ROBUST SEGMENTATION AND ANNOTATION OF FOLK SONG RECORDINGS

Meinard Müller
Saarland University and
MPI Informatik
Saarbrücken, Germany
meinard@mpi-inf.mpg.de

Peter Grosche
Saarland University and
MPI Informatik
Saarbrücken, Germany
pgrosche@mpi-inf.mpg.de

Frans Wiering
Department of Information and
Computing Sciences, Utrecht University
Utrecht, Netherlands
frans.wiering@cs.uu.nl

ABSTRACT

Even though folk songs have been passed down mainly by oral tradition, most musicologists study the relation between folk songs on the basis of score-based transcriptions. Due to the complexity of audio recordings, once having the transcriptions, the original recorded tunes are often no longer studied in the actual folk song research though they still may contain valuable information. In this paper, we introduce an automated approach for segmenting folk song recordings into its constituent stanzas, which can then be made accessible to folk song researchers by means of suitable visualization, searching, and navigation interfaces. Performed by elderly non-professional singers, the main challenge with the recordings is that most singers have serious problems with the intonation, fluctuating with their voices even over several semitones throughout a song. Using a combination of robust audio features along with various cleaning and audio matching strategies, our approach yields accurate segmentations even in the presence of strong deviations.

1. INTRODUCTION

Generally, a folk song is referred to as a song that is sung by the common people of a region or culture during work or social activities. Since many decades, significant efforts have been carried out to assemble and study large collections of folk songs [7, 12]. Even though folk songs were typically transmitted only by oral tradition without any fixed symbolic notation, most of the folk song research is conducted on the basis of notated music material, which is obtained by transcribing recorded tunes into symbolic, score-based music representations. After the transcription, the audio recordings are often no longer studied in the actual research. Since folk songs are part of oral culture, one may conjecture that performance aspects enclosed in the recorded audio material are likely to bear valuable information, which is no longer contained in the transcriptions. Furthermore, even though the notated music material may be more suitable for classifying and identifying folk songs using automated methods, the user may want to listen to the original recordings rather than to synthesized versions of the transcribed tunes.

It is the object of this paper to indicate how the original recordings can be made more easily accessible for folk song researches and listeners by bridging the gap between the symbolic and the audio domain. In particular, we present a procedure for automatically segmenting a given folk song recording that consists of several repetitions of the same tune into its individual stanzas. Using folk song recordings of the *Onder de groene linde* (OGL), main challenges arise from the fact that the songs are performed by elderly non-professional singers under poor recording conditions. The singers often deviate significantly from the expected pitches and 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. As our main contribution, we introduce a combination of robust audio features along with various cleaning and audio matching strategies to account for such deviations and inaccuracies in the audio recordings. Our evaluation on folk song recordings shows that we obtain reliable segmentations even in the presence of strong deviations.

The remainder of this paper is organized as follows. In Sect. 2, we describe the relationship of these investigations to folk song research and describe the folk song collection we employ. In Sect. 3, we show how the recorded songs can be segmented and annotated by locally comparing and aligning the recordings' feature representations with available transcriptions of the tunes. In particular, we introduce various methods for achieving robustness to the aforementioned pitch fluctuations and recording artifacts. Then, in Sect. 4, we report on our systematic experiments conducted on a representative selection of folk song recordings. Finally, in Sect. 5, we indicate how our segmentation results can be used as basis for novel user interfaces, sketch possible applications towards automated performance analysis, and give prospects on future work. Further related work is discussed in the respective sections.

2. FOLK SONG RESEARCH

Folk song reseach has been carried out from many different perspectives. An important problem is to reconstruct and understand the genetic relation between variants of folk songs [12]. Furthermore, by systematically studying entire collections of folk songs, researchers try to discover musical connections and distinctions between different national or regional cultures [7]. To support such research, several databases of encoded folk song melodies have been assembled, the best known of which is the Essen folk song database, [1] which currently contains roughly 20000 folk songs from a variety of sources and cultures. This collection has also been widely used in MIR research.

Even though folk songs have been passed down mainly by oral tradition, most of the folk song research is conducted on the basis of notated music material. However, various folk song collections contain a considerable amount of audio data, which has not yet been explored at a larger scale. One of these collections is *Onder de groene linde* (OGL), which is part of the *Nederlandse Liederenbank* (NLB). The OGL collection comprises several 7277 Dutch folk song recordings along with song transcriptions as well as a rich set of metadata. [2] This metadata includes date and location of recording, information about the singer, and classification by (textual) topic. OGL contains 7277 recordings, which have been digitized as MP3 files. Nearly all of recordings are monophonic, and the vast majority is sung by elderly solo female singers. When the collection was assembled, melodies were transcribed on paper by experts. Usually only one strophe is given in music notation, but variants from other strophes are regularly included. The transcriptions are somewhat idealized: they tend to represent the presumed intention of the singer rather than the actual performance. For about 2500 melodies, transcribed stanzas are available in various symbolic formats including LilyPond, [3] from which MIDI representations have been generated (with a tempo set at 120 BPM for the quarter note).

An important step in unlocking such collections of orally transmitted folk songs is the creation of content-based search engines. The creation of such a search engine is an important goal of the WITCHCRAFT project [8]. The engines should enable a user to search for encoded data using advanced melodic similarity methods. Furthermore, it should also be possible to not only visually present the retrieved items, but also to supply the corresponding audio recordings for acoustic playback. One way of solving this problem is to create robust alignments between retrieved encodings (for example in MIDI format) and the audio recordings. The segmentation and annotation procedure described in the following section exactly accomplishes this task.

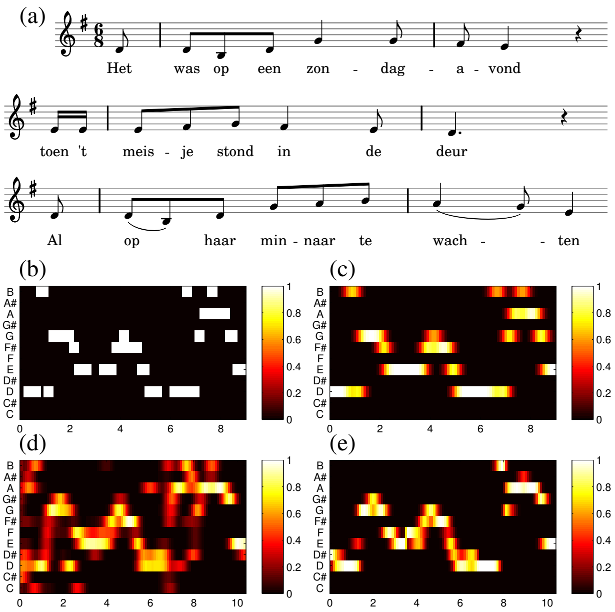

Figure 1. Representations of the beginning of the first stanza of NLB73626 **(a)** Score representation. **(b)** Chromagram of MIDI representation. **(c)** Smoothed MIDI chromagram (CENS). **(d)** Chromagram of audio recording (CENS). **(e)** F0-enhanced chromagram (see Sect. 3.4).

3. FOLK SONG SEGMENTATION

In this section, we present a procedure for automatically segmenting a folk song recording that consists of several repetitions of the same tune into its individual stanzas. Here, we assume that we are given a transcription of a reference tune in form of a MIDI file. Recall from Sect. 2 that this is exactly the situation we have with the songs of the OGL collection. In the first step, we transform the MIDI reference as well as the audio recording into a common mid-level representation. Here, we use the well-known chroma representation, which is summarized in Sect. 3.1. On the basis of this feature representation, the idea is to locally compare the reference with the audio recording by means of a suitable distance function (Sect. 3.2). Using a simple iterative greedy strategy, we derive the segmentation from local minima of the distance function (Sect. 3.3). This approach works well as long as the singer roughly follows the reference tune and stays in tune. However, this is an unrealistic assumption. In particular, most singers have significant problems with the intonation. Their voices often fluctuate by several semitones downwards or upwards across the various stanzas of the same recording. In Sect. 3.4, we show how the segmentation procedure can be improved to account for poor recording conditions, intonation problems, and pitch fluctuations.

3.1 Chroma Features

In order to compare the MIDI reference with the audio recordings, we revert to chroma-based music features, which have turned out to be a powerful mid-level representation for relating harmony-based music, see [1, 6, 9, 11].

[1] http://www.esac-data.org/

[2] The OGL collection is currently hosted at the Meertens Institute in Amsterdam. The metadata of the songs are available through www.liederenbank.nl

[3] www.lilypond.org

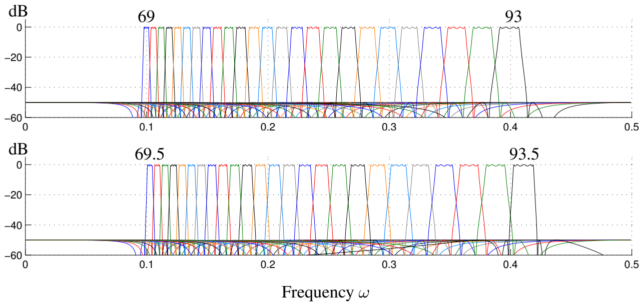

Figure 2. Magnitude responses in dB for some of the pitch filters of the multirate pitch filter bank used for the chroma computation. **Top:** Filters corresponding to MIDI pitches $p \in [69 : 93]$ (with respect to the sampling rate 4410 Hz). **Bottom:** Filters shifted half a semitone upwards.

Here, the chroma refer to the 12 traditional pitch classes of the equal-tempered scale encoded by the attributes $\mathrm{C}, \mathrm{C}^\sharp, \mathrm{D}, \ldots, \mathrm{B}$. Representing the short-time energy content of the signal in each of the 12 pitch classes, chroma features do not only account for the close octave relationship in both melody and harmony as it is prominent in Western music, but also introduce a high degree of robustness to variations in timbre and articulation [1]. Furthermore, normalizing the features makes them invariant to dynamic variations.

It is straightforward to transform a MIDI representation into a chroma representation or chromagram. Using the explicit MIDI pitch and timing information one basically identifies pitches that belong to the same chroma class within a sliding window of a fixed size, see [6]. Fig. 1 shows a score and the resulting MIDI reference chromagram. For transforming an audio recording into a chromagram, one has to revert to signal processing techniques. Most chroma implementations are based on short-time Fourier transforms in combination with binning strategies [1]. In this paper, we revert to chroma features obtained from a pitch decomposition using a multirate pitch filter bank as described in [9]. The employed pitch filters possess a relatively wide passband, while still properly separating adjacent notes thanks to sharp cutoffs in the transition bands, see Fig. 2. Actually, the pitch filters are robust to deviations of up to ± 25 cents[4] from the respective note's center frequency. The pitch filters will play an important role in Sect. 3.4. Finally, in our implementation, we use a quantized and smoothed version of chroma features referred to as CENS features [9] with a feature resolution of 10 Hz (10 features per second), see (c) and (d) of Fig. 1. For technical details, we refer to the cited literature.

3.2 Distance Function

We now introduce a distance function that expresses the distance of the MIDI reference chromagram with suitable subsegments of the audio chromagram. More precisely, let $X = (X(1), X(2), \ldots, X(K))$ be the sequence of chroma features obtained from the MIDI reference and let $Y = (Y(1), Y(2), \ldots, Y(L))$ be the one obtained from the audio recording. In our case, the features $X(k)$, $k \in [1 : K]$, and $Y(\ell)$, $\ell \in [1 : L]$, are normalized 12-dimensional vectors. We define the distance function $\Delta := \Delta_{X,Y} : [1 : L] \to \mathbb{R} \cup \{\infty\}$ with respect to X and Y using a variant of dynamic time warping (DTW):

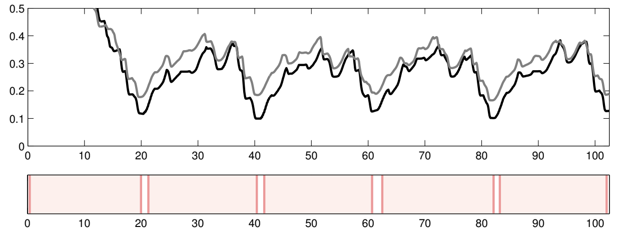

Figure 3. **Top:** Distance function Δ for NLB73626 using original chroma features (gray) and F0-enhanced chroma features (black). **Bottom:** Resulting segmentation.

$$\Delta(\ell) := \frac{1}{K} \min_{a \in [1:\ell]} \Big(\mathrm{DTW}\big(X\,,\, Y(a:\ell)\big)\Big), \qquad (1)$$

where $Y(a : \ell)$ denotes the subsequence of Y starting at index a and ending at index $\ell \in [1 : L]$. Furthermore, $\mathrm{DTW}(X, Y(a : \ell))$ denotes the DTW distance between X and $Y(a : \ell)$ with respect to a suitable local cost measure (in our case, the cosine distance). The distance function Δ can be computed efficiently using dynamic programming. For details on DTW and the distance function, we refer to [9]. The interpretation of Δ is as follows: a small value $\Delta(\ell)$ for some $\ell \in [1 : L]$ indicates that the subsequence of Y starting at index a_ℓ (with $a_\ell \in [1 : \ell]$ denoting the minimizing index in (1)) and ending at index ℓ is similar to X. Here, the index a_ℓ can be recovered by a simple backtracking algorithm within the DTW computation procedure. The distance function Δ for NLB73626 is shown in Fig. 3 as gray curve. The five pronounced minima of Δ indicate the endings of the five stanzas of the audio recording.

3.3 Audio Segmentation

Recall that we assume that a folk song audio recording basically consists of a number of repeating stanzas. Exploiting the existence of a MIDI reference and assuming the repetitive structure of the recording, we apply the following simple greedy segmentation strategy. Using the distance function Δ, we look for the index $\ell \in [1 : L]$ minimizing Δ and compute the starting index a_ℓ. Then, the interval $S_1 := [a_\ell : \ell]$ constitutes the first *segment*. The value $\Delta(\ell)$ is referred to as the *cost* of the segment. To avoid large overlaps between the various segments to be computed, we exclude a neighborhood $[L_\ell : R_\ell] \subset [1 : L]$ around the index ℓ from further consideration. In our strategy, we set $L_\ell := \max(1, \ell - \frac{2}{3}K)$ and $R_\ell := \min(L, \ell + \frac{2}{3}K)$, thus excluding a range of two thirds of the reference length to the left as well as to the right of ℓ. To achieve the exclusion, we modify Δ simply by setting $\Delta(m) := \infty$ for $m \in [L_\ell : R_\ell]$. To determine the next segment S_2,

[4] The *cent* is a logarithmic unit to measure musical intervals. The semitone interval of the equally-tempered scale equals 100 cents.

the same procedure is repeated using the modified distance function, and so on. This results in a sequence of segments $S_1, S_2, S_3, \ldots$. The procedure is repeated until all values of the modified Δ lie above a suitably chosen *quality threshold* $\tau > 0$. Let N denote the number of resulting segments, then $S_1, S_2, \ldots, S_N$ constitutes the final segmentation result, see Fig. 3 for an illustration.

3.4 Enhancement Strategies

Recall that the comparison of the MIDI reference and the audio recording is performed on the basis of chroma representations. Therefore, the segmentation algorithm described so far only works well in the case that the MIDI reference and the audio recording are in the same musical key. Furthermore, the singer has to stick roughly to the pitches of the well-tempered scale. Both assumptions are violated for most of the songs. Even worse, the singers often fluctuate with their voice by several semitones within a single recording. This often leads to poor local minima or even completely useless distance functions as illustrated Fig. 4. To deal with local and global pitch deviations as well as with poor recording conditions, we use a combination of various enhancement strategies.

In our first strategy, we enhance the quality of the chroma features similar to [4] by picking only dominant spectral coefficients, which results in a significant attenuation of noise components. Dealing with monophonic music, we can go even one step further by only picking spectral components that correspond to the fundamental frequency (F0). More precisely, we use a modified autocorrelation method as suggested in [3] to the estimate the fundamental frequency for each audio frame. For each frame, we then determine the MIDI pitch having a center frequency that is closest to the estimated fundamental frequency. Next, in the pitch decomposition used for the chroma computation, we assign energy only to the pitch subband that corresponds to the determined MIDI pitch—all other pitch subbands are set to zero within this frame. Finally, the resulting sparse pitch representation is projected onto a chroma representation and smoothed as before, see Sect. 3.1. The cleaning effect on the resulting chromagram, which is also referred to as *F0-enhanced chromagram*, is illustrated by (d) and (e) of Fig. 1.

Even though the folk song recordings are monophonic, the F0 estimation is often not accurate enough in view of applications such as automated transcription. However, using chroma representations, octave errors as typical in F0 estimations become irrelevant. Furthermore, the F0-based pitch assignment is capable of suppressing most of the noise resulting from poor recording conditions. Finally, local pitch deviations caused by the singers' intonation problems as well as vibrato are compensated to a substantial degree. As a result, the desired local minima of the distance function Δ, which are crucial in our segmentation procedure, become more pronounced. This effect is also illustrated by Fig. 3.

Next, we show how to deal with global pitch deviations and continuous fluctuation across several semitones. To account for a global difference in key between the MIDI reference and the audio recording, we revert to the observation by Goto [5] that the twelve cyclic shifts of a 12-dimensional chroma vector naturally correspond to the twelve possible transpositions. Therefore, it suffices to determine the shift index that minimizes the chroma distance of the audio recording and MIDI reference and then to cyclically shift the audio chromagram according to this index. Note that instead of shifting the audio chromagram, one can also shift the MIDI chromagram in the inverse direction. The minimizing shift index can be determined either by using averaged chroma vectors as suggested in [11] or by computing twelve different distance functions for the twelve shifts, which are then minimized to obtain a single transposition invariant distance functions. We detail on the latter strategy, since it also solves part of the problem having a fluctuating voice within the audio recording. A similar strategy was used in [10] to achieve transposition invariance for music structure analysis tasks.

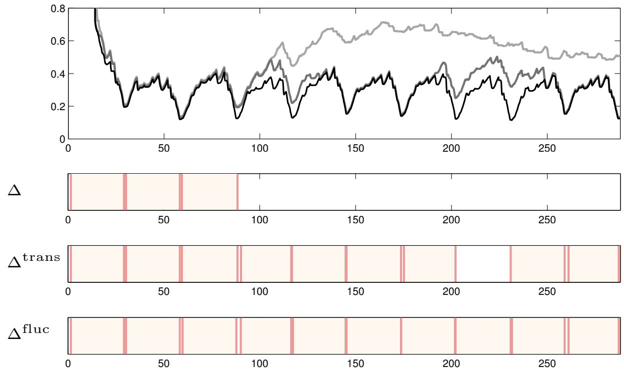

Figure 4. Distance functions Δ (light gray), Δ^{trans} (dark gray), and Δ^{fluc} (black) for the song NLB73286 as well as the resulting segmentations.

Stanza	1	2	3	4	5	6	7	8	9	10
12 shift	5	5	5	4	4	4	4	3	3	3
24 shift	5.0	5.0	4.5	4.5	4.0	4.0	3.5	3.5	3.0	3.0

Table 1. Shift indices (cyclically shifting the audio chromagrams upwards) used for transposing the various stanzas of the audio recording of NLB73286 to optimally match the MIDI reference, see also Fig. 4. The shift indices are given in semitones (obtained by Δ^{trans}) and in half semitones (obtained by Δ^{fluc}).

We simulate the various pitch shifts by considering all twelve possible cyclic shifts of the MIDI reference chromagram. We then compute a separate distance function for each of the shifted reference chromagrams and the original audio chromagram. Finally, we minimize the twelve resulting distance functions, say $\Delta^0, \ldots, \Delta^{11}$, to obtain a single *transposition invariant* distance function $\Delta^{\mathrm{trans}} : [1:L] \rightarrow \mathbb{R} \cup \{\infty\}$:

$$\Delta^{\mathrm{trans}}(\ell) := \min_{i\in[0:11]}\Big(\Delta^i(\ell)\Big). \qquad (2)$$

Fig. 4 shows the resulting function Δ^{trans} for a folk song recording with strong fluctuations. In contrast to the original distance function Δ, the function Δ^{trans} exhibits a number of significant local minima that correctly indicate the segmentation boundaries of the stanzas.

So far, we have accounted for transpositions that refer to the pitch scale of the equal-tempered scale. However, the above mentioned voice fluctuation are fluent in frequency and do not stick to a strict pitch grid. Recall from Sect. 3.1 that our pitch filters can cope with fluctuations of up to ±25 cents. To cope with pitch deviations between 25 and 50 cents, we employ a second filter bank, in the following referred to as *half-shifted filter bank*, where all pitch filters are shifted by half a semitone (50 cents) upwards, see Fig. 2. Using the half-shifted filter bank, one can compute a second chromagram, referred to as *half-shifted chromagram*. A similar strategy is suggested in [4, 11] where generalized chroma representations with 24 or 36 bins (instead of the usual 12 bins) are derived from a short-time Fourier transform. Now, using the original chromagram as well as the half-shifted chromagram in combination with the respective 12 cyclic shifts, one obtains 24 different distance functions in the same way as described above. Minimization over the 24 functions yields a single function Δ^{fluc} referred to as *fluctuation invariant distance function*. The improvements achieved by this novel distance function are illustrated by Fig. 4. Table 1 shows the optimal shift indices derived from the transposition and fluctuation invariant segmentation strategies, where the decreasing indices indicate to which extend the singer's voice rises across the various stanzas of the song.

4. EXPERIMENTS

Our evaluation is based on a dataset consisting of 47 representative folk song recordings selected from the OGL collection, see Sect. 2. The evaluation audio dataset has a total length of 156 minutes, where each of the recorded song consists of 4 to 34 stanzas amounting to a total number of 465 stanzas. The recordings reveal significant deteriorations concerning the audio quality as well as the singer's performance. Furthermore, in various recordings the tunes are overlayed with sounds such as ringing bells, singing birds, or barking dogs, and sometimes the songs are interrupted by remarks of the singers. We manually annotated all audio recordings by specifying the segment boundaries of the stanzas' occurrences in the recordings. Since for most cases the end of a stanza more or less coincides with the beginning of the next stanza and since the beginnings are more important in view of retrieval and navigation applications, we only consider the starting boundaries of the segments in our evaluation. In the following, these boundaries are referred to as *ground truth boundaries*.

To assess the quality of the final segmentation result, we use precision and recall values. To this end, we check to what extent the 465 manually annotated stanzas within 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 given by a parameter δ measured in seconds. Otherwise, the computed boundary is referred to as a *false positive*. Furthermore, a ground truth boundary that is not in a δ-neighborhood of a computed boundary is referred to as a *false negative*. We then compute the precision P and the recall R boundary identification task. From these values one obtains the F-measure $\mathrm{F} := 2 \cdot \mathrm{P} \cdot \mathrm{R}/(\mathrm{P} + \mathrm{R})$.

Strategy	F0	P	R	F	α	β	γ
Δ	$-$	0.898	0.628	0.739	0.338	0.467	0.713
Δ	$+$	0.884	0.688	0.774	0.288	0.447	0.624
Δ^{trans}	$-$	0.866	0.817	0.841	0.294	0.430	0.677
Δ^{trans}	$+$	0.890	0.890	0.890	0.229	0.402	0.559
Δ^{fluc}	$-$	0.899	0.901	0.900	0.266	0.409	0.641
Δ^{fluc}	$+$	0.912	0.940	0.926	0.189	0.374	0.494

Table 2. Performance measures for various segmentation strategies using the tolerance parameter $\delta = 2$ and the quality threshold $\tau = 0.4$. The second column indicates whether original $(-)$ or F0-enhanced $(+)$ chromagrams are used.

δ	P	R	F
1	0.637	0.639	0.638
2	0.912	0.940	0.926
3	0.939	0.968	0.953
4	0.950	0.978	0.964
5	0.958	0.987	0.972

τ	P	R	F
0.1	0.987	0.168	0.287
0.2	0.967	0.628	0.761
0.3	0.950	0.860	0.903
0.4	0.912	0.940	0.926
0.5	0.894	0.944	0.918

Table 3. Dependency of the PR-based performance measures on the tolerance parameter δ and the quality threshold τ. All values refer to Δ^{fluc} using F0-enhanced chromagrams. **Left:** PR-based performance measures for various δ and fixed $\tau = 0.4$. **Right:** PR-based performance measures for various τ and fixed $\delta = 2$.

Table 2 shows the PR-based performance measures of our segmentation procedure using different distance functions with original as well as F0-enhanced chromagrams. In this first experiment, the tolerance parameter is set to $\delta = 2$ and the quality threshold to $\tau = 0.4$. Here, a tolerance of up to $\delta = 2$ seconds seems to us an acceptable deviation in view of our intended applications. For example, the most basic distance function Δ with original chromagrams yields an F-measure of $\mathrm{F} = 0.739$. Using F0-enhanced chromagrams instead of the original ones results in $\mathrm{F} = 0.774$. The best result of $\mathrm{F} = 0.926$ is obtained when using Δ^{fluc} with F0-enhanced chromagrams. Note that all of our introduced enhancement strategies result in an improvement in the F-measure. In particular, the recall values improve significantly when using the transposition and fluctuation-invariant distance functions.

A manual inspection of the segmentation results showed that most of the false negatives as well as false positives are due to deviations in particular at the stanzas' beginnings. The entry into a new stanza seems to be a problem for some of the singers, who need some seconds before getting stable in intonation and pitch. A typical example is NLB72355. Increasing the tolerance parameter δ, the PR-based performance measures improve substantially, as indicated by Table 3 (left). For example, using $\delta = 3$ instead of $\delta = 2$, the F-measure increase from $F = 0.926$ to $F = 0.953$. Other sources of error are that the transcriptions sometimes differ significantly from what is actually sung, as is the case for NLB72395. Here, as was already mentioned in Sect. 2, the transcripts represent the presumed intention of the singer rather than the actual performance. Finally, structural differences between the various stanzas are a further reason for segmentation errors. The handling of such structural differences constitutes an interesting research problem, see Sect. 5. In a further experiment, we investigated the role of the quality threshold τ on the final segmentation results, see Table 3 (right). Not surprisingly, a small τ yields a high precision and a low recall. Increasing τ, the recall increases at the cost of a decrease in precision. The value $\tau = 0.4$ was chosen, since it constitutes a good trade-off between recall and precision.

Finally, to complement our PR-based evaluation, we introduce a second type of more softer performance measures that indicate the significance of the desired minima. To this end, we consider the distance functions for all songs with respect to a fixed strategy and chroma type. Let α be the average over the cost of all ground truth segments (given by the value of the distance function at the corresponding ending boundary). Furthermore, let β be the average over all values of all distance functions. Then the quotient $\gamma = \alpha/\beta$ is a weak indicator on how well the desired minima (the desired true positives) are separated from possible irrelevant minima (the potential false positives). A low value for γ indicates a good separability property of the distance functions. As for the PR-based evaluation, the soft performance measures shown in Table 2 support the usefulness of our enhancement strategies.

5. APPLICATIONS AND FUTURE WORK

Based on the segmentation of the folk song recordings, we now sketch some applications that allow folk song researchers to include audio material in their investigations. Once having segmented the audio recording into stanzas, each audio segment can be aligned with the MIDI reference by a separate MIDI-audio synchronization process with the objective to associate note events given by the MIDI file with their physical occurrences in the audio recording, see [9]. The synchronization result can be regarded as an automated annotation of the entire audio recording with available MIDI events. Such annotations facilitate multimodal browsing and retrieval of MIDI and audio data, thus opening new ways of experiencing and researching music [2]. Furthermore, aligning each stanza of the audio recording to the MIDI reference yields a multialignment between all stanzas. Exploiting this alignment, one can implement interfaces that allow a user to seamlessly switch between the various stanzas of the recording thus facilitating a direct access and comparison of the audio material [9]. Finally, the segmentation and synchronization techniques can be used for automatically extracting expressive aspects referring to tempo, dynamics, and articulation from the audio recording. This makes the audio material accessible for performance analysis, see [13].

For the future, we plan to extend the segmentation scenario dealing with the following kind of questions. How can the segmentation be done if no MIDI reference is available? How can the segmentation be made robust to structural differences in the stanzas? In which way do the recorded stanzas of a song correlate? Where are the consistencies, where are the inconsistencies? Can one extract from this information musical meaningfully conclusions, for example, regarding the importance of certain notes within the melodies? These questions show that the automated processing of folk song recordings constitutes a new challenging and interdisciplinary field of research with many practical implications to folk song research.

Acknowledgement. The first two authors are supported by the Cluster of Excellence on Multimodal Computing and Interaction at Saarland University. Furthermore, the authors thank Anja Volk and Peter van Kranenburg for preparing part of the ground truth segmentations.

6. REFERENCES

[1] M. A. Bartsch and G. H. Wakefield, *Audio thumbnailing of popular music using chroma-based representations*, IEEE Trans. on Multimedia, 7 (2005), pp. 96–104.

[2] D. Damm, C. Fremerey, F. Kurth, M. Müller, and M. Clausen, *Multimodal presentation and browsing of music*, in Proceedings of the 10th International Conference on Multimodal Interfaces (ICMI 2008), 2008.

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

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

[5] M. Goto, *A chorus-section detecting method for musical audio signals*, in Proc. IEEE ICASSP, Hong Kong, China, 2003, pp. 437–440.

[6] N. Hu, R. Dannenberg, and G. Tzanetakis, *Polyphonic audio matching and alignment for music retrieval*, in Proc. IEEE WASPAA, New Paltz, NY, October 2003.

[7] Z. Juhász, *A systematic comparison of different European folk music traditions using self-organizing maps*, Journal of New Music Research, 35 (June 2006), pp. 95–112(18).

[8] F. Wiering, L. P. Grijp, R. C. Veltkamp, J. Garbers, A. Volk, and P. van Kranenburg, *Modelling folksong melodies*, Interdiciplinary Science Reviews, 34.2 (2009), forthcoming.

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

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

[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, 16 (2008), pp. 1138–1151.

[12] P. van Kranenburg, J. Garbers, A. Volk, F. Wiering, L. Grijp, and R. Veltkamp, *Towards integration of music information retrieval and folk song research*, Tech. Report UU-CS-2007-016, Department of Information and Computing Sciences, Utrecht University, 2007.

[13] G. Widmer, S. Dixon, W. Goebl, E. Pampalk, and A. Tobudic, *In search of the Horowitz factor*, AI Mag., 24 (2003), pp. 111–130.