COMPUTING STRUCTURAL DESCRIPTIONS OF MUSIC THROUGH THE IDENTIFICATION OF REPRESENTATIVE EXCERPTS FROM AUDIO FILES

BEE-SUAN ONG[1], PERFECTO HERRERA[2]

[1] Institut Universitari de l'Audiovisual, Universitat Pompeu Fabra, SPAIN
beesuan@iua.upf.es
[2] Institut Universitari de l'Audiovisual, Universitat Pompeu Fabra, SPAIN
pherrera@iua.upf.es

With the rapid growth of audio databases, many music retrieval applications have employed metadata descriptions to facilitate better handling of huge databases. Music structure creates the uniqueness identity for each music piece. Therefore, structural description is capable of providing a powerful way of interacting with audio content, and serves as a linkage between low-level description and higher-level descriptions of audio (e.g. audio summarization, audio fingerprinting, etc.). Identification of representative musical excerpts is the primary step towards the goal of generating structural descriptions of audio signals. In this paper, we will provide a systematic review of the existing work on extracting musical structure descriptors from music files and will present, discuss, and evaluate various approaches in identifying representative musical excerpts of music audio signals.

INTRODUCTION

With the recent explosion in the quantity of digital audio library and databases, content descriptions of standard metadata, such as MPEG-7, play an important role in efficiently managing and retrieving of those audio files. It is believed that music structural descriptions, which subsume temporal, harmonic, rhythmic, melodic, polyphonic, motivic and textual information, may improve efficiency and effectiveness in handling huge music audio databases, as repetition and transformations of music structure have created a uniqueness identity of music itself. Moreover, such structural description can also provide a better quality access and powerful ways of interacting with audio content, such as better quality audio browsing, audio summarizing, audio retrieving, audio fingerprinting etc., which would be very much useful and applicable for music commercial and movie industries.

Humans assimilate information at a semantic level with remarkable ease. Nevertheless, the limitation of human memory makes them impossible to recall every single detail of all incidents that happen in their daily life. As a human, we may only recall certain events, which have created a "strong" impression in our mind. Same with the aspect of audio music, we do not recall the music that we hear in its entirely but through a small number of distinctive excerpts (e.g. chorus, verse, intro, etc.) that have left an impression on our mind. It is usually the case that we only need to listen to one of those distinctive excerpts in order to recall the title for the musical piece, or, at least, to tell if we have heard this song before. Hence, identifying representative musical excerpts of audio signals would be a primary step going towards the goal of generating music structure metadata.

One may ask, "What is the criteria required for a music section to be acknowledged as a representative musical excerpt of an audio music?". As mentioned earlier, the repetition and transformation of music structures has created the uniqueness of the music itself, hence we assume that the most representative sections of music are frequently repeated within a song. Therefore, in finding representative excerpts of audio signal, detection of the repetition patterns is necessary. So far, there has been some research in identifying representative musical excerpts of audio signal in the area of music content analysis research [1],[2],[3]. Several methods have also been proposed to obtain reliable representative excerpts from audio signal. In this paper, we provide a critical literature review and evaluate the current methods aimed to deal with these issues. A general framework of the existing approaches will be advanced. This framework forms the guiding principle in identifying musical excerpts of audio signals.

By reviewing the current proposed techniques in identifying representative musical excerpts of audio signals, it is clear that all the proposed methods share a similar general framework, which consists of three main processes: feature extraction, signal segmentation and

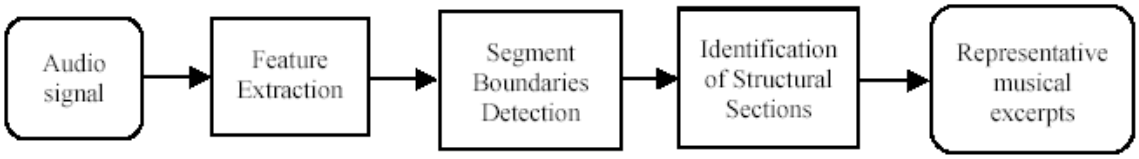

Figure 1: Automatic identification of representative musical excerpts from audio files

identification of structural sections (see flowchart in figure 1). The paper is organized according to these three processes. In section 1 we present the different extracted feature attributes by grouping them according to the features' similarities and differences. In addition, we discuss two different feature extraction approaches. Section 2 reviews the existing processes for achieving temporal separation of section candidates. Section 3 reviews the current techniques aimed to discovering those excerpts that could be considered as representative of a music title.

1 FEATURE EXTRACTION

As mentioned earlier, finding repetitive patterns in music is a key issue in this research. Hence, extracting a kind of music representation from the audio signal is crucial in finding those repetitions in music. Extracting symbolic score-like representation in music could be a possible way to complete the task. However, due to the demanding constraints in extracting symbolic score-like representation in polyphonic music, this approach is practically infeasible. Instead, extracting low-level representations of audio signal for musical content description is found to be an alternative way for completing this task. Low-level feature attributes, which describe the musical content of sound signal, have been widely used in audio identification, boundary detection, classification, sound and music retrieval, etc.

In identifying representative musical excerpts of sound signals, proper selection of feature attributes is crucial to obtain appropriate musical content description for a later identification process. Nevertheless, effective description of musical content not only depends on the best feature attributes, but sometimes also on the use of different feature attributes in a combined manner. Therefore, the application of musical knowledge into the selection process would further improve the quality of musical content description.

Current feature attributes used in representative musical excerpts identification are characteristically computed on frame basis in order to obtain the short-term descriptions of the sound signal. The music signal is cut into frames. For each of these frames, a feature vector of low-level descriptors is computed. In accordance with the similarities and differences of the generated content descriptions, these feature attributes can be roughly classified into three groups: timbre-related features, melody-related features, and dynamics-related features. Figure 2. illustrates the overall structure of feature extraction.

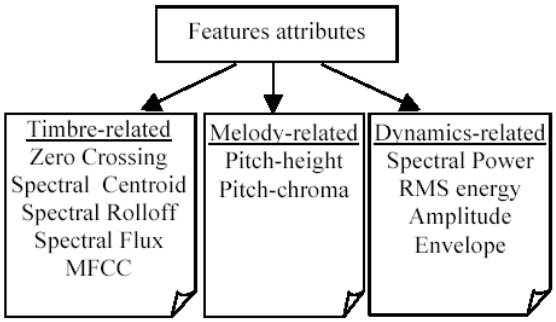

Figure 2: Illustration of categories of feature attributes.

1.1 Timbre-related features

Timbre content descriptions are of general importance in describing audio. Most of the existing research work uses timbre content descriptions in order to differentiate music and speech besides music classification applications. Hence, lots of timbre-related features have been proposed in this research. In fact, timbre-related features are the most widely used among the three groups mentioned above. So far, the most employed timbre-related features are:

- Zero Crossings [4]: A measure of the number of time-domain zero crossings within a signal. It gives an approximate measure of the signal's noisiness.

$$Z_t = \frac{1}{2}\sum_{n=1}^{N} |\, sign(x[n] - sign(x[n-1])) \,| \qquad (1)$$

  where *sign* function is 1 for positive *x[n]* and –1 for negative *x[n]*.

- Spectral Centroid [4]: A representation of the balancing point of the spectral power distribution within a frame that is computed as follows:

$$SC = \frac{\sum_k kX[k]}{\sum_k X[k]} \qquad (2)$$

where *k* is a correspond index to a frequency bin, within the overall measure spectrum, and *X[k]* is the amplitude of the corresponding frequency bin.

- Spectral Rolloff [4]: A measure of frequency, which is below 95 percentile of the power spectral distribution. It is a measure of the "skewness" of the spectral shape – the value is higher for right-skewed distributions

$$SR = K, \text{ where } \sum_{k<K} X[k] = 0.95 \sum_{k} X[k] \quad (3)$$

- Spectral Flux (also known as Delta Spectrum Magnitude) [4]: A measure of spectral difference, thus it characterizes the shape changes of the spectrum. It is a 2-norm of the frame-to-frame spectral magnitude difference vector

$$SF = \| X[k] - X[k-1] \| \quad (4)$$

where *X[k]* is the complete spectral magnitude of a frame.

- MFCC, also called Mel-Frequency Cepstral Coefficients [2],[5],[6],[7],[8],[9],[10]: A compact representation of an audio spectrum that takes into account the non-linear human perceptual of pitch, as described by the Mel scale. It is the most widely used feature in speech recognition. MFCC is particularly useful for analysing complex music due to its low-dimensionality, smooth version of log spectrum, the ability to discriminate between different spectral contents [6] and to somehow discard differences due to pitch evolution. MFCC calculation can be done through the following steps [11]:

    1. Convert signal into short frames
    2. Compute discrete Fourier transform of each frame
    3. The spectrum is converted to the log scale
    4. Mel scaling and smoothing the log scale spectrum
    5. Discrete cosine transform is calculated (to reduce the spectrum to 40 coefficients)

1.2 Melody-related features

Melody, together with harmony, rhythm, timbre and spatial location makes up the main dimension for sound descriptions [12]. With the implicit information that it carries, melody has played an important role in music perception and music understanding. According to Selfridge-Field [13], it is the melody that makes music memorable and enables us to distinguish one work from another. Current research in music content processing such as music transcription, melody similarity, melodic search, melodic classification and query-by-humming, works closely with melodic information. So far, there are several ways of defining and describing a melody. Solomon [14] and Goto [15],[16] define melody as a pitch sequence. While some others define music as a set of attributes that characterize the melodic properties of sound, a set of musical sounds in a pleasant order and arrangement etc. [12]. Among those viewpoints, melody as a pitch sequence would be the most appropriate for finding repetition in music.

In pitch perception, humans recognize pitch as having two dimensions, which refer to pitch height and pitch chroma, respectively. Pitch chroma embodies the perceptual phenomenon of octave equivalence, by which two sounds separated by an octave (and thus relatively distant in term of pitch height) are nonetheless perceived as being somehow equivalent. Therefore, pitch chroma provides a basis for presenting acoustic patterns (melodies) that do not depend on the particular sound source. In contrast, pitch height varies directly with frequency over the range of audible frequencies. Hence, it provides a basis for segregation of notes into streams from separated sound sources. The function of these two pitch dimensions is illustrated when the same melody is sung by a male or female voice [17].

So far, we can consider two approaches in extracting melody-related features for the identification of representative musical excerpts, according to the dimension they emphasize. The first one emphasizes the pitch-height dimension. This approach uses features that carry pitch-height information to find repetitive patterns of music. Dannenberg and Hu, [18] used this approach to estimate pitch and identify the note boundaries of monophonic music. Dannenberg's pitch estimation was performed using an autocorrelation technique on overlapping windows. Autocorrelation computes the correlation between the signal and a time-shifted version of it. When the shift is equal to a multiple of fundamental periods, the correlation will be high. Dannenberg's pitch estimation algorithm also employed several heuristics to find the fundamental peak from others, which may occur due to strong partials above the fundamental and noise or sub-harmonics. This approach is only applicable for single pitch monophonic music. Nevertheless, for a real-world polyphonic music with a complex mixture of pitches, extracting the predominant one is highly complicated and practically infeasible.

The second approach emphasizes the pitch-chroma dimension and focuses on features that carry pitch-chroma information. Compared to the first approach, it

is practically more useful in music audio analysis since pitch-chroma holds the information related to the harmony or the melodic content of music. As it captures the overall pitch class distribution of music [3], the description it yields can be similar even if accompaniment or melody lines are changed in some degree. With this unique characteristic of pitch-chroma, there is no constraint of using this approach to analyze polyphonic music. In fact, application of harmonic or melodic content-related information in music content processing is not a novel strategy. The pitch histogram proposed by Tzanetakis [19] for measuring similarity between songs would be an example. Tzanetakis's pitch histogram is composed of a set of global statistical features related to the harmonic content. This set presents the most common pitch class used in the piece, the occurrence frequency of the main pitch class, and the octave range of the pitches of a song.

In identifying representative musical excerpts research, several authors ([3],[18],[20],[1],[18]) have employed chroma-based vectors to find the repetitive patterns of music. A chroma-based vector is basically an abstraction of the time varying spectrum of audio. It is computed mainly through restructuring a sound frequency spectrum into a chroma spectrum. Octave information is discarded through folding frequency components in order they fall into twelve distinct chroma bins which correspond to the twelve pitch classes [21]. Bartsch and Wakefield [1] performed autocorrelation to the chroma-based vector in order to identify the song extract, which holds the most repeated "harmonic structure". With a different formulation, Goto's [3] RefraiD method employed a similar 12-element chroma-based vector to the one that is used in [1], in order to analyze relationships between various repeated sections, and finally detecting all the chorus parts in a song and estimating their boundaries.

1.3 Dynamics-related features

In human auditory perception, loudness contrast captures listeners' ears. The music term "dynamics", which refers to relative loudness or quietness measurement of the sound, holds a significant role in expressive musical structure formation. In music composition and music performance, artists use dynamics to emphasize and shape the structure of music. Current research studies in music expressive performance analyze dynamics behaviour to evaluate the expressiveness of the performance. A real-time expressive music performance visualizing system, based on tempo and loudness spaces, has been built to help studying performance expressiveness. It depicts the dynamics and tempo behaviour of each performance done by different interpreters on the same piece of music [22]. Considering the significance of music dynamics in marking the occurrence of new music events, dynamics-related features have become unique and useful in music segmentation. When finding repetitions in music, proper identification of repetitions' boundaries is highly significant. So far, three dynamics-related features frequently appear in the existing work: Spectral Power, RMS and amplitude envelope.

- Spectral power [5]: For a music signal *s(n)*, each frame is weighted with a Hann window, *h(n)*:

$$h(n) = \frac{\sqrt{8/3}}{2}[1 - \cos(2\pi \frac{n}{N})] \tag{5}$$

where N is the number of the samples of each frame.

$$SP(k) = 10\log_{10}\left[\frac{1}{N}\left\|\sum_{n=0}^{N-1} s(n)h(n)\exp(-j2\pi\frac{n}{N})\right\|^2\right] \tag{6}$$

- RMS energy [4];[6]: A measure of loudness of the sound frame

$$RMS = \sqrt{\frac{1}{N}\sum_{k=0}^{N-1} x[k]^2} \tag{7}$$

where N is the number of samples in each frame.

- Amplitude Envelope [5]: A description of the signal's energy change in the time domain. Xu et al. [5] computed the signal envelope with a frame-by-frame root mean square (RMS) and a low 3rd order Butterworth lowpass filter [23] with empirically determined cutoff frequencies.

1.4 Extraction Approach

So far, there are two approaches in using the above mentioned low-level features attributes to obtain useful descriptions for detecting repetitions in music. They are the static and the dynamic approach. The first one computes low-level descriptions directly from the sound signal to represent the signal around a given time. Hence, in order to detect repetitive patterns in music, it is essential to find identical evolution of the features. The dynamic approach, proposed by Peeters et al. [8], uses features that model directly the temporal evolution of the spectral shape over fixed time duration. The difference between the two approaches is that the earlier one uses features that do not model any temporal evolution and only provide instantaneous representations around a given time window (i.e. only the successive sequence of the features models the temporal evolution of the descriptions).

Following the static approach, Steelant et al. [6] propose the use of statistical information of low-level features, instead of the features themselves, to find the repetitive patterns of music. These statistics are mainly the average and the variance of the instantaneous features over the whole signal. According to Steelant et al., global representations of the low-level features, which consist of their statistical information, can overcome the problem of very similar passages having different extracted coefficients, due to the large frame-step during features extraction process. In their research, they use feature sets, which contain mean and standard deviation of MFCCs, to find the repetitive patterns of music. Their algorithm, tested on a database of only 10 songs, showed a slight improvement when using the statistical information of the low-level features instead of using the features themselves.

In the dynamic approach side, Peeters et al. [8] compute dynamic-features through passing audio signal, $x(t)$ through a bank of N Mel filters. Short-Time Fourier Transform (STFT) is then used to analyse the temporal evolution of each output signal $x_n(t)$ of the $n \in N$ filters, noted $X_{n,t}(w)$, with a window size of *L*. According to Peeters et al., the window size that is used for STFT analysis determines the kind of structure (i.e. *short-term* or *long-term*) that can be derived from the signal analysis. At the end, only coefficients (*n,w*) with high Mutual information are kept. Even though this approach may greatly reduce the amount of data used for identifying repetition of music, it is only noticeable when one deals with a high dimensionality of feature attributes.

2 SEGMENTATION

Signal segmentation facilitates partitioning audio streams into short regions for further analysis. Finding appropriate boundary truncations is an indispensable process for certain content-based applications, such as audio content analysis, audio summarization, audio notation, etc. In this section, we will discuss different methods implemented for segmenting audio signals for later structural identification. In addition, we have grouped them according to their similarities and differences regarding implementation (i.e. model-free segmentation versus model-based segmentation).

In identifying representative musical excerpts from music audio files, we can distinguish between two segmentation processes: pre-analysis segmentation and post-analysis segmentation. As showed by its name, pre-analysis segmentation (sometimes also called frame segmentation) appears before content analysis process. In fact, pre-analysis segmentation is a crucial primary step for content analysis description. Normally, it partitions audio streams into fixed-length short regions for later content analysis. These short regions may sometimes partially overlap. As arbitrary fixed resolution segmentation of audio streams may cause unnatural partitions, high-level audio descriptions (such as, beat or note-onset information) could be useful in finding natural segmentation points and improve the overall pre-analysis segmentation performance [1].

On the other hand, post-analysis segmentation appears subsequent to content analysis process. The aim of this segmentation process is to identify appropriate boundaries for partitioning the audio streams into sections. This sections comprise a non-fixed number of successive short regions that have been output from earlier segmentation processes (as shown in Figure 3), based on their feature changes. Hence, the partitions we obtain using post-analysis have a longer duration than those from pre-analysis segmentation. Post-analysis segmentation assumes that the boundaries between two consecutive partitions segments should consist of abrupt changes in their features contents. Meanwhile, the feature values of the signals inside each partition segments are supposed to vary little or slowly. Since appropriate boundary truncations are rather significant for music structure, this segmentation process holds an important role in music pattern discovery, music excerpts identification, music summarization, etc.

Post-analysis segmentation strategies can be categorised into two groups, according to the similarities and differences in their implementations. Hence, we speak of model-free segmentation and of model-based segmentation. Model-free segmentation algorithms partition signals without requiring any training phase. In the case of model-based segmentation, a training phase is necessary in order to learn the models for segmenting. The building of a model is done by using a collection of examples, which correspond to the desired output from segmentation algorithms, as training samples. A few examples of model-based methods for segmenting music based on musical instruments are support vector machines (SVM), Neural Networks, Bayesian Classifiers and Hidden Markov Models (HMM).

A widely used model-free segmentation technique takes advantage of (dis)similarity measures [1], [8],[18],[3],[24],[25]. Foote [25] first proposed the use of local self-similarity in spotting musically significant changes in music. It is done by measuring the distance measure between feature vectors using Euclidean distance or the cosine angle between the parameter vectors. Similarity matrix is a two-dimensional representation that contains all the distance measures for all the possibilities of frame combinations. As every frame will be maximally similar to itself, the similarity matrix will have a maximum value along its diagonal. In

addition, if the distance measure is symmetric, the similarity matrix will be symmetric as well. A visual rendering of a similarity matrix, with a given grey scale value proportional to the distance measure, gives a clear image display of the occurrences of different sections in audio, as shown in figure 4. With the used of a cosine metric, similar regions will be close to 1 while dissimilar regions will be closer to –1. According to Foote, by correlating a similarity matrix with a kernel which is composed of self-similar values on either side of the centre points and of cross-similarity values between the two regions, along the diagonal of similarity matrix, it yields the time instant of audio novelty $N(i)$, which is useful for identifying the immediate changes of audio structure. Audio novelty can be represented by

$$N(i) = \sum_{-L/2}^{L/2} \sum_{-L/2}^{L/2} C(m,n) S(i+m, i+n) \tag{8}$$

where *i* is the frame number.

Given that novelty detection is based on the correlation process, the width of the kernel affects the resolution of the detection outcome. A small kernel, which detects novelty on a short time scale, is capable of identifying detailed changes in the audio structure such as the individual note events. On the other hand, a large kernel, which takes a broader view of the audio structure, compensates its indefinite detection with a better identification for longer structural changes, such as music transitions, key modulations, etc. A large kernel can be constructed by forming the Kronecker product of **C** with a matrix of one and applying a window to smooth the edge effects. Finally, segment boundaries are extracted by detecting peaks where the novelty score exceeds a local or global threshold. Binary tree structure is then constructed to organize the index points of the segment boundaries by the novelty score.

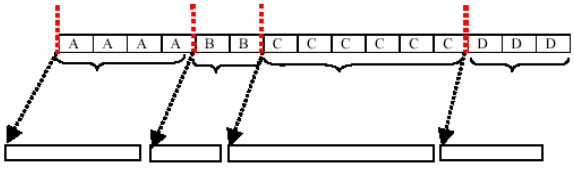

Figure 3: Illustration of post-analysis segmentation

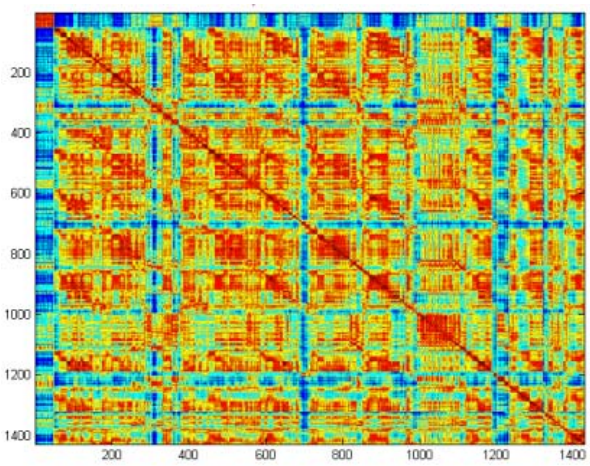

Figure 4: Similarity matrix computed from an audio excerpt from the soundtrack of "Beauty and the Beast". The MFCC derivatives were used as low-level features.

Hidden Markov Models (HMM), a well-known technique in pattern discovery and speech processing, is an example of model-based segmentation used in the research aimed to identifying representative musical excerpts. Aucouturier and Sandler [26] trained a 4-state ergodic HMM with all possible transitions to discover different regions in music, based on the presence of steady statistical texture features. In their experiments, they used the classic Baum-Welsh algorithm to train the HMM. The algorithm provided them with the Gaussian mixture distribution parameters and the transition probabilities for each state of the model. Finally, segmentation was deduced by interpreting the results from a Viterbi decoding algorithm for the sequence of feature vectors for the song. One of the two approaches used in Logan & Chu [7] is another example of applying Hidden Markov Models in a post-analysis segmentation task. Since the segmentation and the identification processes are closely related, HMM is capable of integrating both the segmentation process and identification process into a unified process. In other words, it completes both tasks by using a single algorithm. The application of HMMs for solving identification tasks will be discussed in the following section.

3 IDENTIFICATION OF STRUCTURAL SECTIONS

Identification of structural sections, the final task in identifying representative musical excerpts, aims to discover the patterns that are somehow repeated along a given music title. So far, there appear different approaches, including those which used in pattern recognition and image processing with the objective of completing the identification task. Here, we organize these approaches into three main groups: Self-similarity Analysis, Clustering, and Hidden Markov Modeling. In

the forthcoming subsections we discuss these approaches, including pros and cons of their specific algorithms.

3.1 Self-similarity Analysis

The occurrence of repetitive sections in the structure of music audio has caused researchers to relate music audio structure with fractal geometry phenomena in mathematics. A few methods with a self-similarity approach have been employed for identifying representative musical excerpts. One of them is the two-dimensional self-similarity matrix [25]. Seeing that self-similarity measurement is capable in expressing local similarity in audio structure, Bartsch & Wakefield [1] used a restructured time-lag matrix to store the filtering results that were obtained through applying a uniform moving average filter along the diagonals of the similarity matrix, for the aim of computing similarity between extended regions of the song. Finally, they selected chorus section of music by locating the maximum element of a time-lag matrix subject based on two defined restrictions: (1) the selection's time position indexes must have a lag greater than one-tenth of the song; (2) It happens less than three-fourths of the way into the song.

Goto's [3] RefraiD method is another example of using time-lag similarity analysis in identifying representative musical excerpts from audio music. Goto used 2-dimensional plot representations having time-lag as their ordinate, in order to represent the similarity and the possibility of containing line segments at the time lag. With an automatic threshold selection method, which was based on a discriminant criterion measure [27], time-lags with high possibility of containing line-segments are selected. These selected time lags are then used to search on the horizontal time axis on the one-dimensional function for line segments using the same concept of the previous threshold selection method. After that, groups are used to organize those line segments, with each group consisting of the integration of the line segments having common repeated sections. With the use of the corresponding relation between circular-shifted of chroma vector and performance modulation, Goto further improved the overall detection performance by tackling the problem in identifying modulated repetition. According to Goto, when an original performance is modulated by *tr* semitones upwards, its modulated chroma vectors satisfy,

$$\vec{v}(t) \doteq S^{tr}\vec{v}(t)' \tag{9}$$

where

$\vec{v}(t)$ = chroma vectors of modulated performance,

$\vec{v}(t)'$ = chroma vectors of original performance,

S^{tr} = shift matrix

$$\begin{pmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & & 0 & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 1 & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \\ 1 & 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix} \tag{10}$$

By using this strategy, Goto computed twelve kinds of extended similarity using the shift matrix and chroma vectors of original performances in order to represent the modulation of twelves semitones upwards. For each kind of extended similarity, the above-mentioned process of listing and integrating the repeated sections is performed, with the exception that the threshold adjusted for original performance vectors is used for modulation vectors as well. Goto later unfolded each line segments in each group to obtain unfolded repeated sections and $\lambda_{ij,..}$ its possibility of being chorus sections. Before the possibility of λ_{ij} of each repeated section is used for later calculation for V_i, a total possibility of each group for being a chorus section is adjusted, based on three heuristic assumptions,

i. The length of the chorus section has an approximate range. If the length is out of range. λ_{ij} is set to 0.
ii. Long repeated sections may correspond to a long-term repetition (e.g. the verseA, verseB and chorus) and it is likely that a chorus section is located near its end. Hence, if there exists a repeated section whose end is close to the end of another long repeated section (longer than 50 sec), its λ_{ij} is doubled.
iii. Because a chorus section tends to have two half-length repeated sub-sections within its section, a section that has those sub-sections is likely to be the chorus section. If there is a repeated section that has those sub-sections in another group, half of the mean of the probability of those two sub-sections is added to its $\lambda_{ij.}$

Finally, the group with the highest possibility value of V_i is selected as the chorus section.

3.2 Clustering

Clustering is a grouping technique that has been extensively used in image processing, machine learning and data mining. Clustering organizes a set of objects into groups, which all objects inside each group are somehow similar to each other. Logan and Chu [7] used a clustering technique to discover the key phrase of music. They then divided the sequence of features for the whole song into fixed-length contiguous segments,

as a starting point. Then an iterative algorithm proceeded according the following step:

1. Compute mean and covariance for each cluster with the assumption that each cluster has a Gaussian distribution.
2. Compute and store the distortion between each pair of cluster using Kullback-Leibler distance measurement.
3. Select the pair of clusters with the lowest distortion between them.
4. If it is less than a predefine threshold, combine these two clusters and go to step1.
5. If not, quit
6. Each distinct cluster is assigned a label, with all the frames inside this clusters are given this label.
7. Determine the most frequent label that occurs in the song.

By using this approach, Logan and Chu selected the longest section (which consists of the most frequent label that appears in the first half of the song) as the key phrase of the song. Results from the evaluation test showed that clustering approach performed the best when compared to Hidden Markov Modeling and randomly taken from the song. Nevertheless, the selected key phrase through clustering approach contains an unnatural starting and ending point, which was caused by a fixed amount of resolution in the segmentation process.

Other than using K-means, Foote and Cooper [27] proposed using Singular Value Decomposition (SVD). SVD is a dimension-reduction technique originally developed for still image segmentation, which can also be used for completing the task of segment clustering. SVD works by performing decomposition on a similarity matrix. In other words, it finds the repeated or substantially similar groups of segments through factoring a segment-indexed similarity matrix. Since it is not clear which clustering methods perform better, it would be worthwhile to make some objective back-to-back comparison between these two techniques.

3.3 Hidden Markov Modeling

Hidden Markov Modeling is another approach used in determining representative musical excerpt of music. HMM consists of three main parameters, number of states, state likelihood representation and transition matrix. The selection of the number of states and a good initialisation of the training data highly influence its outcome. With the application of HMM, the segmentation process and the identification process are integrated into a single unified process [7]. Hence, there is not necessary to perform any segmentation prior from using the HMM technique as the system learns the segmentation from the data itself. Unsupervised Baum-Welsh is used to train the HMM given the sequence of feature attributes of the songs. In HMM, each state corresponds to a group of similar frames in the song. With Viterbi decoding, the most likely state of each frame is determined, with each given a label. Finally, continuous segments are constructed by concatenating consecutive frames that have the same given label. Logan and Chu [7] chose the key phrase based on the duration and frequency of occurrence of these segments. In their studies, HMM overcame the problem of unnatural keyphrase beginning that was observed using the clustering approach, even though the HMM did not achieve a satisfactory performance in the evaluation test. Nevertheless, using a fixed number of states in HMMs may not be an optimal solution since in the real world music, the number of sections in music may vary significantly from one title to another one.

In identifying representative musical excerpt of music, insufficient number of states results in poor representations of data, whilst an excessive amount of states causes too detailed representations. Hence, previous knowledge of these parameters will definitely improve the overalll performance. Considering this factor, Peeters et al [8] proposed a multi-pass approach for automatic dynamic generation of audio summaries, with a combination of K-means clustering to help in estimating these parameters in order to further improve the overall performance of the HMM algorithm. Peeters et al. utilized restructured information boundaries, which were obtained from previous post-analysis segmentation processes through similarity measurement, in order to achieve a better estimation of the number of classes and their potential states for a K-means clustering algorithm. K-means is a type of unsupervised clustering technique for generating a specific desired number of disjoint and non-hierarchical clusters. K-means algorithm works by assigning each data point to the nearest mean, which is then adjusted to match the sample means of the data points that they are responsible for. For its initialisation, K-means uses random values.

The output from the K-means clustering is then used to initialise HMM learning model. As Logan & Chu ‘s approach, classic Baum-Welch algorithm is used to train the model. The output of the training are the state observation probabilities, the state transition probabilities and the initial state distribution are the output parameter. Finally, decoding using Viterbi algorithm with the given HMM and the signal features vectors, they obtained the state sequence corresponding to the piece of music. Peeters et al. research proved that an integrative (K-means & HMM) approach can

overcome the quick state-jump between states and produce a better and smoother state sequence.

4 DISCUSSION

It is common that each approach has its own benefits and drawbacks. In this section, we discuss the pros and cons on each approach used in identifying representative musical excerpts of audio music. Self-similarity analysis approach has the advantage of providing a clear and intelligibility view of audio structure. Nevertheless, it is not efficient for spotting repetition with certain degree of tempo change. A fixed resolution in its features representations may give a different representation view on the tempo changed repeated sections compared with its original section. Another problem with this approach is its threshold dependency in reducing noise for line segment detection. Threshold setting may vary from one song to another. Hence, a general setting threshold may not be valid for a wide range of audio. The clustering approach manages to overcome the problem of the sensitivity to tempo changes that suffers the previously mentioned approach, so long as boundary truncations are appropriate. However, one has to take notice that clustering, which organizes objects into groups based on their similarity, may produce complex representations of audio structure when there exists large variety of similarity measures in its features contents. Hence, this approach is not appropriate for music that has non-homogenous features contents. Moreover, clustering segments based on feature similarities may not have a clear meaning from the musical point of view. HMM approach with its transition statistical parameters is capable of handling the problem caused by non-homogenous segments that we have to face with music content analysis. Other advantages of HMMs are their efficiency in handling non-fixed length input and their independency in completing both segmentation task and identification task without any external support. Nevertheless, this approach has disadvantage in its expensive computation. In addition, HMM's performance efficiency highly depends on the number of states and a good initialisation. Insufficient number of states causes poor representations of data, whilst excessive states number causes very detailed representations. As the number of states in HMM can roughly correspond to the amount of different sections in the song, using a fixed number of states in HMMs may cause unsatisfactory outcomes.

By reviewing the existing approaches, we observe a few limitations in the aspect of algorithm evaluation. Databases consisting of a few hundreds of songs with different diversity and complexity will not be able to reflect the real-world music. Hence, by using such a database, it is quite impossible to obtain an objective evaluation on the algorithm efficiency for most of the existing music. Online evaluation tests, by giving access permission to internet users to run the algorithm on their audio files and acquiring feedback from the users, would be a way to have the algorithm executes on an unlimited range of music and obtain an appropriate assessment on the algorithm efficiency to the real-world music, even though there may have some other factors that would need to be considered. Another limitation is the method in weighting the importance of extracted music sections. The significance of the musical excerpts in audio signal highly depends on human perception. Musicians and non-musicians may not have a same viewpoint on "which sections are the representative excerpts of a piece of music". A musician may have a strong impression on the solo instrumental sections whereas it may not be the case for a non-musician. Hence, it would be useful to have two groups of listening subjects and taking into consideration the differences between these two groups when evaluate the significance of the extracted music sections.

5 CONCLUSION

It is apparent that identifying representative musical excerpts of audio files has relevance to music summarization, music content retrieval and the structural understanding of music. After reviewing the current developments in this area, we will attempt to explore more audio descriptors and make use of some other high-level descriptors in order to improve the current approach.

From this review paper, we notice that research in this area is still in an early stage. Much improvement still needs to be done to further develop the current research state and its practical applications. At the end of this review, we suggest that more attention can be given to the following factors:

1. Integration of some previously disregarded lower-level feature attributes (such as Pitch class profile, loudness, etc.) for better description of musical content of sound signal.
2. Make use of higher-level analysis techniques, such as beat detection, phrase detection, etc. in order to achieve better segment truncation.
3. Objective evaluation of the structural significance of musical excerpts in audio signals.
4. Integration of different approaches to discover the representative musical excerpts from audio files, in order to compensate each other's defects within themselves.

The impact of the above mentioned factors on the identification of representative musical excerpts will not

only improve the efficiency and effectiveness of the current process but also will yield a better semantic representation of musical audio signal. This will be practically useful for applications in audio indexing, audio browsing and audio database managing such as those that are being addressed in the SIMAC project, the development of which the presented overview is expected to contribute to.

6 ACKNOWLEDGEMENTS

This research has been partially funded by the EU-FP6-IST-507142 project SIMAC (Semantic Interaction with Music Audio Contents). More information will be found at the project website http://www.semanticaudio.org/. The authors would like to thank members of SIMAC and AUDIOCLAS projects at the Music Technology Group in the UPF for their useful comments and discussions.

REFERENCES

[1] M. A. Bartsch and G. H. Wakefield, "To Catch a Chorus: Using Chroma-Based Representations for Audio Thumbnailing," in *Proceedings Workshop on Applications of Signal Processing to Audio and Acoustics* pp. 15–18 (2001).

[2] J.J. Aucouturier, and M. Sandler, "Finding Repeating Patterns in Acoustic Musical Signals: Applications for Audio Thumbnailing," in *AES22 International Conference on Virtual, Synthetic and Entertainment Audio* (2002).

[3] M. Goto, "A Chorus-Section Detecting Method for Musical Audio Signals," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* vol. V, pp. 437-440 (2003).

[4] G. Tzanetakis and P. Cook, "Multifeature Audio Segmentation for Browsing and Annotation," in *Proceedings IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (1*999).

[5] Xu, Y. Zhu, and Q. Tian, "Automatic Music Summarization Based on Temporal, Spectral and Cepstral Features," in *Proceedings of IEEE International Conference on Multimedia and Expo* pp. 117-120 (2002).

[6] D.V. Steelant, et al., "Discovering Structure and Repetition in Music Audio, " *Eurofuse* (2002).

[7] B. Logan and S. Chu, "Music summarization using key phrases," in *International Conference on Acoustics, Speech and Signal Processing* pp. II–749–752 (2000).

[8] G. Peeters, A.L. Burthe, X. Roder, "Toward Automatic Music Audio Summary Generation from Signal Analysis," in *Proceedings of ISMIR: Third International Conference on Music Information Retrieval* pp. 94–100 (2002).

[9] J. Foote, "Visualizing Music and Audio using Self-Similarity," *ACM Multimedia* (1999).

[10] M. Cooper, and J. Foote, "Automatic Music Summarization via Similarity Analysis," in *Proceedings of ISMIR* pp. 81-85 (2002).

[11] L. Rabiner, and B. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall (1993).

[12] E. Gómez, A. Klapuri, B. Meudic, "Melody Descriptions and Extraction in the Context of Music Content Processing," *Journal of New Music Research* vol. 32.1 (2003)

[13] E. Selfridge-Field, Conceptual and Representational Issues in Melodic Comparison. In *Melodic Similarity – Concepts, Procedures, and Applications*, MIT Press (1998).

[14] L. Solomon, Music theory glossary. *Web publication*, last updated 2002, http://solo1.home.mindspring.com/glossary.htm (1997).

[15] M. Goto, "A Real-Time Music Scene Description System: Detecting Melody and Bass Lines in Audio Signals," in *Proceedings of the IJCAI Workshop on Computational Auditory Scene Analysis* pp. 31-40 (1999).

[16] M. Goto, "Robust Predominant-F0 Estimation Method for Real-time Detection of Melody and Bass Lines in CD Recordings," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* pp. II-757-760 (2000).

[17] J. D. Warren, S. Uppenkamp, R. D. Patterson, and T. D. Griffiths, "Separating Pitch Chroma and Pitch Height in the Human Brain," in *Proceedings of the National Academy of Sciences of the United States of America* vol. 17, pp. 10038–10042 (2003).

[18] R.B. Dannenberg, N. Hu, "Discovering Musical Structure in Audio Recordings," *International conference on Music and Artificial Intelligence*,

pp. 43-57 (2002).

[19] G. Tzanetakis, "Pitch Histograms in Audio and Symbolic Music Information Retrieval," *International Symposium on Music Information Retrieval* (2002).

[20] W.P. Birmingham, et al., "MUSART: Music Retrieval via Aural Queries," in *Proceedings Second International Symposium on Music Information Retrieval* pp. 73-81 (2001).

[21] R.B.Dannenberg and N.Hu "Pattern Discovery Techniques for Music Audio," in *Proceedings of the ISMIR: Third International Conference on Music Information Retrieval* pp. 63-97 (2002).

[22] G. Widmer, S. Dixon, W. Goebl, E. Pampalk, and A. Tobudic, "In Search of the Horowitz Factor" *AI Magazine* vol. 24, no. 3, pp. 111-130, (2003).

[23] G.M. Ellis, *Electronic Filter Analysis and Synthesis*, Artech House (1994).

[24] W. Chai, and B. Vercoe, "Structural Analysis of Musical Signals for Indexing and Thumbnailing," in *Proceedings of ACM/IEEE Joint Conference on Digital Libraries* (2003).

[25] J. Foote**,** "Automatic Audio Segmentation using a Measure of Audio Novelty," *In Proceedings of IEEE International Conference on Multimedia and Expo* vol. I, pp. 452-455 (2000).

[26] J. J. Aucouturier and M. Sandler, "Segmentation of Musical Signals using Hidden Markov Models.," in *AES 110th Convention* (2001).

[27] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Trans. SMC* vol. SMC-9, no. 1, pp. 62–66 (1979).

[28] J. Foote and M. Cooper., "Media Segmentation using Self-Similarity Decomposition," in *Proceedings of SPIE*, vol. 5021:1, pp. 67-75 (2003).