Discovering multi-level structure in folk music

Ciril Bohak **Matija Marolt**

University of Ljubljana,
Faculty of Computer and Information Science
E-mail: {ciril.bohak,matija.marolt}@fri.uni-lj.si

Abstract

In this paper we present a novel approach for discovering multi-level structure in folk song recordings. While this problem was previously addressed for commercial music, only single-level structure in form of stanzas was addressed for folk music. Our approach is an extension of our previous work on finding repeating stanzas in folk song recordings which relied on detection of vocal pauses and was therefore not suitable for choir and instrumental music. Instead of detecting vocal pauses, we calculate similarities between several short song excerpts and the entire song to obtain a repetition patterns. Next we estimate repetitions according to their degree of similarity and use a time series motif discovery method to extract the multi-level song structure. In the end we align the obtained structure to the audio. We tested our approach on a collection of folk songs with single-level and multi-level structure and we show that our approach finds motifs for both types of songs.

1 Introduction

One of the reasons why humans find music interesting is its structure in form of repeating patterns. While the musicologists can find this structure by listening to the music, much research was also done on automatic structure finding with use of music information retrieval (MIR) methods. An overview of such methods is presented in [11]. Finding a structure in music can be presented as search for repeating patterns in signals and thus methods of time series analysis can be applied to the domain as well. An overview of methods, used in time series data mining [5], gives an insight into how such methods can be used in different research domains.

Musical structure usually consists of repeating patterns on different level of perception. Structure of commercial music typically consists of verses and choruses. However this is not usually true for folk music. Since folk music is very simple its structure mostly consists of repeating parts - stanzas. In context of discovering the structure in the music this would indicate that discovering structure might be easier for folk music than it is for commercial music. However, this shows not to be true due to specific properties of folk music. While commercial music is mostly recorded in studios with professional singers and in dedicated recording environments, this is not true for folk music. Folk music is recorded in the field where level of noise is much higher. Performers are mostly not trained singers which results in inaccurate singing regarding the pitch as well as the tempo variance. Recordings might also contain interruptions or pauses. While most of folk music consists of repeating stanzas, some have more complex single-level structure (e.g $ABAB$), or even more complex multi-level structure (e.g. $[A_1A_2A_2]\,[B_1B_2]\,[A_1A_2A_2]$). In this paper we address problem of discovering single- and multi-level structure in folk songs. Presented approach is partly based on our previous work [1].

In the following section we present the related work. Section 3 contains description of our method followed by experiments and results in Section 4. At the end we present our conclusions and pointers for the possible future work in Section 5.

2 Related Work

Segmentation is one of main approaches for discovering structure. Broad overview of current state-of-the-art methods are presented in [11]. Most of segmentation tasks were done on commercial music such as chorus detection [6]. Media segmentation methods which use self-similarity matrix decomposition are presented in [4]. Evaluation of such methods has been made possible through Music Information Retrieval Evaluation eXchange (MIREX) task presented in [3]. In recent years some contributions addressed segmentation of folk music as well. Discovering meaningful parts in folk music is presented in [7]. Robust segmentation of folk songs into repeating stanzas with use of symbolic template is presented in [10]. Folk song segmentation method that combines vocal pauses detection and shifted chroma features is presented in [1]. A fitness measure that address problem of discovering repetitive structure in commercial as well as in classical and folk music is presented in [9].

Musical signal can also be represented as time series. Time series data mining was used for structure analysis and motif discovery in form of dynamic time warping (DTW) for different purposes such as query by humming [2] or finding repeating patterns [1,10]. More recent approach uses time series structure features for unsupervised detection of music boundaries [12].

3 Methodology

Our method is an extension of our previous work [1]. While our previous approach was designed for single-level structure detection in form of repeating stanzas, current approach is designed for multi-level structure discovery in folk music.

In our method we first extract chromatic features for the track. Next we randomly select 10 positions equally distributed throughout the whole track and calculate similarity between 10 second parts at selected positions and the whole track at 1 second resolution. With this we obtain 10 similarity curves of repetitions in the track. We next align the obtained curves according to the time lag between peaks in similarity curves. From aligned curves we calculate the average curve. Obtained curve represents the similarity structure in the track. We align the curve to the audio track and extract the structure using time series motif extraction algorithm.

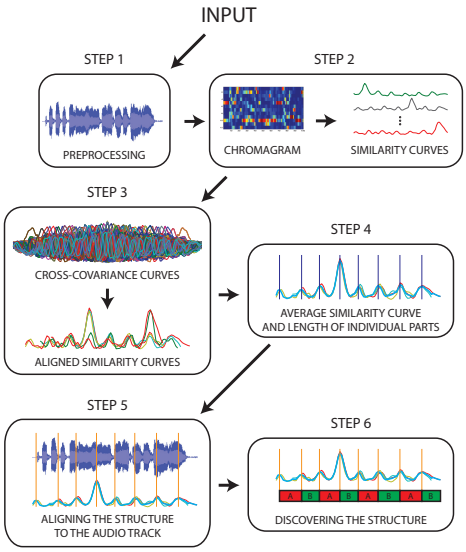

Figure 1: Outline of the method.

The method consists of several steps: (1) preprocessing, (2) calculation of similarity curves, (3) alignment of similarity curves, (4) calculation of average similarity curve and length of individual parts, (5) alignment with audio track and (6) structure discovery. Method outline is presented in Figure 1.

3.1 Preprocessing

In preprocessing step we merge the input stereo audio into single channel. Next we resample the input to 11025 Hz and normalize it. Preprocessing is necessary for our calculation of chromatic features in the next step.

3.2 Calculating similarity curves

Similarity curve in this context represent the similarity between randomly selected part (of defined length) in track and the whole track. Similarity is calculated between CENS chromagram of selected part and chromagram of same length at every second of the audio track. Chromagrams are also shifted for 2 tones up and down the scale by step of one semitone to obtain shift-invariant similarity in a track. More details on calculating the similarity measure can be found in [1].

By calculating the similarity measure between selected part and whole track at every second of track we obtain the similarity curve. We calculate the curves for all 10 randomly selected parts. Example of such calculated curves are presented in Figure 2. The highest peaks of similarity curves represent the selected part in track, while other peaks represent similar parts of track according to presented similarity measure. Distances between peaks represent lengths of repeating parts in audio track. To obtain more accurate lengths of individual repeating part, we will align the curves as described in following section. The problem one can spot is that from obtained similarity curves we can not determine the beginnings of individual part. We will address this problem in later section 3.5.

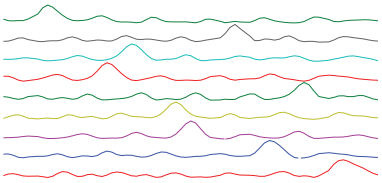

Figure 2: Aligned similarity curves.

3.3 Aligning similarity curves

With alignment of similarity curves we want to get more accurate assessment of lengths of individual parts in audio track. During alignment process we have to make sure we align same parts with one another. The alignment process is done in several steps described below.

3.3.1 Step 1: Selection of most representative curves

First we want to select similarity curves that are most similar to all the others. To obtain such curves we calculate cross-covariance for all pairs of normalized curves. We must take into the account that our cross-covariance method ignores any *NaN* value in the data. According to highest peaks in cross-covariance curves we select those that are most similar to all others. We do this by selecting the similarity curve that is most similar to all others, that is the curve whose mean of highest peaks of cross-covariance with other curves has the highest value. To obtain most relevant similarity curves we select those curves for which the highest value of cross-covariance is above 0.5 (at interval $[0, 1]$).

3.3.2 Step 2: Aligning the curves

Before we do the actual alignment of similarity curve we smooth selected curves with a low-pass filter. The alignment will be made to the most representative curve. We extract the peaks from each similarity curve that meet the conditions:

$$p_i = \begin{cases} min\_peak\_distance = 2\ sec, \\ min\_peak\_value = \max\left(mean(curve), 0.2\right) \end{cases},$$

where p_i represents i-*th* peak of selected curve. The parameter value 0.2 was obtained by testing on separate set of music.

The final alignment is done as follows: we select the highest peak in most representative curve and for every curve check whether a curve has a peak in vicinity of 5 seconds. If curve has a peak there we align the peaks and consequently curves as well. We do so for all the similarity curves that meet the condition. Such alignment of similarity curves is presented in Figure 3, where one can see that different *high peaks* might get aligned with different peak of most representative curve.

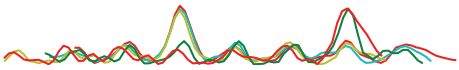

Figure 3: Aligned similarity curves.

3.4 Calculating average similarity curve and length of individual parts

Once we have successfully aligned the selected similarity curves we will calculate the average similarity curve which will be the base for discovering further structure of audio track. The average similarity curve is calculated from all selected and aligned similarity curves by calculating average value at selected time according to number of curves at that time as is presented in following equation:

$$avg\_sim\_curve(t) = \frac{\sum_{c \in Curves} c(t)}{n_c(t)},$$

where $c(t)$ is value of single similarity curve from set of all selected similarity curves $Curves$ at time t and $n_c(t)$ is number of defined similarity curves at time t.

From average similarity curve we can obtain lengths of individual repeating parts. One can do that by calculating distances between most prominent peaks in the similarity curve as shown in Figure 4. With such segmentation we can obtain single-level structure of individual audio track. However we do not know where the exact beginning of first part is in the audio and consequently we also do not know where all the other boundaries are according to the audio track.

Figure 4: Average similarity curve (left) and extracted boundaries for individual parts (right).

3.5 Aligning with track

We can use the obtained information to align the obtained structure to the audio track. We use the average length of parts to calculate how many repeated parts there should be in the audio track (as shown in experimental results the obtained number of parts is usually one part less than the real number of parts in the audio track).

From average length of individual part we calculate how many such parts should be in audio track according to its length: $n_{parts} = \frac{audio\_length}{avg\_part\_len}$. With calculated number of parts in track n_{parts}, we can make a simple alignment between the obtained structure and audio track. We calculate the estimated length of track from number of parts and their average length. Alignment is made by aligning the structure with beginning of non-silence in the audio track.

3.6 Discovering the structure

We use simple method for discovering motifs in time series - Symbolic Aggregate approXimation (SAX) [8]. Main idea of this approach is to discretise the signal into few classes according to their average value in the individual time segment. Method takes four inputs: *data*, *sliding window size* (N), *low-dimensional time approximation* (n) of data in selected window and *alphabet size* (a), which defines the amplitude low-dimensional approximation.

We used training set for determining what values of SAX parameters N, n and a are representing certain structure type. We have tested parameters on next intervals: $N = n \in [3, 10]$ and$a \in [3, 10]$, where parameters N and n were of same value, since we did not want to lose any time resolution.

For each set of parameters we have calculated the quotient q_{sax} between number of maximal repetitions of discovered motif max_motif in the time series and number of detected first-level parts n_{parts} in a track as described in first experiment. Next we calculated the mean value of q_{sax} quotients for individual parameter set. We predicted that obtained quotients are representing the complexity of structure. To determine the threshold that would define which quotient values determine single- and which multi-level structure we calculated the F1-measure for range of possible q_{sax} values on interval $[0, 3]$.

4 Experiments and Results

For testing purposes we have put together a collection of folk music from *Onder de groene linde* availible through Dutch Song Database and *EthnoMuse* [13] archives. Collection consists of different type of folk music as shown in Table 1. Overall it contains 437 minutes of manual annotated audio. Annotation contains beginning and ending times of all parts in a track.

4.1 Experimental setup

We have conducted several independent experiments. First experiment addresses single-level structure extraction. We have compared the results of presented method and our previous method. Second experiment was designed for discovering multi-level structure in music. Each experiment is individually described in following sections.

4.2 Comparing the proposed method with previous approach

Goal of this experiment was obtaining single-level structure in form of repeating stanzas as was done for our

Table 1: Composition of test set.

solo singing (OGL)	47
solo singing (EthnoMuse)	26
2 or 3 singers (EthnoMuse)	27
choir (EthnoMuse)	29
instrumental (EthnoMuse)	20
instrumental & singing (EthnoMuse)	16
Together	**165**

previous approach. Results of our previous method were limited to solo singing folk songs only due to limitations of the method. Method was not designed for music that does not contain vocal pauses which means that its use on choir, instrumental of mixed music is not meaningful. That is also the reason that we do not give head-to-head comparison of both methods on complete data-set but rather the results for data-set that methods were designed for. One other difference between methods is also that previous method was designed for single-level structure discovery only, while method presented in this paper addresses multi-level structure as well.

We do use same evaluation method for grading the results from both methods. We are grading how well our method is in locating the beginnings of individual parts. True positive cases are when method predicts beginning in 2 second neighborhood. Other predicted cases are treated as false positives. For determining the performance of our method we use already presented F1-measure.

Our previous method achieved F1-measure of 0.525 on solo singing data-set. Our presented method achieved F1-measure of 0.346. We have tested our method for higher tolerance as well. For 3 second tolerance F1-measure is 0.445 and for 5 second tolerance it is 0.628. This might seem like low score but we must take into the account that method was not designed for single type of folk music such as solo singing. There is also much room for further improvements of our method which are described in future work.

4.3 Experiment of discovering multi-level structure in folk music.

For this experiment we have manually labeled the tracks that have multi-level structure. We were to see whether our method will recognize the tracks with multi-level structure. We have manually separated the collection to music that has single-level structure and music that has multi-level structure. In our collection 78 tracks have single-level structure and the remaining 87 tracks have multi-level structure.

In our experiment we have randomly selected 12 tracks with single-level structure and 12 tracks with multi-level structure for our learning set. On this set we have trimmed the parameters of our structure discovery method for best results. Next we applied the method on the remainder of the set (141 tracks) was used as test set.

Best obtained F1-measure value for our training set was 0.8 and corresponding q_{sax} value was used on the test set resulting in final F1-measure of 0.69 which is significantly better than random result of 0.5.

With this experiment we have shown that even simple methods can be used for determining the complexity of track structure. Results might improve with use of more elaborate method. System with such results could be used as support for recommendation of track according to its structure complexity.

5 Conclusion and Future Work

We have presented a novel method for discovering multi-level structure in folk music. While this is still not a bulletproof method for extracting the multi-level structure it can still be used as support method for ethnomusicologists that are manually searching for such structure in audio tracks.

We have shown that our method can be used for two purposes: (1) as a single-level segmentation method and (2) as a method for determining the complexity of music structure. Results do not meet current state-of-the art methods, but method shows great potential and there is much space where method can be improved.

In the future we are planning to use more complex methods for discovering motifs in time series, since SAX can not cope the temporal changes which are quite common in folk music. We also want to introduce more advanced methods for aligning the obtained structure to the accompanying audio track. Another improvement is considered for selection of method for structure discovery. One future goal is to try such methods on our data as well. We also want to extend the test set with songs from other countries as well as test our method on bigger set of commercial music.

References

[1] C. Bohak, M. Marolt. Finding Repeating Stanzas in Folk Songs. ISMIR 2012, pages 451–456, 2012.

[2] W. A. Arentz, M. L. Hetland, and B. Olstad. Retrieving Musical Information Based on Rhythm and Pitch Correlations. *Journal of New Music Research*, 34(2):151–159, 2005.

[3] A. F. Ehmann, M. Bay, J. S. Downie, I. Fujinaga, and D. De Roure. Music Structure Segmentation Algorithm Evaluation: Expending on MIREX 2010 Analyses and Datasets. ISMIR 2011, pages 561–566, 2011.

[4] J. Foote. Visualizing music and audio using self-similarity. *Proceedings of ACM MULTIMEDIA '99*, pages 77–80, 1999.

[5] T.-C. Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.

[6] M. Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1783–1794, 2006.

[7] P. Van Kranenburg and G. Tzanetakis. A Computational Approach to the Modeling and Employment of Cognitive Units of Folk Song Melodies using Audio Recordings. ICMPC 2010 2010.

[8] J. Lin, E. Keogh, S. Lonardi, and P. Patel. Finding motifs in time series. *Proc. of the 2nd Workshop on Temporal Data Mining*, 2002.

[9] M. Müller, P. Grosche, and N. Jiang. A segment-based fitness measure for capturing repetitive structures of music recordings. ISMIR 2011, pages 615–620, 2011.

[10] M. Müller, P. Grosche, and F. Wiering. Robust Segmentation and Annotation of Folk Song Recordings. ISMIR 2009, pages 735–750, 2009.

[11] J. Paulus, M. Müller, and A. Klapuri. Audio-Based Music Structure Analysis. ISMIR 2010, pages 625–636, 2010.

[12] J. Serrà, M. Müller, P. Grosche, and J.L. Arcos. Unsupervised detection of music boundaries by time series structure features. AAAI 2009, pages 1613–1619, 2012.

[13] G. Strle and M. Marolt. The EthnoMuse digital library: conceptual representation and annotation of ethnomusicological materials. *International Journal on Digital Libraries*, 12(2-3):105–119, 2012.