Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2016, Article ID 8297987, 11 pages
http://dx.doi.org/10.1155/2016/8297987

Research Article

Probabilistic Segmentation of Folk Music Recordings

Ciril Bohak and Matija Marolt

Faculty of Computer and Information Science, University of Ljubljana, SI-1000 Ljubljana, Slovenia

Correspondence should be addressed to Ciril Bohak; ciril.bohak@fri.uni-lj.si

Received 7 August 2015; Accepted 15 February 2016

Academic Editor: Yan-Wu Wang

The paper presents a novel method for automatic segmentation of folk music field recordings. The method is based on a distance measure that uses dynamic time warping to cope with tempo variations and a dynamic programming approach to handle pitch drifting for finding similarities and estimating the length of repeating segment. A probabilistic framework based on HMM is used to find segment boundaries, searching for optimal match between the expected segment length, between-segment similarities, and likely locations of segment beginnings. Evaluation of several current state-of-the-art approaches for segmentation of commercial music is presented and their weaknesses when dealing with folk music are exposed, such as intolerance to pitch drift and variable tempo. The proposed method is evaluated and its performance analyzed on a collection of 206 folk songs of different ensemble types: solo, two- and three-voiced, choir, instrumental, and instrumental with singing. It outperforms current commercial music segmentation methods for noninstrumental music and is on a par with the best for instrumental recordings. The method is also comparable to a more specialized method for segmentation of solo singing folk music recordings.

1. Introduction

Structure is an inherent part of most music we listen to. It is what we recognize as repeating patterns of different musical modalities such as beat, rhythm, melody, harmony, or lyrics. Our appreciation of music is correlated with the ability to understand its underlying structure and predict what will occur next.

On the highest level, structure enables listeners to divide a piece of music into a set of segments. Two such segmentations are shown in Figure 1: a hierarchical structure (a) and a typical popular music structure (b). While understanding the structure of a piece of music is an integral part of our listening experience and is (at least for modern genres) not hard for listeners, accurate algorithms for automatic discovery of structure from audio recordings of various genres have yet to be developed.

Automatic discovery of structure from music recordings plays an important role in machine understanding of music and researchers have done extensive work in the past [1] to develop new segmentation algorithms. New approaches are evaluated annually within the Music Information Retrieval Evaluation eXchange (MIREX) [2], where in 2015 the best systems achieved accuracy of approx. 70% on a set of popular music recordings.

A large number of approaches are based on the concept of self-similarity matrices, describing within-song similarities according to a chosen set of features. One of the first such approaches was presented by Foote [3, 4], where segmentation was derived from a novelty measure calculated from a self-similarity matrix of mel-frequency cepstral coefficients (MFCCs). A similar approach where authors used chroma vectors instead of MFCCs was presented in [5] and an approach with both features combined was proposed in [6]. Jensen presented an approach where similarity was calculated in three separate domains: timbre, rhythm, and harmony [7], while Goto introduced time-lag matrices instead of self-similarity matrices for segmentation [8].

Recently, nonnegative matrix factorization (NMF) is often used to discover musical structure. In [9], the authors use NMF to search for acoustically similar frames in self-similarity matrices, while in [10] the authors use NMF for searching repeating patterns in beat-synchronous chromagrams. A novel adapted matrix factorization technique was presented by the authors in [11], who use convex constraints in the factorization process to decompose the similarity

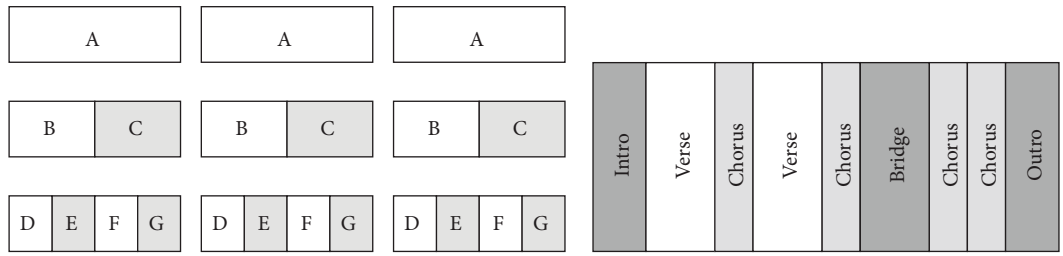


Figure 1: Examples of musical structures: (a) a hierarchical structure with repetitions on different levels and (b) typical structure of popular songs with alternating verse and chorus parts.

matrix in a way that individual centroids can be interpreted as different sections of musical piece.

A graph based musical structure analysis method is presented in [12], where a graph is constructed from a sparse representation of feature vectors. The segmentation of audio features is obtained by applying spectral clustering to the graph. The approach was tested on popular music with promising results.

A general approach for boundary detection in time series was presented in [13, 14]. The method was also applied to audio recordings for detection of repeating parts. It mimics short-term memory by encapsulating the most recent parts of a signal, assesses homogeneities and repetitions by pairwise comparison, and computes structure features and differences of these features, which yield a novelty measure, whose peaks indicate boundary estimates. This method was also combined with a chorus detection method [8] by [15].

Authors of [16] presented a novel approach that does not rely on self-similarity matrices for segmentation but uses the spectral graph theory. It produces low-dimensional encoding of the repetition structure and exposes hierarchical relationships among individual structural components. The same authors presented a different approach [17] to music segmentation that relies on an ordinal linear discriminant analysis method for learning feature projections to improve time-series clustering. They also propose latent structural repetition features, which provide a fixed-dimensional representation of global song structure and facilitate modeling across multiple songs.

The use of 2D-Fourier transform for clustering is presented in [18]. The magnitude coefficients computed from chroma features simplify the clustering problem since they are key and phase shift invariant. Authors explore various strategies to obtain segment boundaries and apply k-means clustering for labeling them.

The presented approaches were all developed for segmentation of popular or classical music. They thus assume that music is professionally recorded, with minimal noise and by professional musicians who deliver accurate performances. Many also rely on additional constraints, such as the presence of strong beats (for computing beat-synchronous features) or nearly constant tempo.

As we show in this paper, performance suffers when these assumptions are broken. We specifically study segmentation of folk music field recordings. Collections of folk music are being digitized rapidly, also due to increased funding for preservation of cultural heritage, so methods for their automatic annotation are sorely needed. These new methods should be reliable, adaptive, and robust to folk music specifics. Folk music recordings contain a number of challenges for automatic processing: they are typically noisy, as they are recorded in the field in everyday conditions. They are also usually performed by amateur singers and musicians, so performances may contain inaccurate singing, pitch drifting, forgotten lyrics, interruptions, large tempo deviations, and so forth. Ensembles in folk music are diverse, ranging from solo and choir singing to instrumental recordings with a variety of instrument families. On the other hand, the structure of folk music is not complex and usually consists of repetitions of the same melodic pattern, which is beneficial for automatic segmentation.

One of the first approaches that dealt specifically with segmentation of folk music recordings was presented by Müller et al. [19]. It requires a symbolic representation of a single repeating part as prior knowledge and bases the segmentation around a distance function computed with the dynamic time warping algorithm (DTW). The algorithm aligns the symbolic representation with $F0$-enhanced Chroma Energy Normalized Statistics (CENS) audio features, tolerating tempo deviations, while cyclic shifting of chroma features enables robustness to pitch drifting. An extension of the method that does not require prior knowledge was presented in [20], where authors introduce a novel fitness measure that can cope with strong variations in tempo, instrumentation, and modulation within and across the related segments of the music. Another adaptation of the method, also removing the need for prior knowledge, was presented by authors in [21].

A novel segmentation method was presented in [22]. The approach uses enhanced self-similarity matrices and a novel dynamic programming based fitness measure to find the most representative segment in a recording and its repetitions. The approach tolerates changes in transposition and tempo and was evaluated on popular and classical music, as well as folk songs. Its application to audio thumbnailing is presented in [23].

Authors have previously presented a probabilistic approach [24] for segmenting long ethnomusicological field recording consisting of different content, such as speech,

Table 1: Test collection details.

Type	Number of songs	Duration (min)
Solo singing (OGL)	47	156
Solo singing (EthnoMuse)	31	72
Two and three voices (EthnoMuse)	30	80
Choir (EthnoMuse)	35	92
Instrumental (EthnoMuse)	33	74
Instrumental and singing (EthnoMuse)	30	60
Total	**206**	**534**

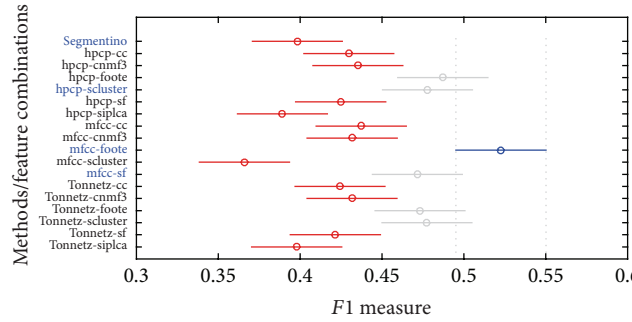


Figure 2: Comparison of segmentation methods with indication of significant differences in $F1$ measures.

multiple songs from different performers, and even bell chiming in single recording, into individual units according to their content type and their labeling. The presented model is designed to segment and label recordings on larger scale. Some obtained individual units (e.g., songs) can be considered as input into the approach presented in this paper.

In this paper, we present a novel method for segmentation of folk music field recordings, which include individual songs. We designed the method to be tolerant to field recording specifics, such as high levels of noise, inaccurate singing, pitch drifting, and tempo variations. We present an evaluation of our method on a collection of 206 folk music recordings and show that it performs well for a variety of ensemble types. It outperforms several state-of-the-art algorithms for segmentation of popular music and performs comparably to state-of-the-art methods for segmentation of solo singing folk song performances.

The paper is organized as follows. We first evaluate several state-of-the-art methods for segmentation of popular music on a collection of folk music recordings in Section 2. In Section 3, we present our method and in Section 4 its evaluation, analysis, and discussion. We conclude the paper and describe future work in Section 5.

2. Evaluating the State of the Art

In this section, we present an evaluation of several state-of-the-art algorithms for music segmentation and structure discovery on a collection of folk music recordings. The collection consists of 206 recordings of different types, as presented in Table 1, mostly taken from the *EthnoMuse* archive [25], as well as from *Onder de Groene Linde* (OGL) [26]. The collection contains approximately 9 hours of recordings with manually annotated segment boundaries. Annotations were made by the authors who analyzed the songs and manually annotated the segment boundaries with an accuracy of ±100 ms.

To analyze the performance of algorithms on our folk music dataset, we gathered several publicly available implementation scenarios of segmentation algorithms: Segmentino [27], MSAF-Foote [4], MSAF-SCluster [16], MSAF-SF [13], MSAF-CNMF3 [11], and MSAF-SI-PLCA [28]. All except Segmentino are available within the Music Structure Analysis Framework (MSAF) and they were tested with three different feature types: MFCCs, HPCP chromagrams, and tonal centroid features, Tonnetz. Figure 2 shows the comparison of the methods according to $F1$ measure, with indicated confidence intervals. An estimated segment boundary was taken as correct (true positive) if it was located within a ±3-second window around an annotated boundary (the same window size is also used in MIREX evaluations). We analyze results of several best performing methods in the following subsections, which are also presented in Table 2 in the Evaluation. More detailed results as well as the data used in the research can be obtained at http://lgm.fri.uni-lj.si/ciril/mpie-bohak-2016/.

2.1. Segmentino. Segmentino [27] is a segmentation algorithm that was originally used for improving the results of a transcription algorithm. It is based on finding paths in a self-similarity matrix calculated from beat-synchronous chroma features. Its performance varies a lot with regard to the type of ensemble. It outperforms all other approaches for instrumental music with mean $F1$ measure of 0.61, with higher recall and lower precision (oversegmentation). On the other hand, for solo singing, it is the worst performer with $F1$ score of 0.3, with high precision but very low recall (undersegmentation).

The main reason for such behavior is that the method uses beat-synchronous features to calculate the self-similarity matrix. Beat-synchronous features are a standard approach to amend tempo variations in a song; however, they are based on the assumption that the beat can be reliably estimated. While this is easier for instrumental music and thus the method's performance is good for this type of recording, beat tracking is very difficult for singing, where there are no strong onsets, pauses between stanzas may be long, and tempo may vary a lot. If beat tracking fails, segmentation will also be poor. In addition, the method uses several fixed constraints when searching for repetitions (e.g., beginnings can start at multiples of four beats; repetitions can be of certain lengths), which do not hold for folk music in the same manner as for popular music. It also does not tolerate pitch drifting, which is common in singing.

2.2. MSAF-MFCC-Foote. Foote [4] presented one of the first methods for music segmentation, which is surprisingly the best performing method on our collection. It is based on a novelty measure calculated from a self-similarity matrix of

Table 2: Evaluation results, where precision (P), recall (R), and $F1$ measure ($F1$) are calculated as an average value of measures calculated per individual song.

Approach		Segmentino	MSAF-MFCC-Foote	MSAF-HPCP-SCluster	MSAF-MFCC-SF	The proposed method
Solo (OGL)	P	0.86	0.38	0.41	0.36	0.84
	R	0.17	0.76	0.51	0.42	0.85
	F	0.25	0.51	0.46	0.39	0.85
Two-Three	P	0.82	0.48	0.44	0.52	0.84
	R	0.27	0.85	0.5	0.6	0.89
	F	0.33	0.61	0.47	0.55	0.84
Instrumental	P	0.55	0.31	0.38	0.33	0.69
	R	0.82	0.97	0.87	0.8	0.63
	F	0.62	0.47	0.53	0.47	0.6
Instr. singing	P	0.55	0.37	0.35	0.42	0.74
	R	0.72	0.86	0.64	0.76	0.62
	F	0.59	0.52	0.46	0.54	0.61
Solo	P	0.92	0.46	0.45	0.46	0.87
	R	0.26	0.78	0.64	0.49	0.87
	F	0.36	0.57	0.53	0.48	0.86
Choir	P	0.74	0.31	0.4	0.4	0.73
	R	0.36	0.76	0.61	0.64	0.9
	F	0.41	0.44	0.48	0.5	0.78
Overall	P	0.74	0.39	0.41	0.41	78
	R	0.4	0.81	0.59	0.56	0.8
	F	0.4	0.52	0.48	0.47	0.76

MFCC features. Its performance on different types of music is relatively constant, ranging from $F1$ measure of 0.44 for choirs to 0.61 on two- to three-voice ensembles. It tends to oversegment all types of music.

The main reason for incorrectly placed boundaries is the fact that MFCCs do not represent harmonic properties but rather timbral properties of sound. As timbre may not change significantly in folk music, this leads to high self-similarity values throughout a song, making the discovery of segment boundaries difficult. Emphasis is also given on distinctions between vocal sounds (e.g., between "A" and "O"), which are not relevant as lyrics are mostly not repeated in folk songs. In addition, the method is very sensitive to the size of the kernel used for calculating the novelty measure, which is difficult to estimate.

2.3. MSAF-HPCP-SCluster. The method [16] uses techniques from spectral graph theory to make hierarchical segmentation based on a recurrence matrix calculated from MFCC and HPCP features. It yields very stable results with $F1$ measure around 0.5 for all ensemble types. It has balanced precision and recall on noninstrumental recordings, but high oversegmentation on instrumental music.

We can give three main reasons for the method's performance: (1) it tolerates only moderate tempo changes, so larger changes may interfere with spectral clustering and thus resulting segmentation, (2) it does not tolerate pitch drifting, and (3) high oversegmentation of instrumental recordings may indicate either improper balancing of local and global connectivity within the segmentation graph or improper selection of final boundaries from hierarchical segmentation.

2.4. MSAF-MFCC-SF. Serrà's approach for segmenting the time series [13] is based on segment features calculated from a filtered time-lag matrix. The method performs comparably to the MSAF-HPCP-SCluster method with an overall $F1$ measure of 0.47. Over- or undersegmentation does not occur much on noninstrumental recordings but is more prominent on instrumental music.

The method is designed to cope with small tempo deviations but does not perform well with larger deviations found mostly in sung materials. This affects two steps of the method: first, the use of blocks of features (delay coordinates) for calculating the recurrence plots yields poor similarity estimates and thus poor segmentation, as tempo may be quite different in repeated segments. Also, the calculation of similarity features is affected, so features and consequently the novelty curve will be smeared in time. The method also does not cope with pitch drifting, as it uses an Euclidean norm to calculate the recurrence plot.

2.5. Discussion. By analyzing the 10% of songs where methods yielded the worst performance, we found that not many songs were common to all methods. For methods that use MFCCs (MSAF-MFCC-Foote and MSAF-MFCC-SF), there are 4 common songs in the bottom 10% (20 songs). There are only 3 mutual songs for MSAF-MFCC-Foote and

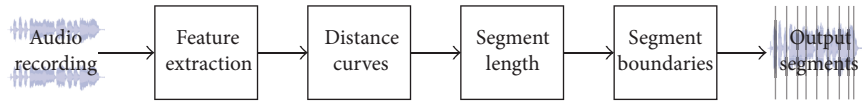


Figure 3: Outline of the proposed method.

MSAF-HPCP-SCluster combination and 6 songs for MSAF-MFCC-SF and MSAF-HPCP-SCluster combination. All three methods only have a single common song in the bottom 10%. This shows that the reason why methods perform poorly is not due to several "difficult" songs but that each method fails for different reasons presented in the previous sections.

Based on shortcomings of existing methods, we decided to consider the following folk music specifics when designing our segmentation method: (1) tolerance to tempo deviations in calculation of between-segment similarities, (2) tolerance to pitch drifting in calculation of between-segment similarities, (3) tolerance to noise and performer errors that may occur at different locations in a song, (4) songs that are structured as repetitions of one melodic or harmonic pattern, and (5) focus on segmentation of noninstrumental music, which represents a greater challenge for current methods than instrumental recordings.

3. The Proposed Method

Our proposed method processes an input audio recording in several phases and calculates a list of segment boundaries as a result. The process is illustrated in Figure 3. The input audio signal is first converted to a chroma representation, which is subsequently used to calculate tempo and pitch drift invariant distance curves indicating distances between different parts of the signal. Because we assume that songs contain repetitions of one pattern, the length of a typical segment is then calculated. Finally, a probabilistic approach is taken to calculate segment boundaries.

3.1. Feature Extraction. An input audio recording is first averaged to a single channel and normalized. To represent the content of the audio signal, we use harmonic chroma features, as they best capture melodic and harmonic between-segment similarities. Chroma features come in several flavors, and after some experiments, we decided to use HPCP features calculated with the Essentia library [29]. Similar to the findings of Serrà et al. [30], we found that increased resolution is beneficial for segmentation, so we used a 24-bin representation (2 bins per semitone) instead of the more standard 12 chroma bins (one per semitone).

3.2. Finding Similarities. Choosing an appropriate similarity measure is the key to successful segmentation. Most current approaches use local similarity measures that compare short-term features such as chromas or MFCCs across the signal, resulting in a self-similarity matrix or recurrence plot. An exception is Serrà's approach [13], who compared sequences of features, so each similarity value already represented the similarity of two (short) sequences of features. In these approaches, variations in tempo had to be processed in subsequent segmentation phases; however as our results show, large tempo variations are not handled well.

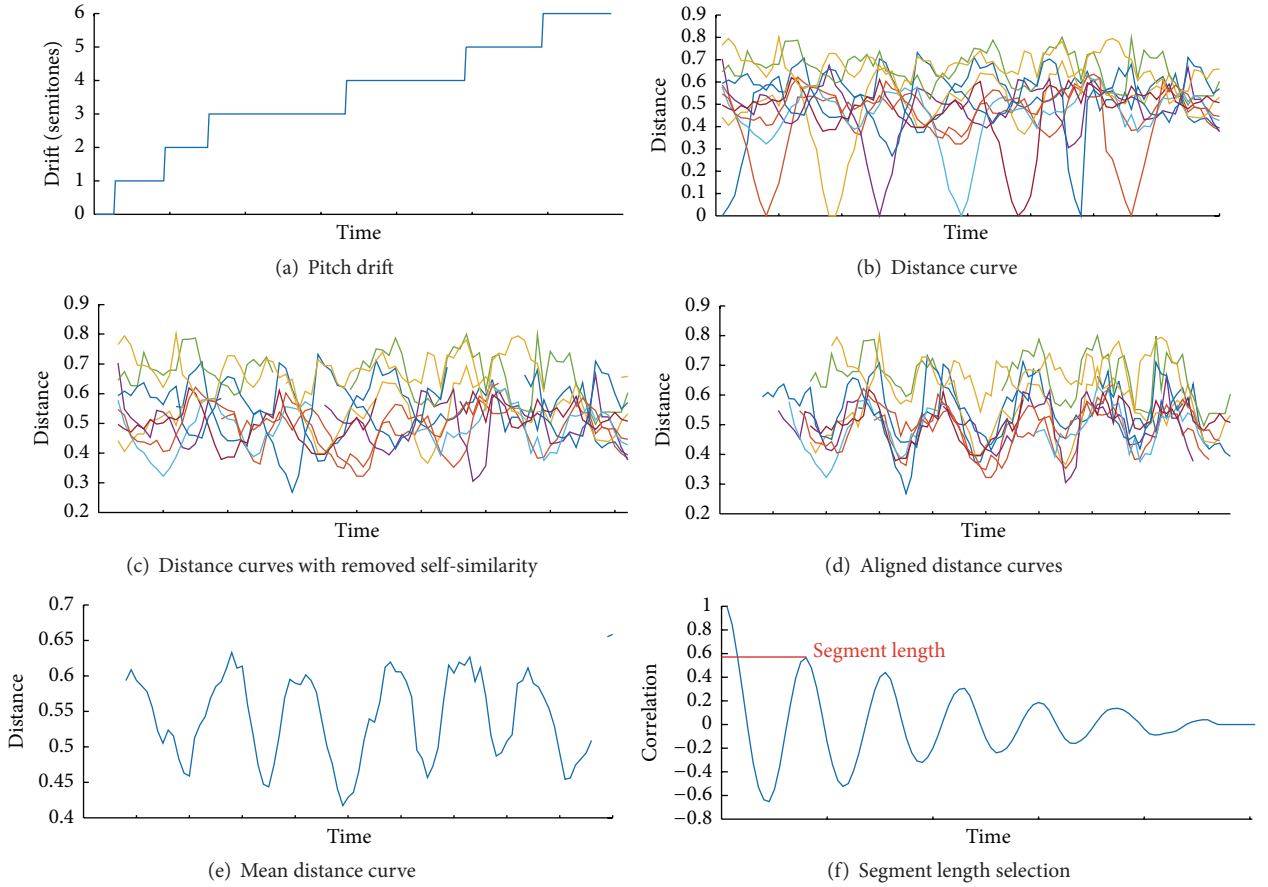
We therefore decided to consider tempo variations already in calculation of similarity values. The idea was taken from Müller et al. [19], who proposed the use of dynamic time warping (DTW) to calculate similarity between two song parts. DTW is a standard technique for measuring similarity between two temporal sequences which may vary in time or speed.

We denote $X_1^{(N)}$ to be a sequence of N HPCP feature vectors $\mathbf{x}_i$, starting at time 1, representing the entire analyzed audio signal. We also define a local cost measure $c(\mathbf{x}_i, \mathbf{x}_j)$ as the distance between feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$. We use the correlation distance (one minus correlation) as our local cost measure, as it compares favorably to other distance functions (the same was also demonstrated by Serrà et al. [30]). To calculate the distance d_{ij} between two sequences of feature vectors $X_i^{(L)}$ and $X_j^{(L)}$, we calculate the optimal time warping path between the two sequences by minimizing the total cost as

$$d_{ij} = \min \frac{1}{|w|} \sum_{(m,n)\in w} c\left(\mathbf{x}_m, \mathbf{x}_n\right), \tag{1}$$

where w represents a time warping path starting at the beginning of both sequences (i, j) and ending at $(i+L-1, j+L-1)$. The problem can be solved with dynamic programming in $O(n^2)$.

The *distance curve* $D_i = (d_{i1}, d_{i2}, \ldots, d_{iN})$ represents the distance of a song segment $X_i^{(L)}$ to all other segments of length L in a song. Ideally, if i represented the beginning of a segment, and L the length of a segment, dips in D_i would reflect all repetitions of the segment in a song (tempo variations are already considered in distance calculation). However, several problems still remain. First, we have no information on location of segment boundaries to set the value of i. It would be tempting to set i to the beginning of the song (as we did in [21]); however, the first stanza often contains mistakes, speech and other artifacts, when singers forget how a song should be performed. In these cases, D_i would be noisy and segmentation poor. We solve this problem by calculating several distance curves on randomly chosen locations $T = (t_1, t_2, \ldots, t_M)$ distributed over the entire length of a song (see Figure 4(b)). The rationale is that we will thus increase the probability of choosing several song locations t_i which are repeated in a song and will yield valid distance curves D_{t_i}. The number of locations M can be arbitrary; however, too frequent sampling only increases computational costs and does not improve the accuracy of segmentation. In practice, we decided to calculate approximately two curves per segment, so the locations T are set to be approximately 10 seconds apart (average segment length in our collection is 20 seconds). We



(a) Pitch drift
(b) Distance curve
(c) Distance curves with removed self-similarity
(d) Aligned distance curves
(e) Mean distance curve
(f) Segment length selection

FIGURE 4: Segmentation steps: (a) cumulative pitch drift, (b) distance curves, (c) distance curves with self-similarity removed, (d) aligned distance curves, (e) mean distance curve, and (f) autocorrelation with estimated segment length.

also set the DTW length L to 10 seconds, which is long enough to yield meaningful DTW distances.

Finally, we need to consider pitch drifting, which often occurs when intonation of performers changes upwards or downwards over the course of a song. Not taking drifting into consideration would lead to inaccurate distance curves and thus poor segmentation. We solve this problem by calculating a series of distance curves D_i^p, which represent the distance between the song and segment $X_i^{(L)}$ circularly shifted by p bins. As chroma features are octave invariant, their circular shift by p bins is akin to a pitch change of p bins. A similar approach was already introduced in [19]; however, calculation of similarity in [19] involves the selection of optimal shift for every DTW calculation d_{ij}. Such approach is not completely realistic, as it does not consider the notion that intonation does not change rapidly over time, but rather gradually.

Instead, we propose to estimate the pitch drift (represented as a sequence of shift values $(p_1, p_2, \ldots, p_N)$) by minimizing a cost function that balances between unconstrained minimization of distances (as in [19]) and gradual drifting through time:

$$\min_{p_j \in [-\zeta \cdots \zeta]} \quad \sum_{j=1}^{N} d_{ij}^{p_j} + C_p \delta\left(p_j - p_{j-1}\right), \tag{2}$$

where ζ is the maximum allowed drift and C_p is the cost of pitch change. Optimization can be solved with dynamic programming and results in a set of pitch drift values which change gradually over time. An example of upward drift is shown in Figure 4(a).

The described procedure results in a set of distance curves $D_i' = (d_{i1}^{p_1}, d_{i2}^{p_2}, \ldots, d_{iN}^{p_N})$, $i \in T$, describing distances between segments starting at times T and the whole song, considering variances in tempo and intonation. Such set of curves is shown in Figure 4(b). In the final step, the self-similar parts of the curves are removed, as they carry no useful information for further segmentation; the result is shown in Figure 4(c). Self-similar parts of the curves are visible as large dips in distance curves in Figure 4(b), where we compare the segment to itself. We remove these dips and in further calculations replace them with the average curve value.

3.3. Calculating the Segment Length. We assume that a song consists of several repetitions of a segment. We can therefore estimate the length of a typical segment, which enables faster and more accurate search for segment boundaries. First, we align all of the distance curves to a reference curve. Alignment is needed, because curves were calculated by comparing the entire song to segments at random time locations, T, and are thus not aligned (this can also be observed in Figure 4(c)).

The reference curve is chosen as the distance curve that has the highest correlation to all other curves. The rationale is that the segment this curve was calculated from is very representative of the entire song; otherwise, its curve would be poorly correlated to others. Alignment is performed by time-shifting each curve according to the distance of its closest valley (representing a similar segment) to the segment time the reference curve was calculated from. An example of alignment can be seen in Figure 4(d).

After the curves are aligned, we calculate the average distance curve D_a by averaging all aligned curves (Figure 4(e)). Autocorrelation of the average distance curve yields segment periodicities. We choose the highest autocorrelation peak as the length of a typical segment l (see Figure 4(f)).

3.4. Segmentation. We cast the segmentation problem into a probabilistic framework similar to hidden Markov models. We first define a set of states $\{s_i,\ i = 1 \cdots N\}$, which correspond to all time instances in the signal and thus represent the set of all possible segment boundaries. We search for an optimal sequence of states $S_{1:Q}$ given state probabilities $P(S_t)$ and transition probabilities $P(S_t \mid S_{t-1})$, maximizing

$$P\left(S_{1:Q}\right) = P\left(S_1\right) \prod_{t=2}^{Q} P\left(S_t \mid S_{t-1}\right) P\left(S_t\right). \tag{3}$$

State probabilities are calculated from the audio signal and are proportional to lengths of low-magnitude regions in the signal, while transition probabilities reflect the estimated segment lengths and between-segment similarities. Both calculations are explained in more detail in the following subsections.

3.4.1. Calculating State Probabilities. State probabilities $P(S_t = s_i)$ are proportional to the likelihood of placing a segment boundary at time i. We follow the rationale that this likelihood is larger if the boundary is preceded by a low-amplitude signal region: for singing, this often corresponds to breathing pauses before stanza beginnings, while for instrumental parts, this may also correspond to phrase endings. The longer the low-amplitude region, the higher the probability of a segment boundary. To calculate state probabilities, we first calculate the amplitude envelope of the music signal A_e. We then low-pass-filter A_e to obtain an adaptive estimate of the average signal amplitude over time (A_a). Setting a threshold A_t that will define the regions of low amplitude ($A_e < A_a + A_t$) is not trivial, as recordings may contain high levels of background noise. We therefore set this threshold adaptively, as the minimal value larger than or equal to −10 dB where low-amplitude regions will cover at least 10% of the entire signal length. These values were chosen so that they prevent sparse distribution of nonzero state probabilities, which would lead to undersegmentation. Finally, the state probability s_i is set to be proportional to the size of the low-amplitude region immediately preceding i, if such region exists and the region ends at i. To reduce the effect of very long regions, we place an upper bound on this value at 1 second, so regions larger than that have the same effect as a 1-second region. If no such region exists, the state probability is set to a constant value τ. Probability of the first state (which has no prior regions) is set to 0.5 and the last state to 1, as we want the segmentation to end at the last state. The process is illustrated in Figure 5, which shows the amplitude envelope of a song, its average, the threshold, and calculated state probabilities.

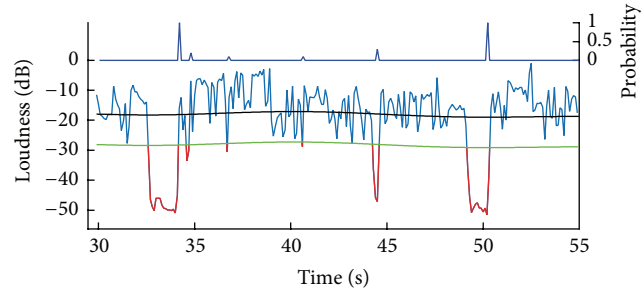


FIGURE 5: The figure shows the amplitude envelope of song A_e (light blue with red sections under the threshold), average amplitude A_a (black), selected threshold $A_a + A_t$ (green), and state probabilities (dark blue with scale on the right side).

3.4.2. Calculating Transition Probabilities. Transition probability $P(S_t = s_i \mid S_{t-1} = s_j)$ is the probability of placing a segment boundary at time i if the previous was located at time j. We incorporate three constraints into calculation of transition probabilities: (a) two segments beginning at i and j should be similar (s_i is a repetition of s_j), (b) the segments should be separated by approximately the estimated segment length l, and (c) only forward transitions are allowed:

$$P\left(S_t = s_i \mid S_{t-1} = s_j\right) \propto \operatorname{sim}\left(s_j, s_i\right) \mathcal{N}\left(l, \sigma\right),$$
$$P\left(S_t = s_i \mid S_{t-1} = s_j\right) = 0, \quad \text{if } i \leq j. \tag{4}$$

$\mathcal{N}$ is the normal distribution with mean l and standard deviation σ that models the expected segment duration and sim is a similarity function. The similarity function is calculated from the average distance curve D_a, which is inverted (subtracted from 1) and scaled to fit the $[0, 1]$ interval. As the resulting curve has no absolute time, similarity between s_j and s_i is obtained by finding a peak nearest to j and looking up the value at $(i - j)$ offset from the peak. Peaks represent repetitions, so we thus model the similarity of the segment starting at j to the segment at i.

We can use the Viterbi algorithm to find the optimal sequence of states, whereby we allow the starting state to occur within the first η seconds, and enforce the ending in the last state. As states are directly mapped to time, the resulting sequence of states represents the set of found segment boundaries.

4. Evaluation

We evaluated our segmentation algorithm on the folk music collection described in Section 2. Table 2 shows average

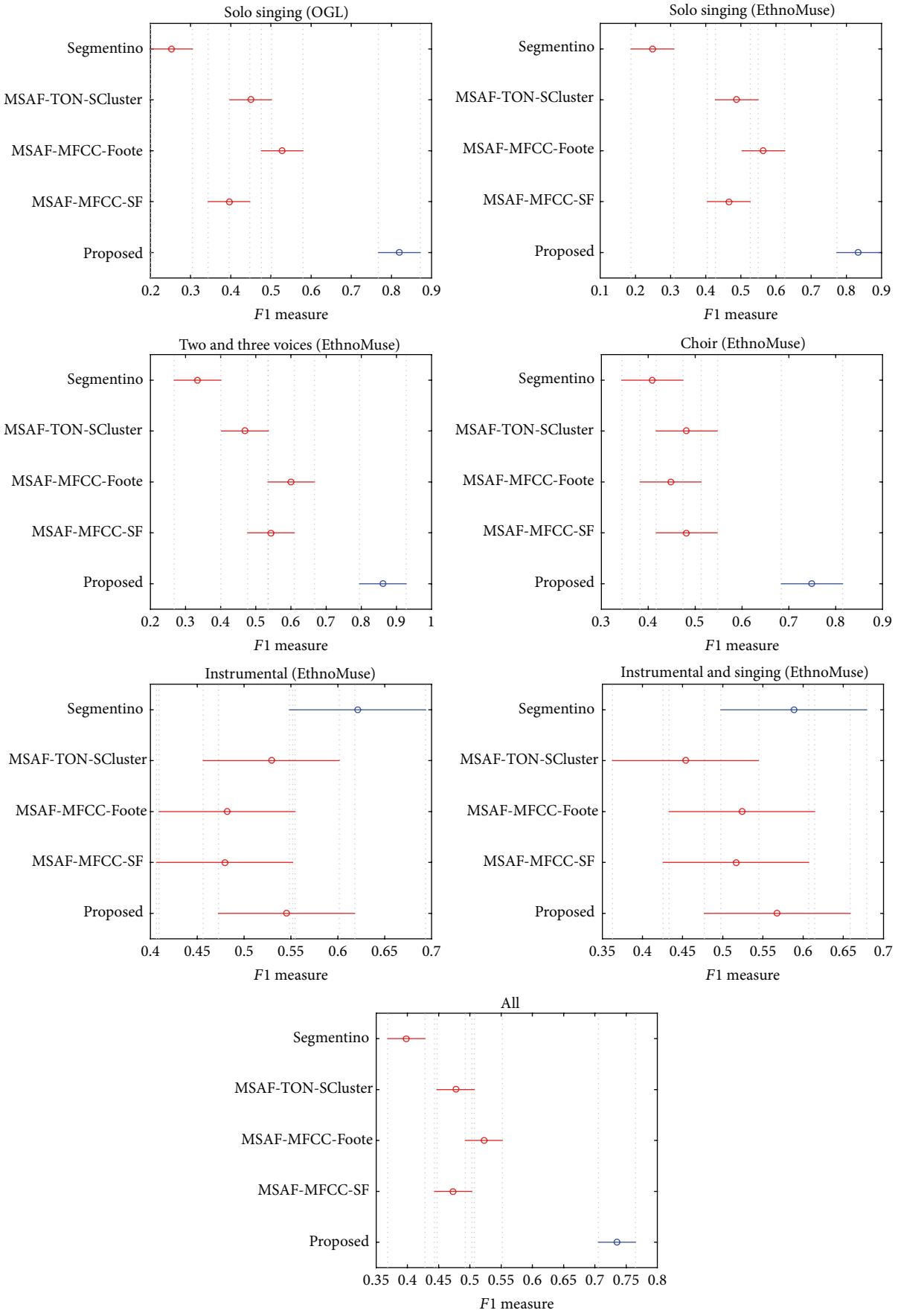


Figure 6: The figure shows statistically significant differences between methods on individual ensemble types, as well as overall values.

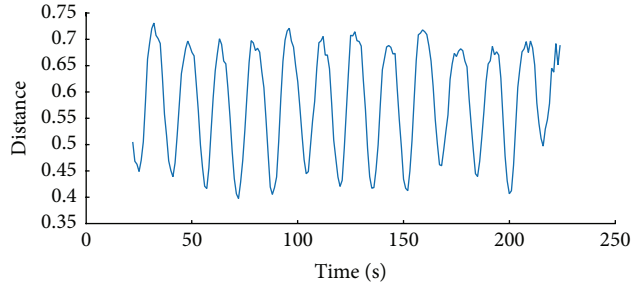
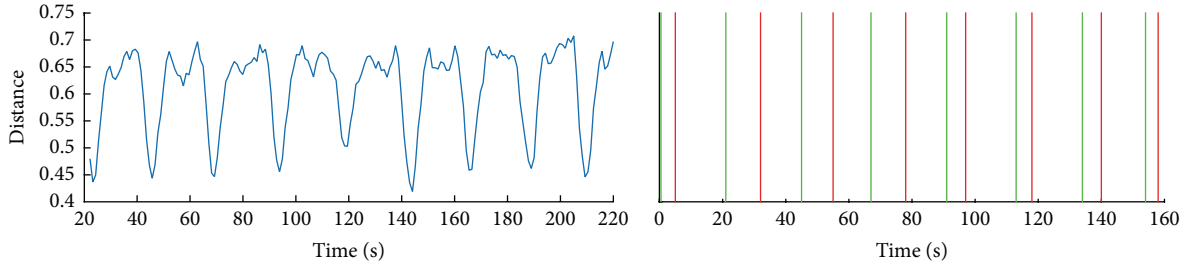


Figure 7: A distance curve, where repetitions are clearly visible as valleys.



(a) Mean distance curve

(b) Ground truth (green) versus calculated beginnings

Figure 8: An example of phase shifting: (a) the mean distance curve and (b) estimated boundaries (red) versus ground truth (green).

precision, recall, and $F1$ measure per song for each ensemble type, as well as for the entire collection. An estimated segment boundary was counted as correct (true positive) if it was located within a ±3-second window around an annotated boundary (the same window size is also used in MIREX evaluations). For comparison, results obtained by methods described in Section 2 are also shown.

For evaluation of the proposed approach, the following parameter values were used: the allowed pitch drift ζ was set to 2 semitones, the minimal state probability $\tau = 0.01$, the width of the normal distribution penalizing for deviations from expected segment duration $\sigma = l/4$, and the allowed set of initial states $\eta = 6$ seconds. The values were chosen according to experience and were not optimized specifically for the collection. Additional tests showed that the method is not very sensitive to changes in these values.

Overall, the proposed approach significantly outperforms others for noninstrumental recordings, while for instrumental, its performance is comparable to the best performer, Segmentino (see Figure 6). As we focused the design of our method on noninstrumental recordings, such results were expected. Our approach also yields the most balanced results in terms of precision and recall, which means that it does not significantly over- or undersegment the recordings.

The results of our method are comparable with results of current state-of-the-art methods for folk music segmentation. We can compare our method to results presented by Müller et al. [23], which tested their segmentation approach on the corpus *Onder de Groene Linde* (Solo OGL), also made available to us by the authors. Their results are slightly better ($F1$ measure of 0.872 for Solo OGL); however, we should note that the method is based on $F0$-enhanced CENS features, specifically tuned for solo singing, so we cannot estimate how it would perform for other ensemble types.

Analysis of results for each song revealed that perfect precision and recall (1.0) were obtained for 86 songs, while segmentation for 6 songs was completely false (0.0). When the mean distance curve is correctly estimated, the segmentation very likely succeeds. An example is shown in Figure 7, where repetitions are clearly visible as valleys in the distance curve.

Analysis of errors revealed three major causes. In several cases, the distance curve and segment length are correctly estimated; however, the final segmentation produces segment boundaries which are out of phase (time-shifted) with the actual beginnings. This type of error was already pointed out by Müller et al. [26] and may occur when performers make mistakes at beginnings of one or several stanzas. This can result in segment boundaries that are placed at points where the performance runs smoothly in all stanzas. An example is shown in Figure 8 where in (a) the mean distance curve does not reveal any peculiarities; however, as shown in (b), the found boundaries (red) are shifted in comparison with the ground truth (green).

Correct segmentation is also dependent on the correctly estimated segment length. In several cases, incorrect length was chosen. When stanzas consist of two similar parts, the estimated length (based on autocorrelation of the distance curve) may be half the true length, which results in oversegmentation by a factor of 2. Additionally, when similarities within individual segments are high (mostly in instrumental recordings), the distance curve does not capture the

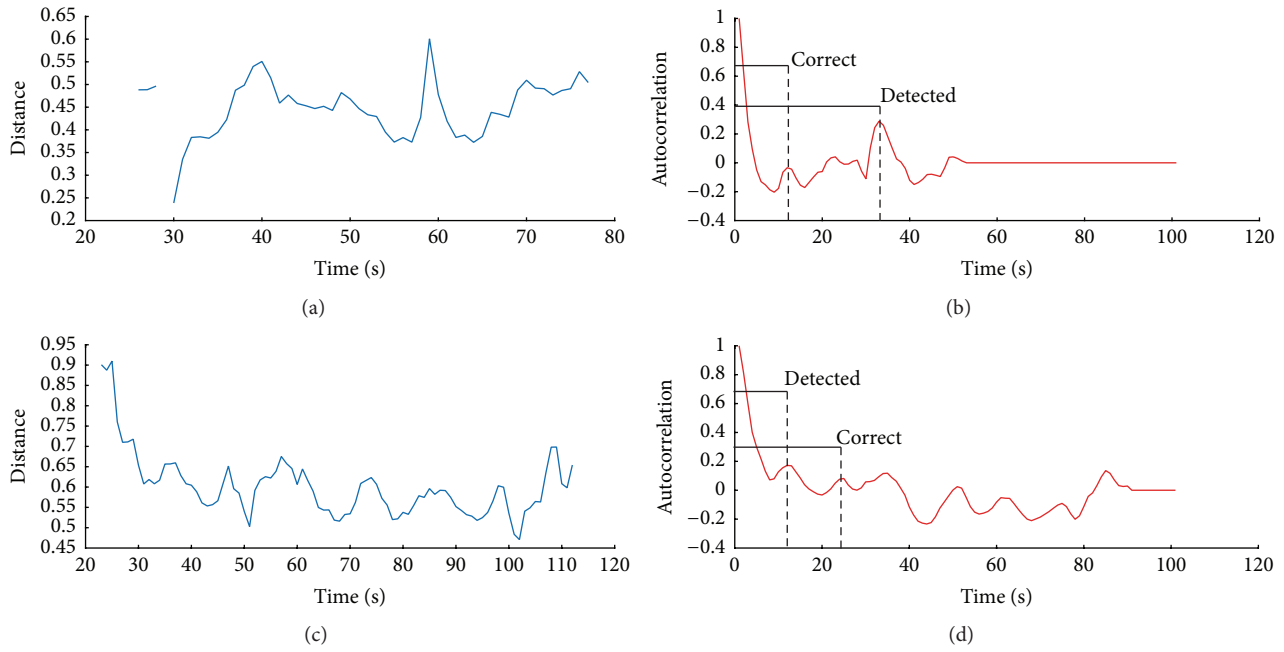


FIGURE 9: Figures display two examples where our method fails. The mean distance curve does not have prominent dips (a) and consequently autocorrelation fails to find the correct segment length (b). There are too many prominent dips in mean distance curve of the second example (c) and consequently autocorrelation finds segment length, which is too short, resulting in 2-time oversegmentation (d).

repetitions well, which results in an incorrectly estimated segment length and poor transition probabilities and final segmentation. Two such examples are shown in Figure 9.

5. Conclusions

Due to rapid digitization of folk music collections, development of dedicated methods for extraction of high-level descriptors from such recordings is needed. Our paper introduces a novel method for segmentation of folk music field recordings. We first analyze why state-of-the-art approaches for segmentation of commercial recordings fail for folk music and then outline a segmentation method which incorporates mechanisms for coping with specifics of folk music. We show that the method significantly outperforms current state-of-the-art approaches and performs at least as well as a state-of-the-art method for segmentation of solo singing folk music.

Future work will be dedicated to improvement of the method, especially for segmentation of instrumental music. We also plan to enlarge the evaluation database and further specialize the method for individual ensemble types, by first automatically determining the ensemble type and then choosing the method parameters accordingly. We also plan to extend the method for discovery of hierarchical music structure.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work would not have been done without field recordings provided by the Institute of Ethnomusicology at Research Centre of Slovenian Academy of Sciences and Arts.

References

[1] J. Paulus, M. Müller, and A. Klapuri, "Audio-based music structure analysis," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR '10)*, pp. 625–636, Utrecht, Netherlands, 2010.

[2] J. S. Downie, "The music information retrieval evaluation exchange (2005–2007): a window into music information retrieval research," *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247–255, 2008.

[3] J. Foote, "Visualizing music and audio using self-similarity," in *Proceedings of the 7th ACM International Conference on Multimedia (Part 1) (MULTIMEDIA '99)*, pp. 77–80, ACM, Orlando, Fla, USA, October-November 1999.

[4] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '00)*, vol. 1, pp. 452–455, IEEE, New York, NY, USA, July-August 2000, Latest Advances in the Fast Changing World of Multimedia (Cat. no. 00TH8532).

[5] M. A. Bartsch and G. H. Wakefield, "To catch a chorus: using chroma-based representations for audio thumbnailing," in *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 15–18, IEEE, New Platz, NY, USA, October 2001.

[6] A. Eronen, "Chorus detection with combined use of MFCC and chroma features and image processing filters," in *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx '07)*, pp. 229–236, September 2007.

[7] K. Jensen, "Multiple scale music segmentation using rhythm, timbre, and harmony," *Eurasip Journal on Advances in Signal Processing*, vol. 2007, Article ID 073205, 2006.

[8] 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*, vol. 14, no. 5, pp. 1783–1794, 2006.

[9] F. Kaiser and T. Sikora, "Music structure discovery in popular music using non-negative matrix factorization ," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR '10)*, pp. 429–434, 2010.

[10] R. J. Weiss and J. P. Bello, "Identifying repeated patterns in music using sparse convolutive non-negative matrix factorization," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR '10)*, pp. 123–128, August 2010.

[11] O. Nieto and T. Jehan, "Convex non-negative matrix factorization for automatic music structure identification," in *Proceedings of the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '13)*, pp. 236–240, Vancouver, Canada, May 2013.

[12] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "$l1$ -graph based music structure analysis," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR '11)*, pp. 495–500, Miami, Fla, USA, October 2011.

[13] J. Serrà, M. Müller, P. Grosche, and J. L. Arcos, "Unsupervised detection of music boundaries by time series structure features," in *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pp. 1613–1619, AAAI Press, Toronto, Canada, July 2012.

[14] J. Serrà, M. Müller, P. Grosche, and J. L. Arcos, "Unsupervised music structure annotation by time series structure features and segment similarity," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1229–1240, 2014.

[15] G. Peeters and V. Bisot, "Improving music structure segmentation using lag-priors," in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR '14)*, pp. 337–342, Taipei, Taiwan, October 2014.

[16] B. McFee and D. P. W. Ellis, "Analyzing song structure with spectral clustering," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR '14)*, pp. 405–410, Taipei, Taiwan, October 2014.

[17] B. McFee and D. P. W. Ellis, "Learning to segment songs with ordinal linear discriminant analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '14)*, pp. 5197–5201, IEEE, Florence,Italy, May 2014.

[18] O. Nieto and J. P. Bello, "Music segment similarity using 2D-fourier magnitude coefficients," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '14)*, pp. 664–668, IEEE, Florence,Italy, May 2014.

[19] M. Müller, P. Grosche, and F. Wiering, "Robust segmentation and annotation of folk song recordings," in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR '09)*, pp. 735–740, Kobe, Japan, October 2009.

[20] M. Müller, P. Grosche, and N. Jiang, "A segment-based fitness measure for capturing repetitive structures of music recordings," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR '11)*, pp. 615–620, Miami, Fla, USA, October 2011.

[21] C. Bohak and M. Marolt, "Finding repeating stanzas in folk songs," in *Proceedings of the International Conference on Music Information Re-Trieval (ISMIR '12)*, pp. 451–456, Porto, Portugal, 2012.

[22] M. Müller and P. Grosche, "Automated segmentation of folk song field recordings," in *Proceedings of the ITG Conference on Speech Communication*, pp. 1–4, VDE, Braunschweig, Germany, September 2012.

[23] M. Müller, N. Jiang, and P. Grosche, "A Robust fitness measure for capturing repetitions in music recordings with applications to audio thumbnailing," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 21, no. 3, pp. 531–543, 2013.

[24] M. Marolt, "Probabilistic segmentation and labeling of ethnomusicological field recordings," in *Proceedings of the 10th International Society for Music Information Retrieval Conference*, pp. 75–80, Kobe, Japan, October 2009.

[25] G. Strle and M. Marolt, "The EthnoMuse digital library: conceptual representation and annotation of ethnomusicological materials," *International Journal on Digital Libraries*, vol. 12, no. 2-3, pp. 105–119, 2012.

[26] M. Müller, P. Grosche, and F. Wiering, "Automated analysis of performance variations in folk song recordings," in *Proceedings of the International Conference on Multimedia Information Retrieval (ISMIR '10)*, pp. 247–256, Philadelphia, Pa, USA, March 2010.

[27] M. Mauch, K. C. Noland, and S. Dixon, "Using musical structure to enhance automatic chord transcription," in *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR '09)*, pp. 231–236, Kobe, Japan, October 2009.

[28] R. J. Weiss and J. P. Bello, "Unsupervised discovery of temporal structure in music," *IEEE Journal on Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1240–1251, 2011.

[29] D. Bogdanov, N. Wack, E. Gómez et al., "ESSENTIA: an audio analysis library for music information retrieval," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR '13)*, pp. 493–498, Curitiba, Brazil, November 2013.

[30] J. Serrà, E. Gómez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 6, pp. 1138–1151, 2008.