Music Searching and Browsing using Time-frequency Dictionaries

Soe Myat Thu
University of Computer Studies,Yangon
thuthu052228@gmail.com

Abstract

Nowadays, music searching and browsing are becoming very important as the huge amount of music are accessible over the Internet. Music recommender systems are essential especially for searching and browsing music catalogs. In this paper, retrieving the required information from acoustic music signal in an efficient way is considered. The system is to determine similarities among songs, particularly, a piece of input music signal compared with storage music song's signal into the database and then to retrieve the similar song. Representing the music signal having sparse nature is accomplished by Matching Pursuit with time-frequency dictionaries. In order to matching a candidate segment with the query segment, the music signal similarity measure is performed by Spatial Pyramid Matching.

1. Introduction

Music searching and browsing from audio signals is an interesting topic that receives a lot of attention these days. These feature sets are designed to reflect different aspects of music such as timbre, harmony, melody and rhythm. In addition, researchers use data from online sources that places music in a social context. Individual sets of audio content and social context features have been shown to be useful for various MIR tasks (e.g., classification, similarity, recommendation). Among them, similarity is crucial for the effectiveness of searching music information and the music segmentation. There exist three general recommendation approaches, namely the collaborative filtering approach, the content-based approach and hybrid approaches. In the early years of MIR (1985-95), research concentrated on rudimentary time and frequency domain features such as (1) windowed amplitude data and derived tempo statistics, and (2) windowed spectra, reduced spectra, and derived spectral statistics ("spectral measures"). The next generation of MIR (roughly 1995-2005), saw the introduction of more sophisticated features involving higher level time and frequency domain features such as beat histograms, Mel-frequency cepstral coefficients (MFCCs) and chromagrams. In addition, researchers began using more sophisticated statistics to aggregate the values of each feature within a song, and using newer machine learning techniques.

This paper presents a content-based approach to determine similar song from a database and retrieve a whole music song according to the input query. Because of the challenge of matching a candidate segment with the query segment, the system could significantly improve similarity measure using Spatial Pyramid Matching. And the retrieval time could considerably improve using Matching Pursuit (MP) Method. Our particular approach to choose music song also makes it possible automatic retrieval using matching pursuit features sets, for example for use in browsing rapidly through a list of possible song of interest returned by a search engine. By guiding us to the most significant parts of a music song, it also allows the development of fast and efficient methods for searching very large collections based purely on the audio content of the song, sidestepping the computational complexity of existing content-based search methods.

The rest of this paper is organized as follows. In Section 2, related work on music structure analysis for music searching and browsing system is discussed. The framework

of the system is introduced in Section 3. The feature extraction technique and similarity matrix from the musical audio signals are addressed in Section 3.1 and Section 3.2. In Section 4, overview of our proposed system is represented an example of music features extraction with a dictionary and pattern similarity discovery with pyramid kernel level. Evaluation study is provided in Section 5 followed by the Conclusion in Section 6.

2. Related Work

Automatic analysis of the structure has been studied mainly for the music information retrieval search engine of creating a meaningful summary of a musical piece. One of the first works operating on acoustic signals was by Logan and Chu, describing an agglomerative clustering and hidden Markov model (HMM) based approaches for key phrase generation [1]. They used mel-frequency cepstral coefficients (MFCCs) from short (26 ms),overlapping frames. The clustering method grouped the frames together iteratively until a level of stability had been reached. In the HMM method they trained an ergodic HMM with only few states, hoping that each state would represent a musical part, and used the Viterbi decoded state sequence as the description of the musical structure. The HMM approach was taken further by Aucouturier and Sandler using spectral envelope as the feature [2]. It was noted in both these studies that when using such short frames, the HMM states did not model musically meaningful parts, as was hoped. Abdallah et al increased the frame length considerably and the number of states up to 80 [3]. After acquiring the state sequence, each frame was provided with some knowledge about the surrounding context by calculating a state histogram in a 15 frame window. The histograms were then used in clustering the frames by optimising a cost function with simulated annealing. Rhodes et al added a term to control the duration of stay in a certain cluster [4], while Levy et al refined the clustering method to a context aware variant of fuzzy C-means [5]. Another popular starting point of the analysis is to calculate frame-by-frame similarities over the whole signal, constructing a self-similarity matrix. Foote proposed to use the similarity matrix for visualising music [6]. It was noted that the parts of music having similar timbral characteristics created visible areas in the similarity matrix. The borders of these areas were sought and used in segmenting the piece in [7]. In [8] Foote and Cooper used a spectral clustering method to group similar segments. When the used feature describes the tonal (pitch) content of the signal instead of general timbre, e.g., chroma instead of MFCCs, repetitions generate off-diagonal stripes to the similarity matrix instead of rectangular areas of high similarity. Such stripes reveal similar sequential structures, e.g. melody lines or chord progressions, instead of just denoting parts having similar timbral characteristics, or sounding the same. The two main approaches (HMM-based “state” method and “sequence” method relying on stripes in the similarity matrix) were compared by Peeters [9]. He noted that as the sequence approach requires a part to occur at least twice to be found, the HMM approach would be more robust analysis method. Still, the stripes have been used in structure analysis by several authors. Bartsch and Wakefield extracted chroma from beat-synchronised frames and used the most prominent off-diagonal stripe to define a thumbnail for the piece [10]. Lu et al proposed a distance metric considering the harmonic content of sounds, and used 2D morphological operations (erosion and dilation) to enhance the stripes [11]. In popular music pieces, the clearest repeated part is often the chorus section. Goto[12] aimed at detecting it using chroma, and presented a method for handling the musical key modulation sometimes taking place in the last refrain of the piece. Music tends to show repetition and similarities on different levels, starting from consecutive bars to larger parts like chorus and verse. Some authors have tried to take this into account and proposed methods operating on several temporal levels. Jehan constructed several hierarchically related similarity matrices [13]. Shiu et al extracted chroma from beat-synchronised frames and then

used dynamic time warping (DTW) to calculate a similarity matrix between all the measures of the piece [14]. The higher level musical structure was then modelled with a manually parametrised HMM. Dannenberg and Hu gathered the shorter repeated parts and gradually combined them to create longer, more meaningful, parts in [15]. Later, Dannenberg used the stripes in similarity matrix to find similar musical sections, and then utilised this information to aid a beat tracker [16]. Chai proposed to take the context into account by matching two windows of frame level features with DTW. Sliding the other window while keeping the other fixed provided a method to calculate the similarity on different lags and to determine the lag of maximum similarity. Gathering this information in a matrix formed stripes of prominent lags, like the stripes in a similarity matrix. The longer stripes were then interpreted information about the repeats of structural parts [17]. Maddage et al proposed a method for analysing a musical piece combining different sources of information. They used beat-synchronised pitch class profile as the feature and detected chords with pre-trained HMMs. Using assumptions of the lengths of the repeated parts, fixed length segments were matched to get a measure of similarity. Finally heuristic rules, claimed to apply on English-language pop songs, were used to deduce the high-level structure of the piece [18].

3. Background

3.1. Matching Pursuit Representation

Matching Pursuit is part of a class of signal analysis algorithms known as Atomic Decompositions. These algorithms consider a signal as a linear combination of known elementary pieces of signal, called atoms, chosen within a dictionary.

MP aims at finding sparse decompositions of signals over redundant bases of elementary waveforms.

3.1.1. Dictionary Approximation

Matching pursuit that decomposes music signal into a linear expansion of waveforms that are selected from a redundant dictionary of functions. Wavelet transforms should be designed as follow: **Dictionary:** A dictionary contains a collection of blocks plus the signal on which they operate. It can search across all the blocks (i.e., all the scales and all the bases) for the atom which brings the most energy to the analyzed signal. **Book:** A book is a collection of atoms. Summing all the atoms in a book gives a signal. **Atoms:** An elementary piece of signal. An atom is organized by its Gabor atoms.

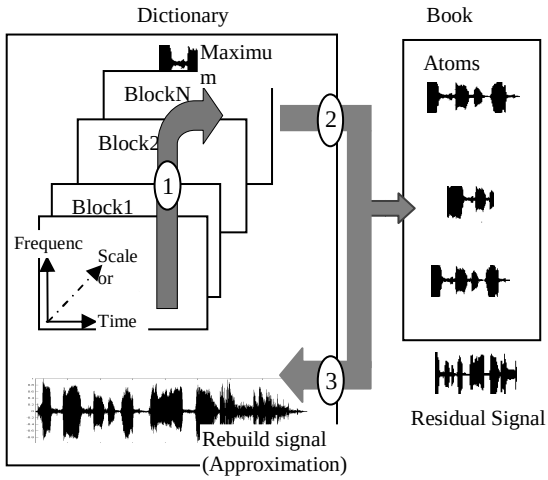


Figure 1. Implementation of our Matching Pursuit Algorithm

The implementation of the Matching Pursuit algorithm uses roughly 3 steps,

1. Update the correlations in the blocks, by applying the relevant correlation computation algorithm to the analyzed signal, and find the maximum correlation in the same loop.
2. Create the atom which corresponds to the maximum correlation with the signal (and store this atom in the book).
3. Subtract the created atom from the analyzed signal, thus obtaining a residual signal, and re-iterate the analysis on this residual.

Using Matching Pursuit method is price of efficiency and convergence. Time compression is quite excellent by extracting prominent atoms (features). In order to achieve the required information in our system, the algorithm could use the following steps:

1. initialization:

$$m = 0, x_m = x_0 = x; \quad (1)$$

2. computation of the correlations between the signal and every atom, using inner products :

$$\forall_w \epsilon\, D: \quad Corr(x_m, w) = |< x_m, w >| \quad (2)$$

3. search of the most correlated atom, by searching for the maximum inner product:

$$\hat{w}_m = \underset{w \in D}{argmax} Corr(x_m, w) \quad (3)$$

4. subtraction of the corresponding weighted atom $\infty_m \hat{w}_m$ from the signal :

$$x_{m+1} = x_m - \infty_m \hat{w}_m \quad (4)$$

$$where \ \propto_m =< x_m, \hat{w}_m >;$$

5. If the desired level of accuracy is reached, in terms of the number of extracted atoms or in terms of the energy ratio between the original signal and the current residual x_{m+1}, stop; otherwise, re-iterate the pursuit over the residual: $m \leftarrow m+1$ and go to step 2.

Music song signal analysis of our system is desirable to obtain sparse representations that are able to reflect the signal structures. The functions used for MP in our algorithm are Gabor function, i.e. Gaussian-windowed sinusoids. The Gabor function is evaluated at a range of frequencies covering the available spectrum, scaled in length (trading time resolution for frequency resolution), and translated in time. Each of the resulting functions is called an atom, and the set of atoms is a dictionary which covers a range of time-frequency localization properties. The Gabor function in our new search model is defined as

$$g_{s,u,\omega,\vartheta}(t) = K_{s,u,\omega,\vartheta}\left(\frac{t-u}{s}\right)\cos[2\pi\omega(t-u)+\theta] \quad (5)$$

where $\gamma = (s, u, \omega, \theta)$ denotes the parameters to the Gabor function, with s, u, ω, θ corresponding to an atom's position in scale, time, frequency and phase, respectively. The Gabor dictionary was implemented with the parameters of atoms chosen from dyadic sequences of integers [19].

3.2. Spatial Pyramid Matching for Pattern Discovery

Spatial Pyramid Matching is to find an approximate correspondence between these two sets by level. At each level of resolution, it works by placing a sequence of increasingly coarser grids over the features.

A pyramid match kernel allows for multi-resolution matching of two collections of features in a high-dimensional appearance space, but discards all spatial information. Another problem with this approach is that the quality of the approximation to the optimal partial match provided by the pyramid kernel degrades linearly with the dimension of the feature space. In our system, the approximate matching pattern discovery (SPM) is constructed the pyramid level and then the number of matches at level L is given by histogram intersection function. In determining SPM, SPM is used step by step level to improve matching musical data space and taking a weighted sum of the number of matches. At any fixed resolution, two points are said to match if they fall into the same cell of the grid. For matching pattern discovery, our system used histogram intersection function. The histogram intersection function is as follows:

$$\mathcal{I}(H_X^l, H_Y^l) = \sum_{i=1}^{D} \min\,(H_X^l\,(i), H_Y^l(i)) \quad (6)$$

In the following, it will be abbreviated $\mathcal{I}(H_X^l, H_Y^l)\ to\ \mathcal{I}^l$. To achieve more definitely pattern matching, our system modified step by

step level pyramid kernel function. Note that the number of matches found at level 'l' also includes all the matches found at the finer level l+1. Therefore, the number of new matches found at level l is given $\mathcal{I}^l - \mathcal{I}^{l+1}$ fo l=0,…….L-1. The weight associated with level is set to $\frac{1}{2^{L-l}}$, which is inversely proportional t cell width at that level. The definition of pyramid kernel is:

$$K^L(X,Y) = \mathcal{I}^L + \sum_{l=0}^{L-1} \frac{l}{2^{L-l}} (\mathcal{I}^l - \mathcal{I}^{l+1}) \qquad (7)$$

[20].

Spatial Pyramid Matching for Patter Discovery is its efficiency, its use of implic correspondences that respect the joint statistic of co-occuring features, and its resistance t 'superfluous' data points. Since pyramid match kernel is also positive-definite function, convergence is guaranteed, model free and effective for finding sparse over-complete representation.

4. Music Searching and Browsing System

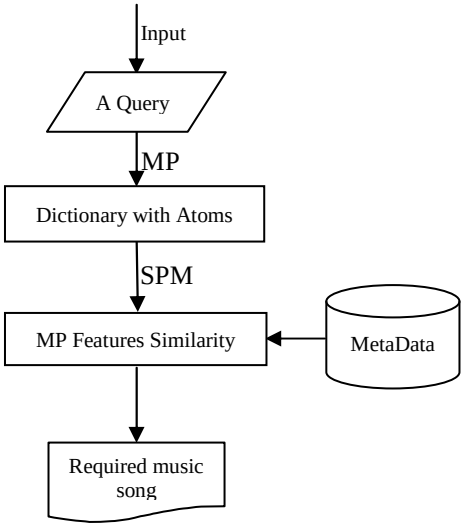


Figure 2. Overview of Music Searching and Browsing System

A block diagram of the system can be seen in Figure 2. The proposed method creates matching pursuit features from a query input music signal. Music signal structural features are represented as a dictionary with most prominent atoms that match their time-frequency signature.

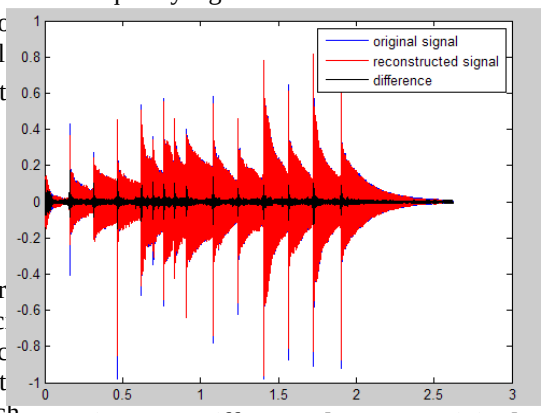
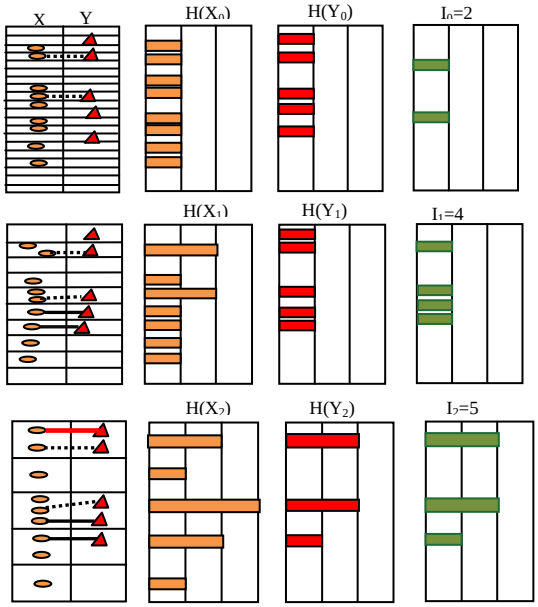


Figure 3. Difference between original signal and reconstructed signal

In Figure 3, it shows an example music song signal dictionary containing maximum prominent 1000 atoms at data sampled 26.2kHz analyzed with Matching Pursuit, an efficient implementation of the algorithm.



(a) Features sets (b) Histogram pyramids (c) Intersections

Figure 4. An Example of Pattern Discovery with histogram intersection function.

A pyramid matching determines a partial correspondence by matching feature points once they fall into the same histogram bin. In this example, two 1-D feature sets are used to form one histogram pyramid. Each row corresponds to a pyramid level. In (a), the set X is on the left side, and the set Y is on the right. (Features Points are distributed along the vertical axis, and these same points are repeated at each level.) Bold dashed lines indicate a pair matched at this level, and bold black lines indicate a match already formed at a former resolution. In (b) multi-features histograms are shown, with bin counts along the horizontal axis. In (c) the intersection pyramid between the histograms in (b) are represented.

Pyramid match kernel measures similarity achieved from a partial matching between two sets as shown in Figure 4. The optimal description of the music song from the meta database is found in respect to the faster and more similar function defined in Spatial Pyramid Matching method as detail described in Section 2. Using Matching Pursuit incooperated with Spatial Pyramid Matching method in this new search engine model, the new model can be optimised by occupying different groups in order which eliminates much of the search space.

5. Evaluation Study

Music songs are meaningful and no societies without music. In religion, sports and work, music songs are essential for social activities. In music searching and browsing system, new search system study and analyze the problem of music searching and browsing to become more and more similarity and effectively. The performance of the new search system will be evaluated in simulations browsing the similar structure of a set popular music pieces.Our proposed method needs ability to test on a collection of tracks including music genres in the database such as punk, hip-hop, jazz, metal, rock, country songs, etc. Our new search system will test South by Southwest Dataset used in previous work [21].Using matching pursuit and spatial pyramid matching method, the approach would be more effective and efficient than existing methods in retrieving similar music information. Performance metric of our proposed method can be measured in terms of accuracy.

6. Conclusion

In music information retrieval, music searching and browsing particular music songs in an efficient manner is still demanding. We demonstrate a promising approach for new search engine model in music information retrieval. The new system would use matching pursuit and spatial pyramid matching for determining significant features of music pieces and retrieving music queries in efficient way. The feature sets will be achieved by matching pursuit method as training and testing data. Retrieving similar music pieces from a database is completed by matching the MP features space by step by step level using spatial pyramid matching. Better speed and accuracy on large collection of musical songs can be expected upon the whole architecture of the system.

References

[1]. B. Logan and S. Chu." Music summarization using keyphrases*". In Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, pages749–752, Istanbul, Turkey, June 2000.*

[2]. J,J.Aucouturier and M. Sandler. "Segmentation of musical signals using hidden Markov models*". In Proc. of 110th Audio Engineering Society onvention, Amsterdam, The Netherlands, May 2001.*

[3]. S. Abdallah, K.Nolad, M. Sandler, M.Casey, and Rhodes." Theory and evaluation of a Bayesian music structure extractor.*" In Proc. of 6th International Conference on Music Information Retrieval, London, UK, Sept. 2005.*

[4]. C. Rhodes, M. Casey, S. Abdallah, and M. Sandler." A Markov-chain Monte-Carlo approach to musical audio segmentation." *In Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 797–800, Toulouse, France, May 2006.*

[5]. M. Levy, M. Sandler, and M. Casey." Extraction of high-level musical structure from audio data and its application to thumbnail generation." *In Proc. of*

IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 13–16, Toulouse, France, May 2006.

[6]. J. Foote." Visualizing music and audio using self-similarity." *In Proc. of ACM Multimedia, pages 77–80, Orlando, Florida, USA, 1999.*

[7]. J. Foote. "Automatic audio segmentation using a measure of audio novelty." *In Proc. of IEEE International Conference on Multimedia and Expo, pages 452–455, New York, USA, Aug. 2000.*

[8]. J. T. Foote and M. L. Cooper. "Media segmentation using self-similarity decomposition. "*In Proc. of The SPIE Storage and Retrieval for Multimedia Databases, volume 5021, pages 167–175, San Jose, California, USA, Jan. 2003.*

[9]. G. Peeters. Deriving musical structure from signal analysis for music audio summary generation:"sequence" and "state" approach. In Lecture Notes inComputer Science, volume 2771, pages 143–166.Springer-Verlag, 2004.

[10]. M. A.Bartsch and G.H Wakefield. To catch a chorus: "Using chroma-based representations for audio thumbnailing." In Proc. of 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pages 15–18, New Platz, New York, USA, Oct. 2003.

[11]. L. Lu, M. Wang, and H.-J. Zhang. "Repeating pattern discovery and structure analysis from acoustic music data." *In Proc. of Workshop on Multimedia Information Retrieval, pages 275–282, New York, USA, Oct. 2004.*

[12]. M. Goto." A chorus-section detecting method for musical audio signals." *In Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 437–440, Hong Kong, 2003*.

[13]. T. Jehan." Hierarchical multi-class self similarities." *In Proc. of 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pages 311–314, New Platz, New York, USA, Oct. 2005*.

[14]. Y. Shiu, H. Jeong, and C.-C. J. Kuo. "Musical structure analysis using similarity matrix and dynamic programming." *In Proc. of SPIE Vol. 6015 - Multimedia Systems and Applications VIII, 2005.*

[15]. R. B. Dannenberg and N. Hu. "Pattern discovery techniques for music audio. "*In Proc. of 3rd International Conference on Music Information Retrieval, pages 63–70, Paris, France, Oct. 2002*.

[16]. R. B. Dannenberg. "Toward automated holistic beat tracking, music analysis, and understanding." *In Proc. of 6th International Conference on Music Information Retrieval, pages 366–373, London, UK, Sept. 2005.*

[17]. W. Chai. "Semantic segmentation and summarization of music: methods based on tonality and recurrent structure." IE*EE Signal Processing Magazine, 23(2):124–132, Mar. 2006*.

[18]. N. C. Maddage, C. Xu, M. S. Kankanhalli, and X. Shao. "Content-based music structure analysis with applications to music semantics understanding*." In Proc. of ACM Multimedia, pages 112–119, New York, New York, USA, Oct. 2004*.

[19]. S.Chu,S.Narayanan and C-C.Jay Kuo,"Environmental Sound Recognition using MP-Based Features",University of Southern California,Los Angeles,CA 90089-2564.

[20].S.Lazebnik,C.Schmid,J.Ponce, "Spatial Pyramid Matching".

[21]. M. Hoffman, D. Blei and P. Cook, "Content-Based Musical Similarity Computation Using The Hierarchical Dirichlet Process", *ISMIR –session 3a Content-Based Retrieval, Categorization and Similarity 1, 2008.*