Automated Segmentation of Folk Song Field Recordings

Meinard Müller, Peter Grosche[*]

Saarland University and MPI Informatik, 66123 Saarbrücken, Germany
Email: {meinard,pgrosche}@mpi-inf.mpg.de
Web: www.mpi-inf.mpg.de/~mmueller

Abstract

In this paper, we introduce an automated procedure for segmenting a given folk song field recording into its constituent stanzas. One challenge arises from the fact that these recordings are performed by elderly non-professional singers under poor recording conditions such that the constituent stanzas may reveal significant temporal and spectral deviations. Unlike a previously described segmentation approach that relies on a manually transcribed reference stanza, we introduce a reference-free segmentation procedure, which is driven by an audio thumbnailing procedure in combination with enhanced similarity matrices. Our experiments on a Dutch folk song collection show that our segmentation results are comparable to the ones obtained by the reference-based method.

1 Introduction

Generally, a folk song is referred to as a song that was sung by the common people of a region or culture during work or social activities. As a result, folk music is closely related to the musical culture of a specific region. The understanding of the genetic relation between folk songs and entire folk song families has been the subject of various research efforts, see, e. g., [1, 2]. Folk songs have been mainly passed down by oral tradition and significant efforts have been made to preserve the cultural heritage by assembling collections of folk song field recordings. Current folk song research is mainly based on score-like symbolic representations, which have been obtained by manually transcribing a representative stanza of the recorded tunes, see Figure 1a. The original audio material, however, is rarely considered, even though it may contain valuable information not reflected by the idealized transcripts.

One reason for folk song researchers to focus on symbolic representations is that, due to its massive data volume and complexity, audio material is generally hard to deal with. Therefore, computer-assisted analysis tools and intuitive user interfaces are needed to make the original audio material more accessible. In this paper, we address a first analysis problem with the objective to automatically segment a given folk song field recording into its constituent stanzas. More precisely, for most folk songs a tune is repeated over and over again with changing lyrics. A typical field recording therefore consists of a sequence $A_1 A_2 \ldots A_K$ of stanzas A_k, $k \in [1:K] := \{1,2,\ldots,K\}$, where each A_k corresponds to the same tune. Given a field recording, the segmentation task consists in identifying the temporal boundaries of the various stanzas. This task is more challenging than one may guess at first sight since the sung stanzas often reveal significant temporal and spectral deviations and deformations. For example, the elderly non-professional singers often hesitate while performing a stanza, deviate significantly from the expected pitches, and have serious problems with intonation. Furthermore, the field recordings are often of poor recording quality with significant background noise.

In [3], a reference-based segmentation algorithm is described that relies on the availability of a manually transcribed reference stanza. The segmentation is then achieved by locally comparing the field recording with the reference stanza. In this paper, we introduce a reference-free segmentation procedure that does not rely on any reference transcript. Our idea is to apply a recent audio thumbnailing approach described in [4] to identify the most "repetitive" segment in a given recording. This so-called *thumbnail* then takes over the role of the reference. The thumbnailing procedure is built upon suitable audio features and self-similarity matrices (SSM). To cope with the above mentioned variations, we introduce various enhancement strategies to absorb a high-degree of these deviations and deformations already on the feature and SSM level. We report on various experiments using challenging field recordings taken from the collection *Onder de groene linde* (OGL), which is part of the *Nederlandse Liederenbank*.[1] Our evaluation shows that the segmentation results of our approach are comparable to the ones obtained from the reference-based segmentation procedure [3].

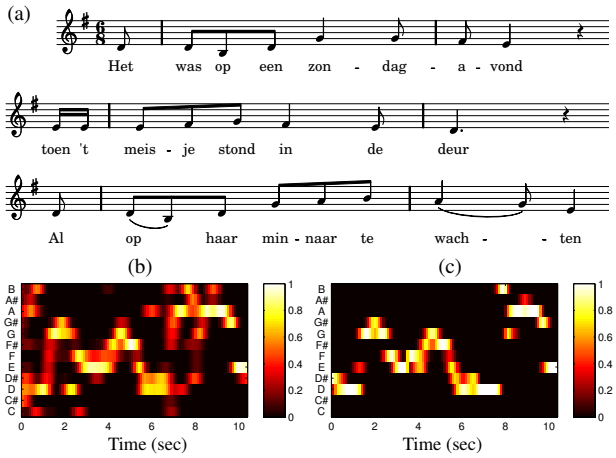

Figure 1: Representations of the beginning of the first stanza of OGL27517. **(a)** Score representation (manual transcript). **(b)** CENS features of the audio recording. **(c)** F0-enhanced CENS features.

The remainder of this paper is organized as follows. We first describe the underlying audio features (Section 2) and then the SSMs along with various enhancement strategies (Section 3). After summarizing the audio thumbnailing procedure (Section 4), we report on our segmentation experiments (Section 5) and give an outlook on future work (Section 6).

2 Audio Features

Chroma-based audio features have become a widely used tool in processing and analyzing music data [5–11]. Assuming the equal-tempered scale, the term chroma refers to the elements of the set $\{C, C^\sharp, D, \ldots, B\}$ that consists of the twelve pitch spelling attributes as used in Western music notation. The resulting 12-dimensional chroma vectors express how the short-time energy of the audio signal is distributed over the twelve chroma bands [5]. There are many ways of computing chroma features, which results in a large number of chroma variants with different properties [5, 7, 8]. Following [5], we employ a multirate filter bank technique to first decompose the audio signal into pitch subbands that correspond to the semitones of the equal-tempered scale (intuitively, these semitones correspond to the keys of a standard piano). Then, adding up the pitch values that belong to the same chroma, one obtains a chroma representation (or chromagram),

[*]The authors have been supported by the Cluster of Excellence on Multimodal Computing and Interaction at Saarland University

[1] www.liederenbank.nl

see also Figure 1b. The resulting chroma features are further processed by applying suitable quantization, smoothing, downsampling, and normalization operations resulting in enhanced chroma features referred to as CENS (Chroma Energy Normalized Statistics) [5].[2] In the following, we use a feature resolution of 2 Hz (two feature per seconds).

Chroma features capture information that closely correlates to harmonic and melodic properties of the audio recording [5]. Since in our scenario we are dealing with monophonic music (one note at a time sung by a single voice), we exploit this fact by adding a fundamental frequency (F0) estimation and quantization step. More precisely, we estimate for each time frame the fundamental frequency as in [12] and then determine the pitch whose center frequency is closest to the fundamental frequency. Then, in the pitch decomposition, we assign energy only to the subband that corresponds to the determined pitch. Doing so, the features gain in robustness towards deviations caused by the singers' intonation problems, vibrato effects, as well as background noise, see also Figure 1c.

3 Self-Similarity Matrices

Let $X := (x_1, x_2, \ldots, x_N)$ be the feature sequence consisting of the chroma features as described above. Furthermore, let $\mathbf{s}$ be a similarity measure that allows for comparing two chroma features.[3] Then one obtains an $N \times N$ *self-similarity matrix* (SSM) by comparing the elements of X in a pairwise fashion:

$$\mathbf{S}(n,m) := \mathbf{s}(x_n, x_m),$$

for $n, m \in [1:N]$. Introduced to the music context in [13], such matrices have turned out to be a powerful tool for revealing repeating patterns of X. In particular, each diagonal path (or stripe) of high similarity running in parallel to the main diagonal of $\mathbf{S}$ indicates the similarity of two audio segments (given by the projections of the path onto the vertical and horizontal axis, respectively), see [14].

For example, Figure 2a shows an SSM for the first eight stanzas $A_1A_2A_3A_4A_5A_6A_7A_8$ of the field recording OGL19101. The highlighted path encodes the similarity between A_2 and A_3. If the eight segments would be close to being exact repetitions, one would expect a "full" path structure as indicated by Figure 2f. However, due to the mentioned spectral and temporal deviations between the sung stanzas, the path structure is in general highly distorted and fragmentary. We now introduce various enhancement strategies to improve on the path structure of the SSM.

Temporal Smoothing

As first matrix enhancement strategy, we apply a smoothing filter along the direction of the main diagonal, which results in an emphasis of diagonal information in $\mathbf{S}$ and a denoising of other structures, see also [7, 9–11] for similar strategies. This form of filtering, however, typically assumes that the tempo across the subsequent stanzas is more or less constant. In the presence of significant tempo differences, however, simply smoothing along the main diagonal may smear out important structural information. To avoid this, we use a strategy that filters the SSM along various gradients as proposed in [9] covering tempo variations of roughly ± 30 percent. Obviously, choosing an appropriate value for the smoothing length parameter constitutes a trade-off between enhancement capability and level of detail. A suitable parameter depends on the kind of audio material.[4] See Figure 2c for an illustration and [9] for details.

[2] An implementation of these features is available at `www.mpi-inf.mpg.de/resources/MIR/chromatoolbox/`, see also [6].

[3] In the following, we simply use the cosine measure, i. e., the inner product between normalized chroma vectors. Since the chroma vectors only have positive entries, this yields a value between 0 and 1.

[4] In our folk song experiments, we use a smoothing length corresponding to 6 seconds. This also takes into account that the length of an individual stanza is above this value.

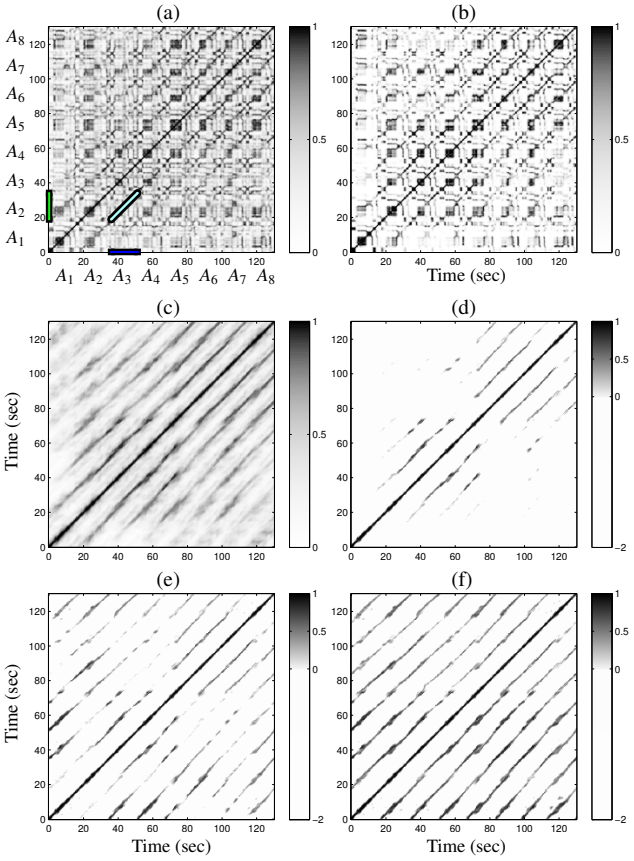

Figure 2: Self similarity matrices for the first eight stanzas of the folk song OGL19101. **(a)** SSM computed from CENS features. The highlighted path encodes the similarity of A_3 and A_2. **(b)** SSM computed from F0-enhanced CENS features. **(c)** Path-enhanced SSM. **(d)** Thresholded and normalized SSM $\mathbf{S}$.[6] **(e)** Transposition-invariant SSM $\mathbf{S}^{\text{trans}}$. **(f)** Fluctuation-invariant SSM $\mathbf{S}^{\text{fluc}}$.

Thresholding and Normalization

We further process the SSM by suppressing all values that fall below a given threshold. Using normalized chroma features and the cosine measure as similarity measure, all values of the SSM are between 0 and 1. Using a suitable threshold parameter $\tau > 0$ and a penalty parameter $\delta \leq 0$, we first set the score values of all cells with a score below τ to the value δ and then linearly scale the range $[\tau : 1]$ to $[0 : 1]$, see Figure 2d. The thresholding introduces some kind of denoising, whereas the parameter δ imposes some additional penalty on all cells of low score. Intuitively, we want to achieve that the relevant path structure lies in the positive part of the resulting SSM, whereas all other cells are given a negative score. Note that different methods can be used for thresholding such as using a predefined threshold or using a relative threshold to enforce a certain percentage of cells to have positive score [11].[7] Again we denote the resulting matrix simply by $\mathbf{S}$.

Transposition and Fluctuation Invariance

As mentioned above, the non-professional singers of the folk songs often deviate significantly from the expected pitches and

[6] For visualization purposes, to make the effects clearer, we have used an absolute threshold $\tau = 0.8$ in (d), (e), and (f).

[7] In our experiments, we choose the threshold in a relative fashion by keeping 40% of the cells having the highest score and set $\delta = -2$. These values were found experimentally. Slight changes of the parameters' values did not have a significant impact on the final segmentation results.

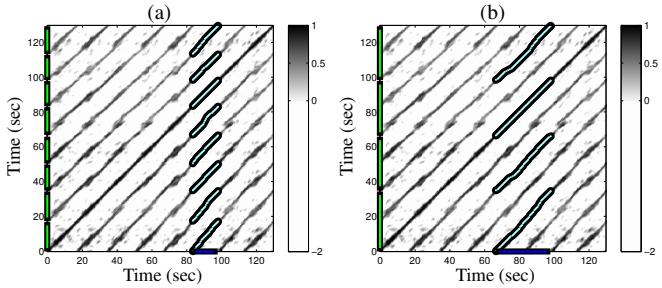

Figure 3: Path families and induced segment families for two different segments α for OGL19101. **(a)** $\alpha = [83:98]$ (thumbnail, maximal fitness, corresponding to stanza A_6). **(b)** $\alpha = [66:98]$ (corresponding to stanzas A_5A_6).

have serious problems with the intonation. Even worse, their voices often fluctuate by several semitones downwards or upwards across the various stanzas of the same recording. For example, in the case of the OGL19101 recording, the singer's voice constantly increases in pitch while performing the stanzas of this song. As a result, many expected paths of the resulting SSM are weak or even completely missing as illustrated by Figure 2d.

Note that one can simulate transpositions (shifts of one or several semitones) on the feature level simply by cyclically shifting a chroma vector along its twelve dimensions [15]. Based on this observation, we adopt the concept of transposition-invariant self-similarity matrices as introduced in [16]. Here, one first computes the similarity between the original feature sequence and each of the twelve cyclically shifted versions of the chromagram resulting in twelve similarity matrices. Then, the *transposition-invariant SSM*, denoted by $\mathbf{S}^{\text{trans}}$, is calculated by taking the point-wise maximum over these twelve matrices. As indicated by Figure 2e, many of the missing paths are recovered this way.

The cyclic chroma shifts account for transpositions that correspond to the semitone level of the equal-tempered scale. However, when dealing with the folk song field recordings, one may have to deal with pitch fluctuations that are fractions of semitones. One strategy may be to introduce an additional tuning estimation step to adjust the frequency bands used in the chroma decomposition [6, 8] and then to compute the SSM from the resulting features. This strategy only works, when one has to deal with a *global* de-tuning that is constant throughout the recording. For the field recordings, however, one often has to deal with *local* pitch fluctuations. Actually, for many recordings such as OGL19101, the singer *continuously* drops or raises with her voice over the various stanzas. This leads to local path distortions and interruptions (see Figure 2e). To compensate for such local detunings, we further sample the space of semitones using different multirate filter banks corresponding to a shift of 0, $1/4$, $1/3$, $1/2$, $2/3$, and $3/4$ semitones, respectively, see [6]. Using the resulting six different chromagrams together with the twelve cyclically shifted versions of each of them, we compute 72 similarity matrices as above and then take the point-wise maximum over these matrices to obtain a single *fluctuation-invariant SSM*, denoted by $\mathbf{S}^{\text{fluc}}$. This strategy leads to further improvements as as illustrated by Figure 2f, which now shows the expected "full" path structure.

4 Thumbnailing Procedure

In view of our folk song segmentation task, the enhancement of the self-similarity is one main step in order to achieve robustness to spectral deviations. To deal we temporal deviations, we apply a segmentation approach as proposed in [4]. Since in our scenario the recording basically consists of repetitions of a single tune, the segmentation problem reduces to a thumbnailing problem. In general, the goal of *audio thumbnailing* is to find the most representative and repetitive segment of a given music recordings, see, e. g., [7, 17, 18]. Typically, such a segment should have many (approximate) repetitions, and these repetitions should cover large parts of the recording. Let $\alpha = [s:t] \subseteq [1:N]$ denote a segment specified by its starting point s, end point t, and length $|\alpha| := t-s+1$. In [4], a fitness measure is introduced that assigns to each audio segment α a fitness value $\varphi(\alpha) \in \mathbb{R}$ that simultaneously captures two aspects. Firstly, it indicates how well the given segment explains other related segments and, secondly, it indicates how much of the overall music recording is covered by all these related segments. The audio thumbnail is then defined to be the segment α^* having maximal fitness φ over all possible segments.

In the computation of the fitness measure, the main technical idea is to assign to each audio segment α a so-called *optimal path family* over α that simultaneously reveals the relations between α and all other similar segments. One main point is that each path family projected to the vertical axis induces a family of segments, where each element of this family defines a segment similar to α. The induced family of segments defines a segmentation of the audio recording.

As an example, Figure 3 shows path families and induced segment families (vertical axis) for two different segments (horizontal axis) for our running example OGL19101. In Figure 3a the segment is $\alpha = [83:98]$, which corresponds to the sixth stanza A_6. The induced segment family consists of eight different segments, which correspond to the eight stanzas $A_1, A_2, \ldots, A_8$. Figure 3b shows the path family and induced segment family for $\alpha = [66:98]$, which corresponds to the two subsequent stanzas A_5A_6. Here, the induced segment family consists of four segments corresponding to A_1A_2, A_3A_4, A_5A_6, and A_7A_8. The fitness value of a given segment is derived from the corresponding path family and the values of the underlying SSM. It is designed to slightly favor shorter segments to longer segments, see [4] for further details. In our example, it turns out that the fitness-maximizing segment is indeed $\alpha^* = [83:98]$. The induced segment family is taken as final result of our segmentation problem.

Strategy	F0	P	R	F
$\mathbf{S}$	$-$	0.668	0.643	**0.652**
$\mathbf{S}$	$+$	0.734	0.704	**0.717**
$\mathbf{S}^{\text{trans}}$	$+$	0.821	0.821	**0.821**
$\mathbf{S}^{\text{fluc}}$	$+$	0.862	0.855	**0.860**
$\mathbf{S}^{\text{fluc}}$, $\lvert\alpha\rvert \geq 10$	$+$	0.871	0.879	**0.872**
$\mathbf{S}^{\text{fluc}}$, $\lvert\alpha\rvert \geq 10$ (modified dataset)	$+$	0.954	0.940	**0.949**
Reference-based method [3]	$+$	0.912	0.940	**0.926**

Table 1: Precision, recall, and F-measure values for the reference-based segmentation method [3] and the reference-free approach described in this paper.

5 Experiments

We now report on our segmentation experiments using the same dataset as used in [3] containing 47 field recordings of the OGL folk song collection with a total runtime of 156 minutes and their manually annotated segment boundaries. The results obtained by the reference-based method [3] serve as baseline in our experiment.

For the evaluation, as in [3], we use standard precision, recall and F-measures expressing the accuracy of the segmentation boundaries. To this end, we check to what extent the 465 manually annotated stanzas of the evaluation dataset have been identified correctly by the segmentation procedure. More precisely, we say that a computed starting boundary is a *true positive*, if it coincidences with a ground truth boundary up to a small tolerance of ± 2 seconds. Otherwise, the computed boundary is referred to as a *false positive*. Furthermore, a ground truth boundary that is not in the tolerance window of a computed boundary is referred to as a *false negative*. We then compute precision P and recall R values and define the F-measure as $\mathrm{F} := 2 \cdot \mathrm{P} \cdot \mathrm{R} / (\mathrm{P} + \mathrm{R})$.

Table 1 shows the results obtained for our reference-free segmentation procedure as well as the results of the reference-based

method for comparison. Using the original self-similarity matrix $\mathbf{S}$ derived from the original CENS features to determine the fitness maximizing segment α^*, our reference-free method yields an F-measure value of $F = 0.652$. Using our F0-enhanced CENS features to increase the robustness against background noise and small local pitch deviations, the F-measure increases to $F = 0.717$. As mentioned before, dealing with field recordings performed by non-professional singers under poor recording conditions, the matrix enhancement strategies as introduced in Section 3 are extremely important for obtaining robust segmentations. In particular, because of the continuous intonation and pitch shifts of the singers, the concepts of transposition and fluctuation invariance significantly improve the segmentation results. For example, using the transposition-invariant SSM $\mathbf{S}^{\text{trans}}$, the F-measure value increases to $F = 0.821$. Furthermore, when using the fluctuation-invariant SSM $\mathbf{S}^{\text{fluc}}$ that even accounts for shifts corresponding to fractions of a semitone, the F-measure value further increases to $F = 0.860$.

Assuming some prior knowledge on the minimal length of a stanza, the results can be further improved. For example, to avoid over-segmentation [19], one may consider only segments α satisfying $|\alpha| \geq 10$ seconds, which results in $F = 0.872$, see Table 1. This result is still worse than the results obtained from the reference-based approach ($F = 0.926$). Actually, a manual inspection showed that this degradation was mainly caused by four particular recordings, where the segmentation derived from α^* was "phase-shifted" compared to the ground truth. Employing a boundary-based evaluation measure resulted in an F-measure of $F = 0$ for these four recordings. Furthermore, we found out that these phase shifts were caused by the fact that in all of these four recordings the singer completely failed in the first stanza (omitting and confusing entire verse lines). In these cases, the stanza transcript used in the reference-based approach corresponds to the remaining "correct" stanzas. As a result, the reference-based approach can better deal with this issue and is able to recover at least the boundaries of the remaining stanzas.

In a final experiment we simulate a similar behavior by replacing the four recordings using a slightly shortened version, where we omit the first stanzas, respectively. Repeating the previous experiment on this modified dataset produced an F-measure of $F = 0.949$, which is already exceeding the quality obtained by the baseline method. However, there are still some boundaries that are incorrectly detected by our approach. A further investigation revealed that most errors correspond to boundaries that are slightly misplaced and do not fall into the ± 2 seconds tolerance. In many of these cases, there is a short amount of silence between two stanzas, which also introduces some uncertainty to the manually annotated ground-truth boundaries.

6 Conclusions

In this paper, we presented an approach for automatically segmenting folk song field recordings in a robust way even in the presence of significant temporal and spectral distortions across repeating stanzas. One crucial step in the overall segmentation pipeline was to employ various enhancement strategies that allow for dealing with theses distortions already on the feature and SSM levels. Our experiments showed that one obtains good segmentation results having a similar quality as the ones obtained from a previously suggested reference-based method. The described segmentation task is only a first step towards making the audio material more accessible to performance analysis and folk song research. For the future, it would be interesting to develop further tools that allow a folk song researcher to conveniently screen a large number of field recordings in order to detect and locate interesting and surprising features worth being examined in more detail by domain experts. This may open up new challenging and interdisciplinary research directions not only for folk song research but also for music information retrieval and music cognition.

References

[1] Z. Juhász, "Motive identification in 22 folksong corpora using dynamic time warping and self organizing maps," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, (Kobe, Japan), pp. 171–176, 2009.

[2] P. van Kranenburg, J. Garbers, A. Volk, F. Wiering, L. Grijp, and R. Veltkamp, "Towards integration of MIR and folk song research," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, (Vienna, AT), pp. 505–508, 2007.

[3] 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)*, (Kobe, Japan), pp. 735–740, 2009.

[4] 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 Conference on Music Information Retrieval (ISMIR)*, (Miami, FL, USA), pp. 615–620, 2011.

[5] M. Müller, *Information Retrieval for Music and Motion*. Springer Verlag, 2007.

[6] M. Müller and S. Ewert, "Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, (Miami, FL, USA), pp. 215–220, 2011.

[7] M. A. Bartsch and G. H. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 96–104, 2005.

[8] E. Gómez, *Tonal Description of Music Audio Signals*. PhD thesis, UPF Barcelona, 2006.

[9] M. Müller and F. Kurth, "Enhancing similarity matrices for music audio analysis," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Toulouse, France), pp. 437–440, 2006.

[10] G. Peeters, "Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, (Vienna, Austria), pp. 35–40, 2007.

[11] 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, pp. 1138–1151, 2008.

[12] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music.," *Journal of the Acoustical Society of America (JASA)*, vol. 111, no. 4, pp. 1917–1930, 2002.

[13] J. Foote, "Visualizing music and audio using self-similarity," in *Proceedings of the ACM International Conference on Multimedia*, (Orlando, FL, USA), pp. 77–80, 1999.

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

[15] 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.

[16] M. Müller and M. Clausen, "Transposition-invariant self-similarity matrices," in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, (Vienna, Austria), pp. 47–50, 2007.

[17] W. Chai and B. Vercoe, "Music thumbnailing via structural analysis," in *Proceedings of the ACM International Conference on Multimedia*, (Berkeley, CA, USA), pp. 223–226, 2003.

[18] M. Levy, M. Sandler, and M. A. Casey, "Extraction of high-level musical structure from audio data and its application to thumbnail generation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Toulouse, France), pp. 13–16, 2006.

[19] H. Lukashevich, "Towards quantitative measures of evaluating song segmentation," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, (Philadelphia, USA), pp. 375–380, 2008.