FINDING REPEATING STANZAS IN FOLK SONGS

Ciril Bohak
University of Ljubljana
ciril.bohak@fri.uni-lj.si

Matija Marolt
University of Ljubljana
matija.marolt@fri.uni-lj.si

ABSTRACT

Folk songs are typically composed of repeating parts - stanzas. To find such parts in audio recordings of folk songs, segmentation methods can be used that split a recording into separate parts according to different criteria. Most audio segmentation methods were developed for popular and classical music, however these do not perform well on folk music recordings. This is mainly because folk song recordings contain a number of specific issues that are not considered by these methods, such as inaccurate singing of performers, variable tempo throughout the song and the presence of noise. In recent years several methods for segmentation of folk songs were developed. In this paper we present a novel method for segmentation of folk songs into repeating stanzas that does not rely on additional information about an individual stanza. The method consists of several steps. In the first step breathing (vocal) pauses are detected, which represent the candidate beginnings of individual stanzas. Next, a similarity measure is calculated between the first and all other candidate stanzas, which takes into account pitch changes between stanzas and tempo variations. To evaluate which candidate beginnings represent the actual boundaries between stanzas, a scoring function is defined based on the calculated similarities between stanzas. A peak picking method is used in combination with global thresholding for the final selection of stanza boundaries. The presented method was tested and evaluated on a collection of Slovenian folk songs from EthnoMuse archive.

1. INTRODUCTION

Folk music is receiving increased attention of the music information retrieval (MIR) community, as our awareness of the need for preserving cultural heritage and making it available to the general public grows. In order to process large quantities of folk song recordings gathered in ehtnomusicological archives, automated methods for analysis of these recordings need to be developed. Usually, such analysis starts with segmentation of recordings. Namely, folk songs are typically found within field recordings, which are integral documents of the process of folk music gathering and can, besides music, contain other kinds of content such as interviews with performers and background information. High-level segmentation of field recordings from EthnoMuse archive [13] into individual units can be done manually or by using automated methods [7].

For accurate analysis of individual songs, further segmentation into shorter parts is desirable. As folk songs typically consist of repetitions of melodically similar stanzas, segmentation boils down to finding the boundaries between repeating stanzas. This is quite different to segmentation of popular music, where songs typically consist of different parts, such as intro, verse, bridge and chorus.

While the structure of a popular song is usually more complex than the structure of a typical folk song, the segmentation of folk songs contains other challenges. While popular music is recorded by professional musicians in studios, folk songs are recorded in an everyday noisy environment (talking in the background, wind and other environmental noises, clapping ...) and singers are mostly untrained and usually older people that may sing out of tune, forget parts of lyrics or melody, interrupt their performances, switch to speaking etc.

In this paper, a novel approach for segmentation of folk songs into stanzas is presented. The algorithm is based on finding the vocal pauses in a folk song recording, derive the likely candidates for stanza beginnings from the pauses, score these candidates and select the best matching ones to obtain the final segmentation.

2. RELATED WORK

Most segmentation algorithms were developed for segmentation of popular or classical music. A broad overview of implemented methods is given in [11], where authors present state-of-the-art approaches and results of segmentation and structure discovery in music recordings. Typically approaches are based on calculating different sound features, which are used for construction of similarity or self-similarity matrices. First use of such matrices in MIR is presented in [4]. By finding repeating parts in matrices the structure of a musical piece can be inferred. Approaches use different features for construction of self-similarity matrices, two of the more popular are Mel-frequency cepstral coefficients [2, 5, 12] and chroma vectors [1, 6].

In recent years, several approaches to segmentation of folk music were presented. In [10] authors present a method for robust segmentation and annotation of folk songs. The

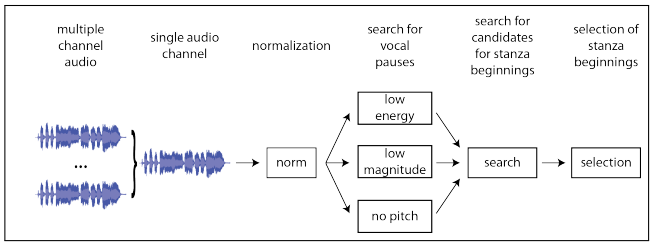


Figure 1. Outline of the proposed segmentation method.

presented approach uses chroma vectors in combination with a newly defined distance function for calculating the distance between individual stanzas and provided MIDI template. The method uses a MIDI representation of a single stanza to determine the length and expected pitch of each individual stanzas. Authors also present enhancements of the presented method, in which shifted chroma vectors are used to improve similarity between parts with shifted pitch.

A newer approach for detecting repetitive structures in music, presented in [9], introduces a novel fitness measure for defining the most representative part of song with the use of enhanced self-similarity matrix constructed from a variation of chroma-based audio features.

In [14], authors do not perform segmentation to search for repeating structure in folk songs, but are rather looking for their meaningful parts. Meaningful parts of folk songs are taken to be separated by breathing pauses, which are defined as parts of audio recording without detectable pitch.

EthnoMuse archive is a collection of audio field recordings, images, video recordings and metadata from Slovenian folklore. Archive contains more than 13.000 manuscripts, 1000 dance recordings, photos and more than 300 field recordings. Archive is not publicly accessible with exception of selected content. However parts of archive are published in book collections.

3. METHODOLOGY

Our segmentation method takes a folk song recording, as its input and outputs a set of boundaries, representing beginnings of individual stanzas in the recording. The method takes into account that performers are not professional singers, which may lead to pitch drifting over the duration of the piece, as well as considerable differences in tempo of individual stanzas.

The method consists of several steps: preprocessing, search for vocal pauses, search for possible beginnings of stanzas and selection of actual stanza beginnings. The general diagram of the suggested method is shown in Figure 1.

3.1 Preprocessing

The input audio signal is mixed from stereo to a single channel, the sample rate reduced to 11025 Hz and the amplitude normalized.

3.2 Detecting vocal pauses

Performers of folk songs are typically amateur singers which make characteristic breathing pauses, reflected in audio recordings as silence. These pauses are longer between stanzas, so they can be used to detect boundaries between stanzas. We use three approaches for detection of vocal pauses: short-term signal energy, amplitude envelope of the signal and the detected pitch. One can confirm such assumptions by listening to audio data from presented collection. This holds for solo and choir singing music, but does not hold for instrumental music.

3.2.1 Detecting vocal pauses according to signal energy

Vocal pauses in the audio signal are determined as parts of the signal where energy is below an experimentally determined threshold. Energy of the signal is computed on 200 ms long frames and the threshold is set to $\xi_1 = \frac{\overline{E}}{120}$, where $\overline{E}$ is the average energy of the signal. Consequent frames with energy values below the specified threshold are merged into one vocal pause. Vocal pauses shorter than $\xi_2 = 0.7$ times the average detected vocal pause length are ignored, to avoid the detection of short breathing pauses during singing. Parameter ξ_2 was also determined experimentally. Endings of detected vocal pauses, displayed red in Figure 2(a) (green are beginnings of vocal pauses), are later used as candidates for beginnings of stanzas.

3.2.2 Detecting vocal pauses according to signal envelope

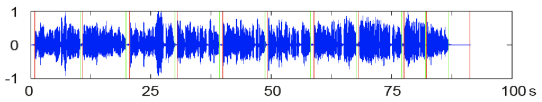
The amplitude envelope of a signal is obtained by filtering the full-wave rectified signal using 4th order Butterworth filter with a normalized cutoff frequency of 0.001. Vocal pauses are parts of the signal where the envelope falls below the threshold $\xi_3 = -60$dB, which was determined experimentally. Such parts of the signal are similarly as before merged into a single vocal pause, whereby we additionally merge all non-consequent parts that are less than $\xi_4 = 0.5$s apart, where the value ξ_4 was defined experimentally as well. As in the previous case, endings of detected vocal pauses are used as candidates for beginnings of stanzas and are displayed red in Figure 2(b) (green are the beginnings of vocal pauses).

3.2.3 Detecting vocal pauses according to relative difference of pitch

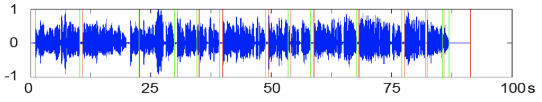
For detecting parts of the signal without any detectable fundamental frequency we are using the approach presented in [14]. The input signal is first resampled to $f_s = 1024$Hz. The resampled signal is then used as input for the YIN algorithm [3] that calculates fundamental frequencies for each frame of the signal. Fundamental frequencies are smoothed with a low-pass filter. Parts of the signal that differ more than 20 semitones from the average signal frequency are selected as vocal pauses.

In our approach we are merging vocal pauses longer than an experimentally obtained value $\xi_5 = 4$ms, while shorter vocal pauses are ignored. We are also taking into account the minimal length of a vocal pause which is in

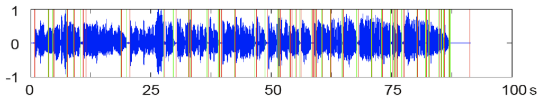
our case $\xi_6 = 250$ms. Again, endings of vocal pauses are used as candidates for stanza beginnings. In Figure 2(c) the detected vocal pauses are shown (green are the beginnings and red are endings) for a sample recording.



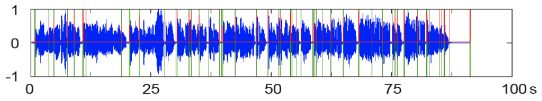
(a) Detected parts with low energy.



(b) Detected parts with low amplitude envelope.



(c) Parts with no detectable fundamental frequency.



(d) All detected stanza boundary candidates.

Figure 2. Comparison of methods for vocal pause detection. In images (a), (b) and (c) green are beginnings of vocal pauses and red are vocal pauses endings. In image (d) green are candidates for stanza boundaries and red is the value of fitness function for the candidates.

3.3 Finding candidates for stanza boundaries

In search for candidates for stanza boundaries we merge all sets of vocal pauses obtained with the previously described methods. An example of such a merged set is displayed in Figure 2(d) where the beginnings of vocal pauses are omitted and only their endings, which we consider as candidates for stanza boundaries, are shown in green. If candidates are present in several sets before merging, they are merged into a single candidate boundary.

We assume that the first candidate from the set represents the actual beginning of the first stanza. We then calculate the distance of the first $\xi_7 = 10$s of this first stanza to the 10s beginnings of all other stanza candidates determined by the candidates for stanza boundaries. The value of ξ_7 was chosen to represent approximately half of the average stanza length in our dataset. The calculation of distances between different stanza candidates takes the pitch drifting and tempo variations into consideration and is composed of four steps and is illustrated in Figure 3.

3.3.1 Step 1

We calculate 12 dimensional chromagrams, as defined in [8], for all stanza candidates using a window size of 50ms.

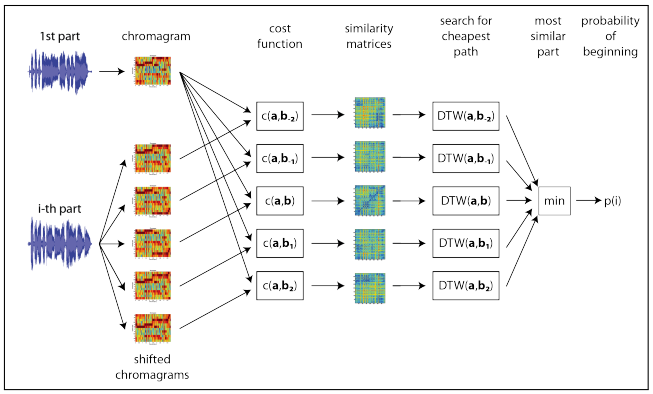


Figure 3. Outline of the algoritm for evaluating candidate stanza beginnings.

3.3.2 Step 2

We define a distance function between each pair of 12 dimensional chroma vectors as the root mean square (RMS) distance, which was also used for chorus detection in [6]:

$$c(\mathbf{a}, \mathbf{b}) = \frac{\sqrt{(\sum_i (a_i - b_i)^2}}{\sqrt{12}}, \tag{1}$$

where c is the distance function between two chroma vectors $\mathbf{a}$ and $\mathbf{b}$, a_i and b_i are i-th elements of chroma vectors.

3.3.3 Step 3

The defined distance function is used by the Dynamic Time Warping (DTW) algorithm for calculation of the total distance between the selected stanzas as:

$$c_p(p_1, p_2) = \sum_{l=1}^{L} c(p_1(l), p_2(l)) \tag{2}$$

where p_1 and p_2 are candidate stanza beginnings. $p_1(l)$ and $p_2(l)$ are the corresponding chroma vectors (previously labeled as $\mathbf{a}$ and $\mathbf{b}$), the index l takes values from the first (1) to the last (L) chroma vector in the selected audio part. The DTW is used for calculating the total distance between two stanza candidates:

$$c_{min}(d_j) = DTW(d_0, d_j) = min\{c_p(d_0, d_j)\}, \tag{3}$$

where c_{min} is the minimal cost between parts d_0 and d_j. A similar approach that uses DTW for calculating the cost was used in [10].

3.3.4 Step 4

To account for pitch drifting during singing, we also calculate distances between stanza candidates with shifted chroma vectors. The chroma vectors are circularly shifted up to two semitones up and down to compensate for the out-of-tune singing. We then select the lowest DTW distance as:

$$\begin{aligned} dist_{min}(d_0) &= 0, \\ dist_{min}(d_j) &= \min_{d_j^f, f \in [-2,2]} \{c_{min}(d_0, d_j^f)\}, \end{aligned} \tag{4}$$

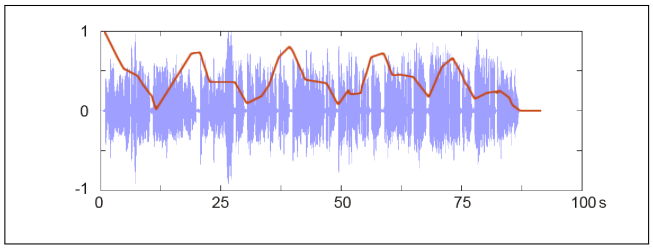


Figure 4. Fitness function for evaluating the candidate stanza beginnings.

where d_j^f represents a rotation of chroma vectors for the selected stanza candidate from two semitones downwards to two semitones upwards in steps of one semitone.

Finally, we define a fitness function for scoring the candidate stanza beginnings k_i as:

$$p(i) = \begin{cases} 0, & d_j \notin D \\ 1 - (\frac{dist_{min}(d_j)}{\max_j dist_{min}(d_j)})^2, & d_j \in D \end{cases} . \quad (5)$$

Figure 4 shows such a fitness function (Eq. 5) plotted on top of the audio signal. As the function is inversely proportional to the distance between the first stanza and a stanza candidate, higher fitness function values correspond to stanza boundaries which are more likely - stanzas are more similar and the candidate thus more likely represents a repetition of the original first stanza.

3.4 Selection of actual stanza beginnings

The selection of actual stanza beginnings is made with a simple peak picking algorithm in combination with a global threshold. In the defined fitness function, peaks represent the most likely stanza beginnings, so all peaks above a global threshold, corresponding to the average value of the fitness function, are picked as the actual boundaries between stanzas.

4. EXPERIMENTS AND RESULTS

The proposed method was tested on a set of folk songs from an ethnomusicological archive labeled as solo or choir singing, totalling 190 minutes in length and containing 135 units of solo or choir singing with an average duration of 100 seconds per unit. The average number of stanzas per unit was approximately 4, the average length of a stanza 18 seconds.

4.1 Evaluation of developed method

We performed an evaluation of vocal pause detection algorithms, as well as an evaluation of the whole segmentation method using the different detection algorithms.

The values of algorithm parameters $\xi_1 \ldots \xi_7$, used in vocal pause detection algorithms, were determined on a small set of recordings by evaluating a range of parameter values and choosing the ones for which the segmentation algorithm performed best. The algorithm itself is not very sensitive to changes in these parameters.

4.1.1 Evaluation of vocal pause detection

We evaluated individual approaches for detecting vocal pauses on the dataset containing 545 annotated vocal pauses. A detected vocal pause was considered as correctly detected, if it was within 2 seconds of the annotated vocal pause. Table 1 shows precision, recall and F-Measures of detection for individual methods and their combination. As shown in the table, combining all of the methods yields high recall and low precision. This is what we are aiming for at this first stage of the segmentation algorithm, because the second stage of the algorithm removes irrelevant vocal pauses and thus finding as many vocal pauses as possible is a priority.

Table 1. Comparison of vocal pause detection algorithms.

Algorithm	Precision	Recall	F-Measure
Energy	0,3336	0,8276	0,4755
Amplitude	0,5066	0,3729	0,4296
NoPitch	0,2793	0,5908	0,3793
Combination	0,0894	**0,9866**	0,1639

4.1.2 Evaluation of the method as a whole

Table 2 shows accuracy of the whole segmentation algorithm by using the four approaches to vocal pause detection described previously. One can see that the method, which uses the combination of all vocal pause detection algorithms, significantly outperforms the others. This result is expected, since individual vocal pause detection algorithms have lower recall, which means that we are already missing a number of annotated segment boundaries.

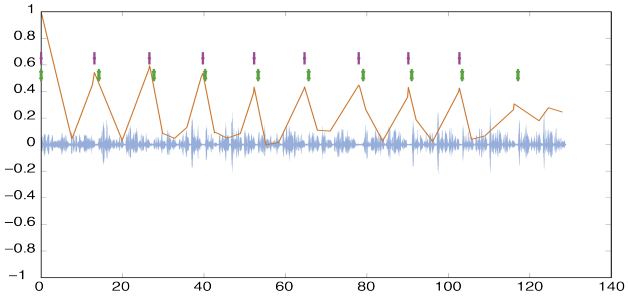
Table 2. Segmentation accuracy with different vocal pause detection algorithms.

Method	Precision	Recall	F-Measure
Energy	0,7592	0,4430	0,5595
Amplitude	0,6886	0,2574	0,3747
NoPitch	0,7447	0,3597	0,3597
Combination	0,6773	0,6435	**0,6600**

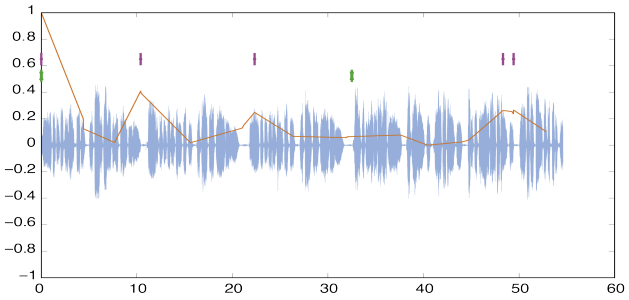
Our method performs well on songs that have strong vocal pauses, songs that consist of melodically similar stanzas and songs where the singing is approximately in tune. One can see an example of such vocal pause detection in Figure 5(a) where our method finds 9 out of 10 annotated vocal pauses. The 10th vocal pause is also clearly seen in the plotted fitness function, however the global thresholding prevents its detection.

The method fails on songs where the first stanza is incorrectly detected, because all stanzas are always compared to the first stanza. It also has difficulties in cases where stanzas are melodically very different, because comparison of chroma vectors relies on melodically similar stanzas. An example of such failure is shown in Figure 5(b).

In this example song consists of melodically significantly distinguished parts. First part of the song consists of three melodically similar parts that are also detected by our method, while same is true for the second part, its repeating parts are not similar to first part of the song.



(a) An example where our method performs well.



(b) An example where our method fails to find any annotated vocal pauses.

Figure 5. The figure shows an example where our method performs well (a) and an example where our method fails (b). The sound signal is plotted in blue, the fitness function is plotted in orange, true stanza beginnings are plotted with green stars (*) and the detected beginnings are plotted with pink plus signs (+).

4.2 Comparison with existing methods

We compared the proposed method with two existing folk song segmentation methods, results are shown in Table 3. The first method [14] only relies on detection of vocal pauses for segmentation and does not consider repetitions. Thus, the method covers most of the annotated vocal pauses (high recall), however it also detects many false positives, resulting in low precision.

We also compared our method with the method presented in [10] that uses symbolic transcription of a typical stanza as its input. Due to this additional prior knowledge, the method outperforms ours by a significant margin, as the approximate stanza melody and length are known to the algorithm. But, as this prior knowledge is not always available, the method cannot be used in all cases.

5. CONCLUSION AND FUTURE WORK

In this paper we presented a novel method for finding repeating stanzas in recordings of folk songs. The method relies on the detection of vocal pauses, which represent candidate stanza beginnings, which are then evaluated according to melodic similarity with the first stanza. The vocal pause detection algorithms and the method as a whole are separately evaluated on a dataset of folk song recordings. The method performs well, however several extensions are planned.

Table 3. Comparison of our method to other approaches.

Method	Precision	Recall	F-Measure
Kranenburg	0,149	**0,930**	0,257
Mueller (Δ^{fluc})	0,865	0,748	**0,802**
Our method	0,442	0,646	**0,525**

Our future work will include improvements with detection of the first stanza and processing of the fitness function, as well as adaptation of the method to instrumental tunes, where vocal pauses are not present. We also plan to integrate the developed method with our algorithm for high-level field recording segmentation, where it could be used to improve the accuracy of high-level segmentation. We will also try using more advanced preprocessing of the audio signal for environmental noise reduction. Other possibility is to try extracting pitch from the raw audio data and try finding repeating stanzas in symbolic domain.

6. ACKNOWLEDGEMENTS

Authors would like to thank all anonymous reviewers for thorough reviews with many suggestions on possible improvements and recommendations on how to improve this paper. This work could not been done without data provided by Institute of Ethnomusicology ar Research Centre of Slovenian Academy of Sciences and Arts. Research work was done as part of basic research and application project ETNOKATALOG: retrieval of semantic data from folk song and music, based on melodic and metro-rhythmic analysis funded by Slovenian research agency.

7. REFERENCES

[1] Mark A. Bartsch and Gregory H. Wakefield. To catch a chorus: Using chroma-based representations for audio thumbnailing. In *Applications of Signal Processing to Audio and Acoustics*, pages 15–19, New Platz, NY , USA, October 2001.

[2] Matthew L. Cooper and Jonathan Foote. Automatic music summarization via similarity analysis. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, pages 81–85, Paris, France, October 2002.

[3] Alain de Cheveigné and Hideki Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[4] Jonathan T. Foote. Visualizing music and audio using self-similarity. In *Proceedings of the 7th ACM international conference on Multimedia*, pages 77–80, 1999.

[5] Jonathan T. Foote and Matthew L. Cooper. Media segmentation using self-similarity decomposition. In *Proceedings of SPIE Storage and Retrieval for Multimedia Databases*, pages 167–175, 2003.

[6] Masataka Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Transactions on Audio, Speech & Language Processing*, 14(5):1783–94, 2006.

[7] Matija Marolt. Probabilistic segmentation and labeling of ethnomusicological field recordings. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pages 75–80, Kobe, Japan, October 2009.

[8] Meinard Müller. *Information Retrieval for Music and Motion*. Springer Verlag, 2007.

[9] Meinard Müller, Peter Grosche, and Nanzhu Jiang. A segment-based fitness measure for capturing repetitive structures of music recordings. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, 2011.

[10] Meinard Müller, Peter Grosche, and Frans Wiering. Robust segmentation and annotation of folk song recordings. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pages 735–740, Kobe, Japan, October 2009.

[11] Jouni Paulus, Meinard Müller, and Anssi Klapuri. Audio-based music structure analysis. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR 2010)*, pages 625–636, Utrecht, Netherlands, August 2010.

[12] Geoffroy Peeters. Toward automatic music audio summary generation from signal analysis. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, pages 94–100, Paris, France, October 2002.

[13] Gregor Strle and Matija Marolt. The ethnomuse digital library: conceptual representation and annotation of ethnomusicological materials. *International Journal on Digital Libraries*, pages 1–15, 2011. 10.1007/s00799-012-0086-z.

[14] Peter van Kranenburg and George Tzanetakis. A computational approach to the modeling and employment of cognitive units of folk song melodies using audio recordings. In *Proceedings of the 11th International Conference on Music Perception and Cognition*, pages 794–797, 2010.