2010-01-01

Audio Thumbnail Generation of Irish Traditional Music

Cillian Kelly
Dublin Institute of Technology, cillian.kelly@dit.ie

Mikel Gainza
Dublin Institute of Technology, mikel.gainza@dit.ie

David Dorran
Dublin Institute of Technology, david.dorran@dit.ie

Eugene Coyle
Dublin Institute of Technology, Eugene.Coyle@dit.ie

Audio Thumbnail Generation of Irish Traditional Music

Cillian Kelly, Mikel Gainza, David Dorran and Eugene Coyle

Audio Research Group
Dublin Institute of Technology

E-mail: cillian.kelly@dit.ie, mikel.gainza@dit.ie, david.dorran@dit.ie, eugene.coyle@dit.ie

Abstract — **An approach is presented which generates an audio thumbnail of Irish Traditional music. An audio thumbnail is considered to be the most representative segment of the music. For popular music, the chorus is considered to be an ideal audio thumbnail, however in Irish Traditional music there is no chorus. An Irish Traditional tune consists of two or more short structural segments called parts. Parts are repeated to extend the tune, and the tune itself is also repeated once or more in its entirety. To further extend a performance, tunes are concatenated to form a set of tunes. As a result, there is plenty of repetition within this music type. The presented approach utilises an existing approach which calculates the structure of Irish Traditional Music. The structural information is used to extract a single rendition of each distinctive part. The resulting parts are concatenated to form the audio thumbnail.**

Keywords — **Audio Thumbnail, Chroma, Irish Traditional Music**

I Introduction

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. However, Irish Traditional music does not contain a chorus, nor is there a segment which is repeated more often than others. Therefore an approach is presented which generates an audio thumbnail for Irish Traditional music by extracting and concatenating specific segments of the music.

Audio thumbnails can be used as short previews of a song provided to users by online music stores such as iTunes. These online music stores such as iTunes provide a user with a free preview of each song, typically 30 seconds in length. Currently the preview provided to the user is either the first 30 seconds or a random 30 seconds of the song. Providing an audio thumbnail would allow the user to listen to the most representative segment of the music. Thus, the user would be more informed about the content of the song before deciding to purchase it or not. This is particularly pertinent for Irish Traditional music. In this ancient genre many tunes have been passed down aurally over generations. As such, the title of a tune may have been altered erroneously or indeed lost altogether over time. This leads to many tunes on commercial recordings with incorrect titles or being given the title *'gan ainm'* (an Irish phrase meaning *'without name'*). Therefore, unlike popular music, simply looking at the title of an Irish Traditional tune may reveal no useful information. Furthermore, as will be explained in Section III, an audio track comprised of Irish Traditional music may contain a number of individual tunes. Consider a user who may only be interested in one particular tune within the music, if a random 30 seconds is provided to the user as a preview of the music, this 30 seconds may not contain any bars of the tune the user is interested in. An approach is presented here which generates an appropriate audio thumbnail for Irish Traditional music.

This paper is structured as follows: Section II provides a literature review of existing approaches toward generating an audio thumbnail. Certain details of these approaches which make them un-

suitable for use with Irish Traditional Music are highlighted. In Section III the structure of Irish Traditional music is detailed to increase understanding of the proposed approach. Section IV details the proposed approach for generating an audio thumbnail of Irish Traditional music. Finally, in Section V the results and conclusions of the proposed approach are provided.

II Literature Review

Previous approaches toward providing an audio thumbnail of music have been concerned with the popular music genre. For this genre, the chorus is considered to be the ideal candidate for the audio thumbnail. As such, chorus detection and audio thumbnailling are synonymous in the popular music domain.

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, [1] segments the audio signal in to frames. Following this, a beat-tracking system is employed to dynamically provide a frame segmentation which is specific to each song. Chroma information is then extracted for each frame. A self-similarity matrix is then computed 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. The chorus is identified by locating the maximum element of the similarity matrix. The thumbnail is then defined by the time-position of this maximum and has a length equal to the length of the window that was used for the correlation filtering.

In [2] an approach is presented which attempts to provide an audio thumbnail of songs by The Beatles. The musical signal is divided into overlapping frames. A pitch estimation algorithm is applied to each resulting frame. Despite the test data being polyphonic, it is said in [2] that estimating a single pitch still captures much information when the signal contains a prominent leading vocal. The resulting feature vector sequence is grouped into overlapping segments. Following this, each segment is matched against the feature vector sequence using a dynamic programming approach to reveal the repetitive structure. The chorus is considered to be the most repeated segment. The approach detailed in [2] is extended in [3] where the audio feature used is key instead of pitch. A local key detection algorithm is employed and distances between audio frames are computed based on the resulting keys rather than single pitches.

An approach toward generating an audio thumbnail is presented in [4]. The approach in [4] is similar to that in [2], however in [4] the audio feature that is used is timbre. This audio feature is extracted by computing a constant-Q spectrum for each audio frame. The resulting spectra are normalised and subjected to Principal Component Analysis which yields 21-dimensional feature vectors for each frame. A Hidden Markov Model (*HMM*) is then trained using the entire sequence of feature vectors for the given piece of music. The features are then Viterbi-decoded using the trained *HMM*. The resulting state sequences gives the most likely sequence of assignments for each beat of the music to one of 40 possible timbres. Once each frame has been classified as a particular timbre, the frames are clustered together according to this timbre resulting in a structural segmentation of the music. To generate the audio thumbnail of the music, the most frequently occurring segment-type is determined. The second occurrence of this segment-type is labelled as the audio thumbnail. It is said in [4] that this is because a musical segment towards the middle of the track is often more representative of the musical piece than a segment which occurs at the very beginning.

In [5] an approach is presented which segments Irish Traditional tunes into their constituent parts and provides a semantic labelling for the resulting parts. Pitch values are determined at specific locations within the music using a pitch detector. This results in a selective pitch contour. Melodic patterns are searched for amongst this pitch contour to determine the overall structure of the music. This approach was tested on a database of monophonic pieces of Irish Traditional music. The approach presented in [5] is extended further in [6] where chroma is calculated at 'set accented tone' locations rather than single pitch values. Following this, the chroma vectors are grouped according to heuristics specific to Irish Traditional Music. The resulting groups of chroma vectors correspond to the structural segments of the music and are compared using three different distance measures to determine which of the segments are similar. Extracting chroma rather than single pitch values at 'set accented tone' locations allows the approach in [6] to be applied to polyphonic music rather than only monophonic music as in [5].

In [1, 2, 3, 4] the most repeated segments are considered to be ideal candidates for an audio thumbnail. These approaches cannot be used to generate an audio thumbnail of Irish Traditional Music. In this music type, there is no structural segment which repeats more often than any other (see Section III). Also, in [2] the audio feature chosen to characterise audio frames is pitch. It is explained in [2] that only music with a prominent vocal can be characterised effectively in this way. Irish Traditional music contains no vocal, therefore pitch is not a suitable audio feature to use

when characterising the audio frames. Similarly in [3] and [4], the audio features used are key and timbre respectively. Within a piece of Irish Traditional music there may be no key changes or timbre changes. Therefore, discerning differences between audio frames characterised by these audio features is not feasible. The approach presented in Section IV is an extension of the approach presented in [7]. The structural information which results from the approach presented in [7] is utilised to generate an audio thumbnail of Irish Traditional Music using chroma at certain locations within the music as the audio feature.

III Irish Traditional Music

Irish Traditional tunes are divided into segments which are referred to as *parts* and are denoted by upper case letters. There are various tune types which are played as part of Irish Traditional Music, the most popular of which are the reel and the jig [8].

A tune may consist of two or more parts. Parts are repeated to extend a performance of the given tune and the tune itself is also repeated in its entirety to extend the performance even further. For example, a three-part tune consisting of an 'A' part, a 'B' part and a 'C' part may be performed according to the musical structure of *AABBCCAABBCC*. Thus, for this example, a three-part tune becomes a piece of music which actually consists of twelve distinct parts. Furthermore, tunes may be concatenated and played one after another in a single performance in what is called a *set*. An example of the structure of a set comprised of a three-part tune followed by a two-part tune is *AABBCCAABBCC-AABBAABB*. For a user who simply wants to preview the music, the ideal audio thumbnail would consist of one rendition of each part. Accordingly, an ideal thumbnail of this example would be represented by the structure *ABABC*. This would provide the user with a reduced representation of the music and it would also provide the user with an example of each individual part of the music. The approach presented here generates an audio thumbnail which consists of a single rendition of each part within the music.

To generate this reduced representation audio thumbnail the music must first be segmented into its constituent parts. This is achieved by extracting audio features at certain locations within the music and comparing parts to determine similarities based on the resulting audio features. While two renditions of the same part may be notated identically, they are rarely performed identically. This is due to the large presence of musical variation inherent with this music type. Embellishments introduced by a musician are encouraged and will render two parts notated as identical as being aurally different. In Irish Traditional music there exist a certain set of notes which are considered impervious to musical variation. These notes are known as 'set accented tones' and are located on the beat of the music [9] as can be seen in Figure 1. These notes are used to characterise each part of the music and are compared to create the part similarity matrix used in Section IV.

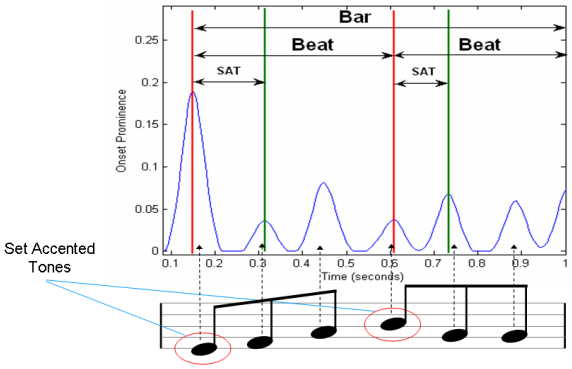

Fig. 1: An onset detection function of one bar of Irish Traditional music. Each 'set accented tone' is located between the start of the beat and the next detected onset. For each 'set accented tone' a window is created between these two points of the onset detection function. Chroma is calculated for each resulting window.

IV Proposed Approach

a) *Overview*

An approach toward generating an audio thumbnail of Irish Traditional tunes is detailed in this section. The approach presented here utilises the approach toward structurally segmenting Irish Traditional music detailed in [7] where the locations and semantic labels of each part of an Irish Traditional music piece are identified. The locations and semantic labels of each part comprise the structural information of the music. The block diagram in Figure 2 outlines the proposed approach. To generate an audio thumbnail consisting of specific parts of the music it is necessary to segment music into its constituent parts. Appropriate parts are then selected and concatenated to form the audio thumbnail. In [7] it is said that musical variation is a prominent characteristic of this genre but that there are a certain set of notes known as 'set accented tones' which are considered impervious to musical variation. The 'set accented tones' are located on the beat of the music. Using a beat tracker, the 'set accented tones' are located and chroma information is extracted at these locations within the music. The resulting chroma vectors are grouped to represent the parts of the music. The parts are then compared with one another in [7] to form a part similarity matrix. Unit kernels which represent the possible structures of an Irish

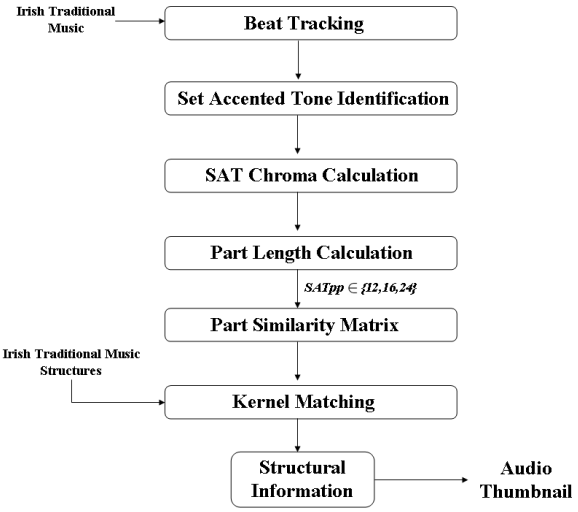

Fig. 2: A block diagram of the proposed approach.

Traditional tune are matched with the part similarity matrix to determine the overall structure of the music. Following this an audio thumbnail is generated by concatenating a single rendition of each part of the music.

b) Beat Tracking and Set Accented Tone Identification

As detailed in [7], to extract chroma at 'set accented tone' locations within the music, the locations of the 'set accented tones' must be defined. Within Irish Traditional Music these notes are considered to be the first note of each beat. Therefore a beat tracker [10] is employed to determine the location of each beat within the music. The beat tracker provides the location of each beat of the music along with an onset detection function which provides the locations of each note within the music. To encapsulate each 'set accented tone' a window is created extending from the start of each beat to the next detected onset as illustrated in Figure 1. This maximises the available harmonic information when determining chroma values at each 'set accented tone' location. Following the creation of each 'set accented tone' window, chroma information is extracted at each of these locations.

c) Chroma Calculation

To create the part similarity matrix detailed in Section e) chroma must be calculated at each 'set accented tone' location. Chroma is a spectral representation of music in which frequencies are mapped onto a set of 12 chroma values which correspond to the 12 notes of the equal tempered scale [11]. To extract chroma for each 'set accented tone', in [7] a Short Time Fourier Transform is applied to each 'set accented tone' window. The local maxima contained within each of the resulting STFT frames are identified using a peak picking algorithm. Following this, the magnitudes of each frequency at each resulting peak location are added to the appropriate chroma bin according to the note of the musical scale to which the frequency most closely corresponds. This results in a chroma vector of twelve elements each containing the amount of each note which was present in the given 'set accented tone' window.

d) Part Length Calculation

Following chroma calculation at each 'set accented tone' location, according to [7] it is necessary to determine how many 'set accented tones' per part there are in the piece of music. This is required in order to determine the correct groups of chroma vectors to use when creating the part similarity matrix in Section e). According to the Irish Traditional Music heuristics detailed in [5] there can only be 12, 16 or 24 'set accented tones' per part (*SATpp*) in an Irish Traditional tune. In [7] determining the part length is achieved by testing each of these three values and calculating a confidence score for each value. The confidence score is based on the similarity between the chroma vectors representing each part. The *SATpp* value yielding the highest confidence score is the *SATpp* value used in [7] when creating the part similarity matrix.

e) Part Similarity Matrix

As detailed in Section d), once the *SATPP* value has been determined, the part similarity matrix is created according to this *SATpp* value. The part similarity values (calculated when determining the confidence score in Section d)) associated with this particular *SATpp* value are used to create the part similarity matrix. These values indicate the similarity of each part of length *SATpp* with every other part of length *SATpp* within the music.

In [7] these values are placed into a matrix which results in a part similarity matrix of size P by P where P is equal to the total number of parts within the music. An example of a part similarity matrix is shown in Figure 4. The part similarity matrix is used along with unit kernels to determine the structure of the music as described in Section f).

f) Kernel Matching

The following section details how unit kernels in [7] are matched with the part similarity matrix created in Section e). Firstly, the kernels that are used for matching are described, along with justifications for using these particular kernels. The process of how the unit kernels are matched with the part similarity matrix is then detailed.

Kernel matching in [7] relies on the availability

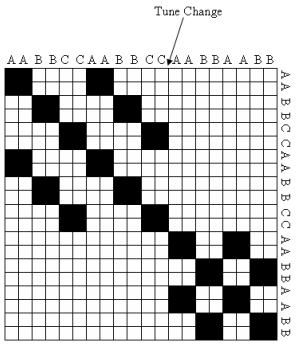

Fig. 3: A unit kernel representing the structure of two tunes. The first tune represented has a structure of *AABBCCAABBCC* the second tune represented has a structure of *AABBAABB*. Black represents similar parts and white represents dis-similar parts.

of pre-existing unit kernels which each represent a specific musical structure. A unit kernel is a matrix consisting of ones and zeros which represent the pattern of a musical structure. A total of 600 unit kernels are used in [7] which represent the musical structures which are most common within Irish Traditional Music. An example of a two-tune kernel is shown in Figure 3.

The unit kernels are correlated with sections of the part similarity matrix as illustrated in Figure 4. In [7] the kernel which yields the highest matching value is the kernel which represents the most likely musical structure present within the given section of the part similarity matrix. The following steps describe the kernel matching technique used in [7]:

1. At location *(i,i)* of the part similarity matrix, each K *x* K unit kernel is matched with a K *x* K section of the part similarity matrix using inner matrix multiplication. For the first iteration only, $i = 1$.

2. The kernel which results in the highest match value is deemed to represent the structure of that section of the part similarity matrix.

3. The value i is updated to be $(i + K)$ where K is equal to the length of the kernel in step 2.

4. Steps 1-3 are repeated until the entire part similarity matrix has been processed.

The outcome of this kernel matching as outlined in Figure 2 is the structural information of the music which is comprised of the location of each part within the music and a semantic label for each part. This structural information is used to generate the audio thumbnail in Section g).

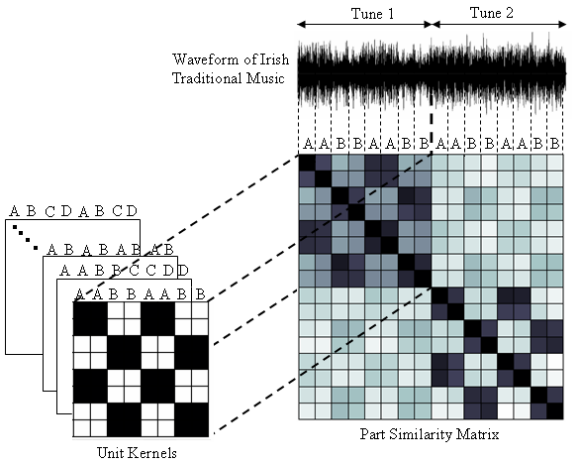

Fig. 4: The section of a part similarity matrix with which a unit kernel is correlated. Each cell of the part similarity matrix corresponds to a part of an Irish Traditional tune. This section of the part similarity matrix is correlated with each unit kernel. The unit kernel which results in the highest matching value corresponds to the structure of this section of the part similarity matrix.

g) Audio Thumbnail Generation

Utilising the approach detailed in [7], the locations and semantic labels of each part within the music are now available. This information is used to extract specific parts within the music which are concatenated to produce the audio thumbnail as shown in Figure 5.

The locations of the start and end of each part along with the semantic labels are used to identify each required part of the music. The first rendition of each part is extracted as this is generally the rendition with the lowest presence of melodic variation. As such, the first rendition of a part will be the most recognisable and identifiable to the user. For the example shown in Figure 5 the audio has been reduced from a track consisting of 20 parts to an audio thumbnail consisting of 5 parts. The resulting audio thumbnail is a reduced representation of the music which contains an example of each distinctive part within the music. This audio thumbnail provides a user with an ideal example of the parts which comprise the music.

V Results & Conclusions

The evaluation of this approach was carried out on a hand annotated database of 30 tracks containing Irish Traditional music. Each music track was comprised of either one, two or three separate tunes. The database used is the same database that was used to test the approach presented in [7].

The generation of the audio thumbnails relies completely on the accuracy of the structural information resulting from the kernel correlation. Therefore, if parts were labelled incorrectly for a

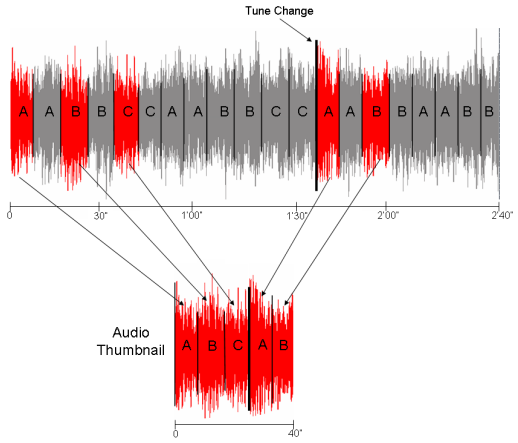

Fig. 5: An audio thumbnail is generated by concatenating a single rendition of each part present within the music. In this example, a three part tune is followed by a two part tune. The original audio track is comprised of 20 separate parts, the audio thumbnail consists of only 5 parts but provides an example of each distinctive part within the music.

particular track, the desired audio thumbnail could not be generated.

The structural information of 17 out of the 30 tracks in the database was calculated to be completely correct. Therefore it was possible to generate the desired audio thumbnail for 57% of the tracks in the test database. The part location times used to generate the audio thumbnails had an accuracy of 90% within a tolerance of 1 second. The approach presented here uses the same method to identify part times as detailed in [5].

For the remaining 13 tracks of the database the calculated part labelling was only partially correct. Therefore the resulting audio thumbnails which were generated based on these part labellings contained some desired parts but also contained some undesired parts. When all 30 audio tracks were evaluated, 86% of the parts which were considered to comprise the ideal audio thumbnails were contained within the automatically generated audio thumbnails.

In conclusion, the presented approach toward generating an audio thumbnail of Irish Traditional music relies on the accuracy of the structural information provided by the approach detailed in [7]. Despite this fact, the structural information is sufficiently accurate for 57% of audio thumbnails to be calculated correctly. When the constituent parts of the thumbnails are evaluated separately, 86% of the parts that constitute the audio thumbnails are considered to be correct.

The concatenation of parts of an audio track can lead to a click or a pop at part transition locations. Currently, no attempt is made to ensure a smooth transition from one part of an audio thumbnail to the next. Future work will aim to ensure smooth transitions between each part contained within the audio thumbnail.

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.

[2] Wei Chai and Barry Vercoe. Structural analysis of musical signals for indexing and thumbnailing. In *ACM/IEEE Joint Conference on Digital Libraries*, Houston, TX, U.S.A., 2003.

[3] Wei Chai. Semantic segmentation and summarization of music: methods based on tonality and recurrent structure. *Signal Processing Magazine, IEEE*, 2006.

[4] Mark Levy, Mark Sandler, and Michael Casey. Extraction of high-level musical structure from audio data and its application to thumbnail generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, 2006.

[5] 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.

[6] Cillian Kelly, Mikel Gainza, David Dorran, and Eugene Coyle. Structural segmentation of irish traditional music using chroma at set accented tone locations. In *127th Audio Engineering Society Convention*, New York, New York, U.S.A., 2009.

[7] Cillian Kelly, Mikel Gainza, David Dorran, and Eugene Coyle. Locating tune changes and providing a semantic labelling of sets of irish traditional tunes. In *Submitted to ISMIR*, Utrecht, The Netherlands, 2010.

[8] Harry Long. *The Waltons Guide to Irish Music*. Waltons Publishing, 2005.

[9] Mícháel Ó'Súilleabháin. *Innovation and Tradition in the Music of Tommie Potts*. PhD thesis, Queen's University, 1987.

[10] Mikel Gainza. On the use of a dynamic hybrid tempo detection model for beat tracking. In *Submitted to IEEE International Conference on Multimedia and Expo*, Singapore, 2010.

[11] Steffen Pauws. Musical key extraction from audio. In *International Conference on Music Information Retrievals*, Barcelona, Spain, 2004.