Results show that structurally segmenting music using chroma instead of individual notes at 'set accented tone' locations significantly increases the accuracy of the resulting structural segmentation. The structural segmentation approach proposed in this paper is outlined in the block diagram in Figure 2.

4.1. Set Accented Tone Identification

Firstly, to determine the locations of the 'set ac-

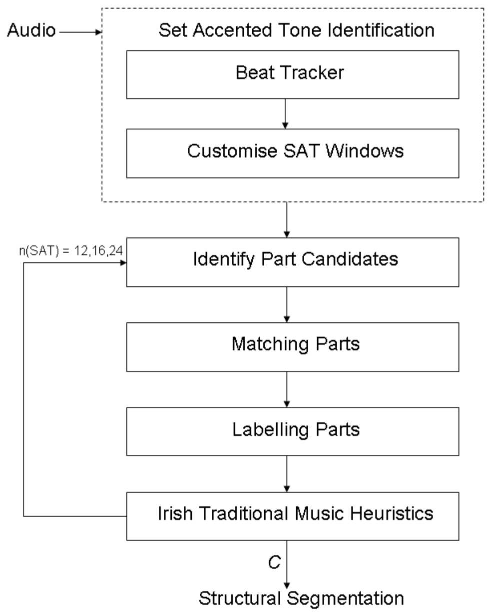

Fig. 2: *A block diagram of the proposed structural segmentation approach where 'C' is equal to the confidence measure associated with a structural segmentation as detailed in Section 4.5. The number of 'set accented tones' per part is represented by 'n(SAT)' as detailed in Section 4.3.*

cented tones' within the music, a beat tracker is applied to the audio signal. For most time signatures within Irish Traditional Music the 'set accented tones' are located on the beats of the music. A notable exception to this are tunes set to a time signature of 4/4. For a tune set in 4/4 there are 4 beats per bar. However it is shown in [2] that despite this fact there only exists two 'set accented tones' per bar. Consequently, for the case of tunes set in 4/4, only the alternating 1^{st} and 3^{rd} beats of the bar are considered during the 'set accented tone' extraction phase. The beat tracker used to determine the beat locations is the same beat tracker used in [3], the details of which are provided in [10]. The beat tracker provides the locations of the beats of the music, also provided is an onset detection function which indicates the locations of all the notes of the music.

Following this, each 'set accented tone' is encapsulated by a window at the beat locations. A customised window size is determined for each 'set accented tone'. The window size is equal to the distance between a detected beat location and the next detected onset after that beat location. The windows used are visualised in Figure 3. This maximises the available harmonic information when determining chroma values at each 'set accented tone' location.

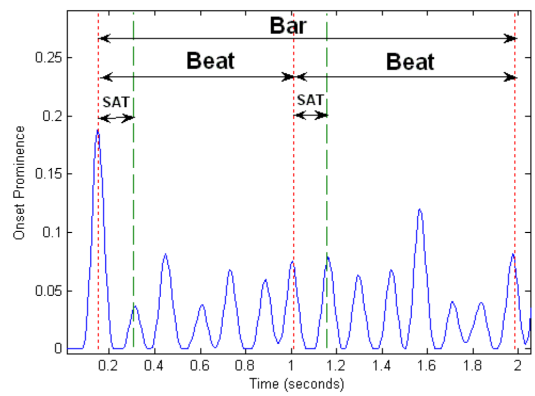

Fig. 3: *A graph showing the onsets detected for the first bar of an Irish traditional jig in 6/8 time. The dotted line corresponds to the beat locations; the dashed line corresponds to the next detected onset after a beat location. Each SAT denotes the window which encapsulates a 'set accented tone'.*

4.2. Chroma Vector Calculation

Following the calculation of the customised window for each 'set accented tone' a chroma vector is calculated for each of these windows. The chromagram is defined in [11] as the restructuring of a spectral representation in which the frequencies are mapped onto a limited set of 12 chroma values in a many-to-one fashion. This is done by assigning frequencies to the 'bin' that represents the ideal chroma value of the equally tempered scale for that frequency. The 'bins' correspond to the twelve notes in an octave. This means for instance, that every occurrence of the note *C* regardless of octave will be represented by the same value. So, *C2*, *C3* and *C4* will all be represented simply as *C*.

To calculate the chroma for each 'set accented tone', the customized 'set accented tone' window is treated as the signal input to a Short Time Fourier Transform (*STFT*) with a frame length of 12 milliseconds. The peaks contained within the resulting $STFT$ frames represent the prominent frequencies present at the 'set accented tone' locations. A peak finding algorithm is applied to each of the resulting $STFT$ frames, where the value at each peak is stored in one of 12 chroma bins depending on the value of the frequency at the peak locations. A chroma vector is calculated for each 'set accented tone' within the tune.

4.3. Identifying Parts

Following the calculation of the chroma vectors, a certain heuristic of Irish Traditional Music is utilised. According to Table 1 there are a finite number of time signatures present within Irish Traditional Music which leads to the hypothesis that there can only be 12, 16 or 24 'set accented tones' per part in an Irish Traditional Tune. Using this heuristic, the algorithm iterates three times dividing the computed 'set accented tones' into 'part candidates' of length 12, 16 and 24 'set accented tones' per 'part candidate' each time. Following this, the chroma vectors of the 'part candidates' are compared to determine which of them are to be considered equivalent. An example of the chroma vectors at 'set accented tone' locations that have been computed for two equivalent 'part candidates' is illustrated in Figure 4.

4.4. Matching Parts

To determine which parts are similar, the chroma vectors of parts are compared to the chroma vectors of other parts. Three distance measures are evaluated separately when calculating similarity values between chroma vectors, the results of which are detailed individually in Section 5. The results of the three distance measures are also combined to provide the best possible structural segmentation for each Irish Traditional tune. For each tune, the distance measure which resulted in the highest confidence measure is chosen as the most likely structural segmentation.

The distance measure values are all normalised so that the resulting values lie between 0 and 1, where 0 is equal to a perfect match. Thus, the measures should be considered as dis-similarity measures. The three dis-similarity measures used are as follows:

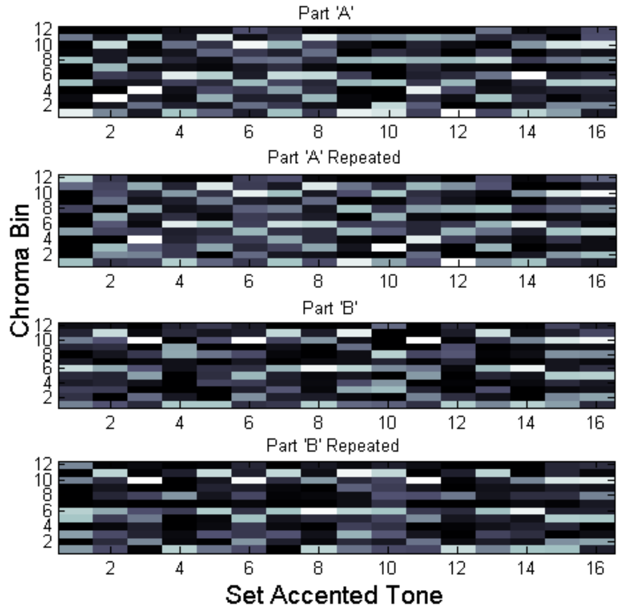

Fig. 4: *An example of the chroma extracted at 'set accented tone' locations for two equivalent 'A' parts and two equivalent 'B' parts of an Irish reel in 4/4 time. There are 16 'set accented tones' per part in this example. Black indicates a low presence of a particular pitch class at a 'set accented tone' location and white indicates a high presence of a particular pitch class.*

The Euclidean Distance Measure as described by Equation 1, Cross-correlation as described by Equation 2 and the mean of the absolute difference between two vectors as described by Equation 3.

$$D(v_1, v_2) = \sqrt{(}\sum_i (v_1(i) - v_2(i))^2) \quad (1)$$

where v_1 and v_2 are the two chroma vectors being compared.

$$D(v_1, v_2)(j) = \begin{cases} \sum_{n=0}^{N-m-1} v1_n + mv2_n^* & (m \geq 0) \\ D^*(v_1, v_2)(-m) & (m < 0) \end{cases} \quad (2)$$

where $v2^*$ is the complex conjugate of v_2.

$$D(v_1, v_2) = \sum_i |v_1(i) - v_2(i)| \quad (3)$$

A dis-similarity value is calculated for each part comparison according to Equation 4.

$$DS(part) = \frac{\sum_{n=0}^{N-1} D(v_1(n), v_2(n))}{N} \quad (4)$$

where N is equal to the number of 'set accented tones' per part.

If this dis-similarity is below a threshold T, the parts are considered to be equivalent. An adaptive threshold is used and is calculated to be the average of all the part dis-similarity values of the entire tune. The threshold T is described by Equation 5.

$$T = \frac{\sum_i D(v_1, v_2)}{i} \quad (5)$$

4.5. Labelling Parts

Labeling is carried out concurrently with the part matching. The first part is always labeled 'A'. The part matching is then carried out, any parts considered to be similar to this first 'A' part are also labeled 'A'. Once all parts have been compared against the first part, the algorithm backtracks to the earliest part which has not yet been assigned a label. This part is labeled 'B'. This 'B' part is then checked for similarity against all unlabelled parts. Following this, parts that are considered similar to the 'B' part are also labeled 'B'. This process of comparing and labeling unlabeled parts repeats until all parts have been assigned a label. Once the matching and labelling process has been completed, the resulting segmentation results are validated to determine if they are plausible within the heuristics of Irish Traditional Music. Certain rules are applied to ensure that the structural segmentation complies with these heuristics.

- A label may not occur three or more times consecutively.
- A label must appear within a distance of p*2 of another instance of that label, where p is equal to the number of unique labels within the tune.
- The amount of parts must be an even number.
- The last label must be the highest alphabetically.

If all of these criteria are met, the segmentation result can be considered plausible within the constraints of Irish Traditional Music.

Following this, an overall structural segmentation confidence is computed by summing the part dis-similarity measures for each part comparison and calculating the reciprocal of the resulting value as described by Equation 6. After the algorithm has attempted to segment the audio using 12, 16 and 24 'set accented tones' per part, the resulting segmentation with the highest confidence measure is chosen as the most likely structural segmentation.

$$C = 1/\frac{\sum_{n=0}^{N-1} DS(n)}{N} \quad (6)$$

where N is equal to the number of part comparisons carried out during the structural segmentation.

5. RESULTS

The proposed approach was tested on a database of 44 Irish Traditional tunes which consist of a total of 266 segments. The segment labels of the tunes in the database have been annotated by an Irish Traditional Musician. The three distance measures described in Section 4.4 are evaluated separately. Following this, the results of the three separate distance measures for each tune are combined, utilising the confidence values associated with each structural segmentation. The distance measure which resulted in the highest confidence value is chosen as the most likely structural segmentation for that tune. The labelling results of the structural segmentation for each tune using the Euclidean Distance measure, correlation and the mean of the difference between two chroma vectors are detailed in 2. The results of combining the three distance measures to produce the best possible structural segmentation for each tune are also detailed in Table 2. Accuracy in Table 2 is calculated using Equation 7.

$$Acc = \frac{N - FP - FN}{N} \quad (7)$$

where *'N'* is equal to the number of annotated structural segments.

6. CONCLUSIONS

The previous approach detailed in [3] relies on two pre-processing stages, beat tracking and pitch detection, before attempting a structural segmentation.

Table 2: *Results of the structural segmentation algorithm. Accuracy is calculated according to Equation 7*

Distance Measure	Total Parts	Detected Parts	GP	FP	FN	Accuracy
Euclidean	266	223	199	24	43	75%
Correlation	266	222	187	35	37	73%
MeanDiff	266	221	174	47	55	62%
Combined	266	266	229	40	8	82%

The approach presented in this paper attempts to eliminate the limitations that pitch detection contributes to the results presented in [3]. Total accuracy for labelling parts in [3] is calculated to be 50%. Computing chroma vectors at 'set accented tone' locations instead of detecting pitches reduces the errors incumbent in determining one single value at each 'set accented tone' location. As can be seen in Table 2, the accuracy of segment labelling using chroma rather than pitch detection at 'set accented tone' locations is calculated to be 82%. This is an increase in accuracy of 32% over the structural segmentation results presented in [3] proving that using chroma information instead of pitch information at 'set accented tone' locations significantly increases accuracy. In addition, the accuracy of results using a single distance measure are noticably lower than when the results of the structural segmentation of each distance measure are combined. However, it was only the improvement of the segment labelling which was attempted here, the segment time results remain the same as in [3]. These segment time results rely on the accuracy of the beat tracker.

Currently, this approach has only been tested for monophonic musical signals, future work will attempt to apply this structural segmentation approach to polyphonic music. Clearly, attempting to attribute a single pitch value at a 'set accented tone' location for polyphonic music is not feasible. Future work will aim to show that using chroma at 'set accented tone' locations for polyphonic music will overcome this problem.

Furthermore, there are a number of very popular musical modes used throughout Irish Traditional Music. Future work will involve a study of the occurrences of these modes within this musical genre, this modal statistical information could then be used to create a chroma filter. The chroma filter could weight each chroma vector according to the popularity of the musical modes within Irish Traditional Music to produce a more accurate structural segmentation for this genre.

7. ACKNOWLEDGMENTS

This work was carried out as part of the IMAAS project funded by Enterprise Ireland under the Commercialisation fund.

8. REFERENCES

[1] Mark A. Bartsch and Gregory H. Wakefield, "To catch a chorus: using chroma-based representations for audio thumbnailing.," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, 2001, pp. 15–18.

[2] Míchéal Ó'Súilleabháin, *Innovation and Tradition in the Music of Tommie Potts*, Ph.d, Queen's University, 1987.

[3] Cillian Kelly, Mikel Gainza, David Dorran, and Eugene Coyle, "Structural segmentation of music using set accented tones," in *124th Audio Engineering Society Convention*, Amsterdam, The Netherlands, 2008.

[4] Bee Suan Ong and Perfecto Herrera, "Semantic segmentation of music audio contents," in *International Computer Music Conference*, Barcelona, Spain, 2005.

[5] Roger B. Dannenberg and Ning Hu, *Discovering Musical Structure in Audio Recordings*, vol. 2445 of *Lecture Notes in Computer Science*, Springer Berlin, 2002.

[6] Simon Dixon, “A lightweight multi-agent musical beat tracking system,” in *Pacific Rim International Conference on Artificial Intelligence*, Melbourne, Australia, 2000, pp. 778–788.

[7] Jouni Paulus and Anssi Klapuri, “Music structure analysis using a probabilistic fitness measure and an integrated musicological model,” in *International Conference on Music Information Retrieval*, Philadelphia, PA, USA, 2008.

[8] Fang Zheng, Guoliang Zhang, and Zhanjiang Song, “Comparison of different implementations of mfcc,” *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582 – 589, 2001.

[9] Dan Ellis, “Identifying ‘cover songs’ with beat-synchronous chroma features,” *MIREX 2006 Audio Cover Song Contest*, 2006.

[10] Mikel Gainza, Dan Barry, and Eugene Coyle, “Dynamic audio beat tracking,” Technical Report TR020208, 2008.

[11] Steffen Pauws, “Musical key extraction from audio,” in *ISMIR*, Barcelona, Spain, 2004.

[12] Kristoffer Jensen, “Multiple scale music segmentation using rhythm, timbre, and harmony,” *EURASIP Journal on Advances in Signal Processing*, 2007.