Towards Automated Processing of Folk Song Recordings

Meinard Müller[1], Peter Grosche[1], Frans Wiering[2]

[1] Saarland University and MPI Informatik
Campus E1-4, 66123 Saarbrücken, Germany
meinard@mpi-inf.mpg.de, pgrosche@mpi-inf.mpg.de
[2] Universiteit Utrecht, Department of Information and Computing Sciences
Centrumgebouw Noord, Padualaan 14, De Uithof, 3584CH Utrecht, Netherlands
fransw@cs.uu.nl

Abstract. Folk music is closely related to the musical culture of a specific nation or region. Even though folk songs have been passed down mainly by oral tradition, most musicologists study the relation between folk songs on the basis of symbolic music descriptions, which are obtained by transcribing recorded tunes into a score-like representation. Due to the complexity of audio recordings, once having the transcriptions, the original recorded tunes are often no longer used in the actual folk song research even though they still may contain valuable information. In this paper, we present various techniques for making audio recordings more easily accessible for music researchers. In particular, we show how one can use synchronization techniques to automatically segment and annotate the recorded songs. The processed audio recordings can then be made accessible along with a symbolic transcript by means of suitable visualization, searching, and navigation interfaces to assist folk song researchers to conduct large scale investigations comprising the audio material.

Keywords. Folk songs, audio, segmentation, music synchronization, annotation, performance analysis

1 Introduction

Generally, a folk song is referred to as a song that is sung by the common people of a region or culture reflecting the people's attitude and life. Such songs were typically performed during work and social activities. Originally, folk songs were spread only by oral tradition without any fixed symbolic notation. Therefore, in the process of oral transmission, folk songs have been reshaped in many different ways [1]. During the previous century significant efforts have been carried out to assemble large collections of folk songs, which are not only part of the nations' cultural heritage but also allow musicologists to conduct folk song research on a large scale. Among others, researchers are interested to reconstruct and understand the genetic relation between variants of folk songs [1]. Furthermore,

by systematically studying entire collections of folk songs, researchers try to discover musical connections and distinctions between different national or regional cultures [2].

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, which is obtained by transcribing recorded tunes into symbolic, score-based music representations. After the transcription, the audio recordings are often no longer used in the actual folk song research. This seems somewhat surprising, since one of the most important characteristics of folk songs is that they are part of oral culture [1]. Therefore, 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 generally wants to listen to the original recordings rather than to synthesized versions of the transcribed tunes.

In general, audio material is hard to access due to its massive data volume and complexity [3]. In a specific folk song recording, musically relevant information such as the occurring notes (specified by musical onset times, pitches, and durations), the melody, or the rhythm are not given explicitly, but are somehow hidden in the waveform of the audio signal. To make things even worse, folk songs are typically performed by non-professional singers, who deviate significantly from the expected pitches and musical note onsets. Therefore, most folk song researchers manually transcribe the recorded material and restrict their research to the notated material, which is an idealized description of the actual performance.

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, thus bridging the gap between the symbolic and the audio domain. Because of the aforementioned deviations and inaccuracies in the audio recordings, it is a hard problem to derive reliable transcriptions in an automatic fashion. Instead, our idea is to exploit the availability of manually generated transcriptions for automatically segmenting, structuring, and annotating the audio material. Here, we revert to music synchronization techniques, which allow for interrelating multiple instances and various representations available for a specific folk song [3, 4]. The generated relations and structural information can then be utilized to create novel navigation and retrieval interfaces [5–7], which assist folk song researcher or listener in conveniently accessing, comparing, and analyzing the audio recordings. Furthermore, the generated linking structures can also be used to automatically locate and capture interesting performance aspects that are lost in the notated form of the song.

The remainder of this paper is organized as follows. In Sect. 2, we outline current directions in folk song research and describe the folk song collection *Onder de groene linde* (OGL), which consists of several thousand Dutch folk song recordings along with song transcriptions as well as a rich set of metadata. In Sect. 3, we describe how the recorded songs can be segmented and annotated

by locally comparing and aligning the recordings' feature representations with available transcripts of the tunes. Finally, in Sect. 4, we indicate how these results can be used to create novel user interfaces and sketch possible applications towards automated performance analysis. Conclusions and prospects on future work are given in Sect. 5. Further related work is discussed in the respective sections.

2 Folk Song Research

In the 19th century, an interest in studying folk song traditions emerged in several European countries. Often the underlying motivation for this research was a desire to trace supposedly original and pure aspects of the national musical character. The groundwork for folk song research consisted in collecting and publishing large amounts of folk song melodies. Here, it turned out that these collections contain many related tunes as well as a large variability within related melodies. This variability is caused by the process of oral transmission of these melodies. The songs were learned not from written notation, but by listening and reproducing the melodies from memory. Because of the nature of human memory, changes in the melodies inevitably occurred leading to considerable differences from the original version after several transmission steps.

Melodic variability was studied in great detail for German folk songs by Walter Wiora [8]. Wiora distinguishes seven categories of change, which include changes in melodic contour and rhythm, insertion and deletion of parts, and last but not least demolition of the entire melody.

An important tool in folk song research is the concept of *tune family*, which was defined by Bayard [9] as follows: *A group of melodies showing basic interrelation by means of constant melodic correspondence, and presumably owing their mutual likeness to descent from a single air that has assumed multiple forms through processes of variation, imitation, and assimilation.* The corresponding term used in Dutch folk song research is *melody norm* (melodienorm). In the melody norm, the emphasis lies with the presumed common historical origin of the melodies. An intrinsic difficulty with this concept is that for most cases there is no documentary evidence to reason from. Therefore, in practice, melody norm classification is performed by experts on the basis of musical and textual similarity.

Computational folk song research emerged as early as 1949, when Bertrand Bronson proposed a method to represent folk songs on punch cards [10]. Several folk song 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 been widely used in MIR research. Computational folk song research is surveyed in [1] and in more detail in [11].

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

2.1 OGL Data Collections

In the Netherlands, folk song ballads have been extensively collected and studied. A long-term effort to record these songs was started by Will Scheepers in the early 1950s and continued by Ate Doornbosch until the 1990s [12]. Their field recordings were usually broadcasted in the radio program *Onder de groene linde* (Under the green lime tree). Listeners were encouraged to contact Doornbosch if they knew more about the songs. Doornbosch would then record their version and broadcast it. In this manner a collection was created that not only represents part of the Dutch cultural heritage but also documents the textual and melodic variation resulting from oral transmission.

The OGL collection is currently hosted at the Meertens Institute in Amsterdam. The metadata of the songs are available through the *Nederlandse Liederenbank* (Dutch Song Database[2]). This metadata is very rich including 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 (stereo, 160 kbit/s, 44.1 kHz). 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.

The transcriptions are encoded by hand using a subset of LilyPond[3]. The encodings contain phrase divisions of melodies. If known, a melody norm is assigned to the melody by the encoder. The encodings are automatically converted to Humdrum [13]. MIDI is available in two versions, one obtained from LilyPond and one from Humdrum. The tempo is always set at 120 BPM for the quarter note. At this date (February 2009) the encoded corpus[4] contains approximately 5800 melodies, including 2500 folk songs from OGL, 1400 folk songs from written sources, and 1900 instrumental melodies from written, historical sources.

A subcorpus of OGL was annotated with similarity judgments done by experts for a number of dimensions. This annotated corpus consists of 360 melodies from 26 melody norms, where each melody norms consists of 9 to 27 members. When assigning melody norms, experts selected a prototypical melody and compared candidate members to this prototype. In this comparison, the experts used six musical dimensions, namely rhythm, contour, motifs, mode, text and form. They expressed their judgment on each of these in a similarity score. The options were 0 (not similar in this dimension), 1 (somewhat similar), or 2 (obviously similar). For rhythm, scores were given at the phrase level, for contour both at the phrase and strophe level. A detailed description is provided in [14]. The scores can for example be used to select a subcorpus of melodies that display similarity based on contour or motif.

[2] `www.liederenbank.nl`

[3] `www.lilypond.org`

[4] All the materials are available for research purposes. For information and conditions please contact Frans Wiering.

An important next step in unlocking these collections of orally transmitted folk songs is the creation of content-based search engines, which allow users to browse and navigate within these collections on the basis of the different musical dimensions. The creations of such search engines is an important goal of the WITCHCRAFT project[5]. The engine 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 using music synchronization techniques [3].

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 the form of a MIDI file. Recall from Sect. 2.1 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 sticks to the pitch scale. However, this is an unrealistic assumption. In particular, most singers have significant problems with the intonation. Their voice often fluctuates even 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 such 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 [3, 15, 16]. Here, the chroma refer to the 12 traditional pitch classes of the equal-tempered scale encoded by the attributes C, C$^\sharp$, D, ...,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 [15]. Furthermore, normalizing the features makes them invariant to dynamic variations.

[5] `http://www.cs.uu.nl/research/projects/witchcraft/`

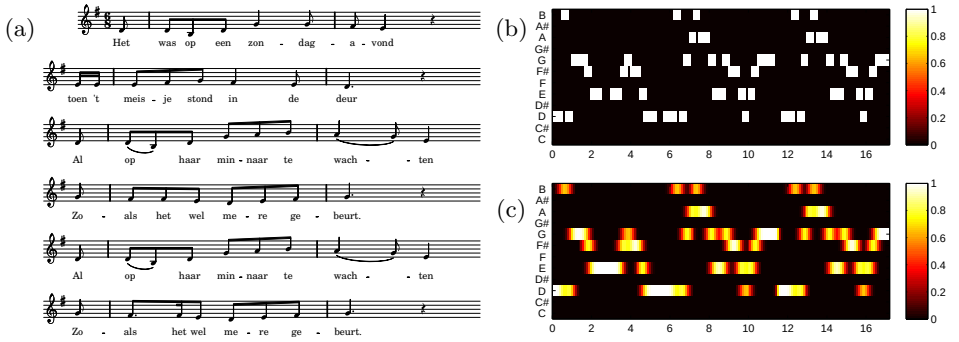

Fig. 1. First stanza of the folk song OGL27517. **(a)** Score representation. **(b)** Chromagram of MIDI representation. **(c)** Smoothed chromagram (CENS).

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 [16]. 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. Here, various techniques have been proposed either based on short-time Fourier transforms in combination with binning strategies [15] or based on suitable multirate filter banks [3]. Fig. 2 (a) shows a chromagram of an audio recording consisting of several stanzas. For technical details, we refer to the cited literature. In our implementation, we use a quantized and smoothed version of chroma features, referred to as CENS features [3] with a feature resolution of 10 Hz (10 features per second), see Fig. 1 (c).

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 $\Delta : [1:L] \to \mathbb{R} \cup \{\infty\}$ with respect to X and Y using a variant of dynamic time warping (DTW):

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

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

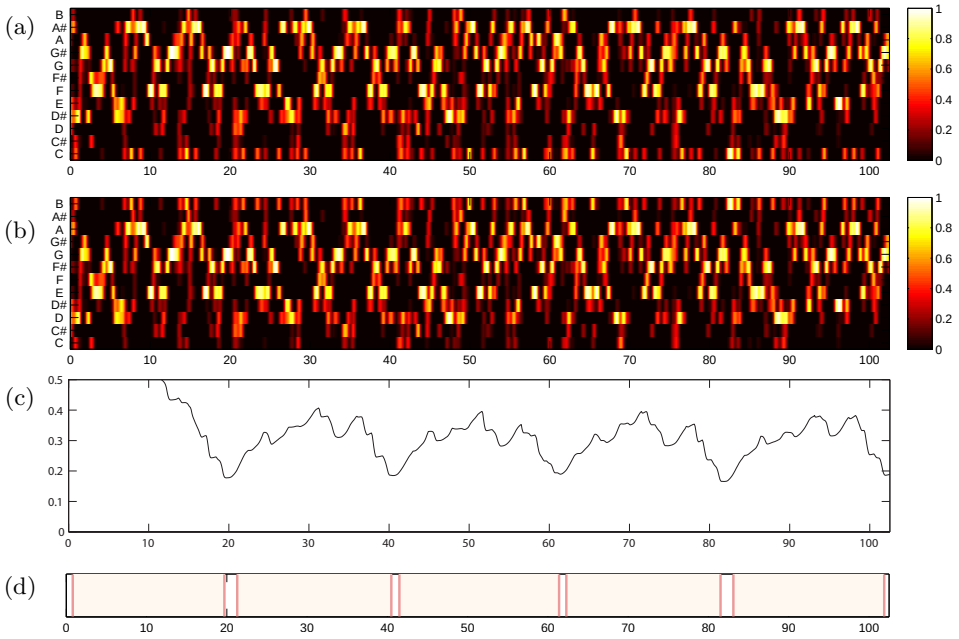

Fig. 2. (a) Chromagram of the audio recording of the folk song OGL27517 consisting of five stanzas. **(b)** Transposed chromagram (cyclically shifted by one pitch downwards to match the key of the MIDI reference). **(c)** Distance function Δ with respect to the MIDI reference chromagram shown in Fig. 1 (c). **(d)** Final segmentation.

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 [3]. 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 back tracking algorithm within the DTW computation procedure. The distance function Δ for the song OGL27517 is shown in Fig. 2 (c). The five pronounced minima of Δ indicate the endings of the five stanzas of the audio recording.

3.3 Audio Segmentation

Recall that the structure of a folk song audio recording is relatively simple, where we assume that it basically consists of a number of repeating stanzas. Exploiting the existence of a MIDI reference and the simple structure of the recording, we can compute the segmentation by the following simple greedy strategy. Using the distance function Δ, we look for the index $\ell \in [1 : L]$ minimizing Δ and

Stanza	Start [sec]	End [sec]	Rank	Cost
1	0.3	20.3	2	0.178
2	20.9	40.8	3	0.185
3	41.3	61.6	5	0.189
4	62.0	81.8	1	0.166
5	82.7	101.9	4	0.186

Table 1. Segmentation result for the audio recording of OGL27517, see also Fig. 2 (d).

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, 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 distance threshold. Let N denote the number of resulting segments. The number $n \in [1 : N]$ is referred to as the *rank* of segment S_n. Fig. 2 (d) and Table 1 show the resulting segmentation of our running example obtained from the distance function shown in Fig. 2 (c).

3.4 Pitch Shift Correction

Recall that the comparison of the MIDI reference and the audio recording is performed on the basis of chroma representations. Therefore, the segmentation algorithm 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. To make things even worse, the singers often fluctuate with their voice by several semitones within a single recording. This may lead to poor or even completely useless distance functions as illustrated Fig. 3.

To account for a global difference in key between the MIDI reference and the audio recording, we revert to the observation by Goto [17] 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 [18] 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

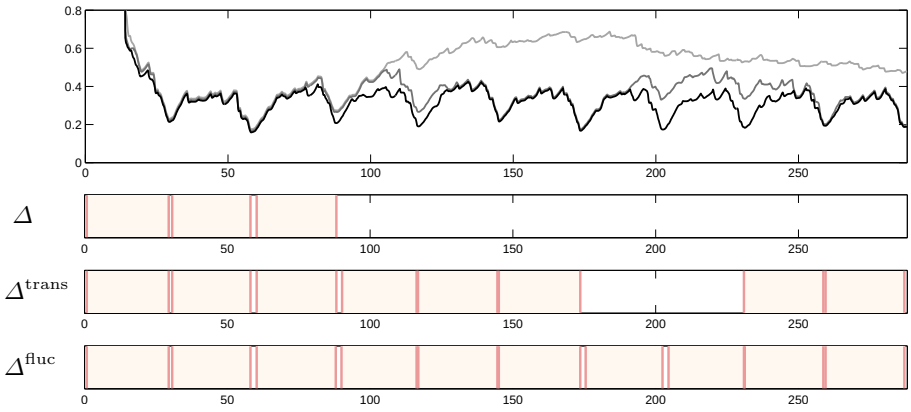

Fig. 3. Distance function Δ (light gray), Δ^{trans} (dark gray), and Δ^{fluc} (black) for the song OGL25010 and the resulting segmentations.

Stanza	1	2	3	4	5	6	7	8	9	10
shift index (semitone)	5	5	5	4	4	4	4	3	3	3
shift index (half semitone)	5.0	5.0	4.5	4.5	4.0	4.0	3.5	3.5	3.0	3.0

Table 2. Transposition of the various stanzas of the audio recording of OGL25010 relative to the MIDI reference. The shift indices are measured in semitones (obtained by Δ^{trans}) and in half semitones (obtained by Δ^{fluc}).

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 [19] to achieve transposition invariance for music structure analysis tasks.

To obtain a transposition invariant distance function, 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 distance function $\Delta^{\text{trans}} : [1:L] \to \mathbb{R} \cup \{\infty\}$:

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

Fig. 3 shows the resulting function Δ^{trans} for a folk song recording with strong fluctuations. In contrast to the original distance function Δ, the *transposition invariant distance 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 singers show a rather poor intonation and often miss the correct pitch. Furthermore, the above mentioned voice fluctuation are fluent in frequency and do not stick to a strict pitch grid. We now

explain how one can deal with such blurred and small-scale pitch deviations. First, in computing the audio chromagrams, we use the multirate filter bank as described in [3]. The employed pitch filters possess a relatively wide pass-band, while still properly separating adjacent notes thanks to sharp cutoffs in the transition bands. Actually, the pitch filters are robust to deviations of up to ± 25 cents[6] from the respective note's center frequency. To cope with 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 a shifted by a half semitone (50 cents) upwards. Using the half-shifted filter bank, one can compute a second chromagram, referred to as *half-shifted chromagram*. A similar strategy is suggested in [20], 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. 3. Here, in regions with a bad intonation, the local minima of Δ^{fluc} are much more significant than those of Δ^{trans}. Table 2 shows the optimal shift indices found for the transposition and fluctuation invariant segmentation strategies. The decreasing indices indicate that the singer's voice constantly rises across the various stanzas of the song.

4 Applications

Based on the segmentation of the folk song recordings, we now sketch some applications that support folk song researchers in including audio material in their investigations. In particular, we show how MIDI-audio synchronization can be used for annotating the audio recordings (Sect. 4.1). Such annotations not only facilitate novel ways for browsing and navigation in audio data (Sect. 4.2) but also yield the basis for performance analysis (Sect. 4.3).

4.1 Audio Annotation

The goal of *MIDI-audio* synchronization is to associate note events given by the MIDI file with their physical occurrences in the audio recording, thus creating musically meaningful cross-links between the two representations [3, 4, 21–24]. The synchronization result can be regarded as an automated annotation of the audio recording with available MIDI events. Once having segmented the audio recording into stanzas, each stanza can be aligned with the MIDI reference by a separate MIDI-audio synchronization process. This can be done in a similar manner as described in Sect. 3.2, where one now globally aligns the chromagrams of the MIDI reference and of a stanza by DTW. From the computed

[6] The *cent* is a logarithmic unit to measure musical intervals. The interval between two adjacent pitches or semitones of the equal-tempered scale equals 100 cents.

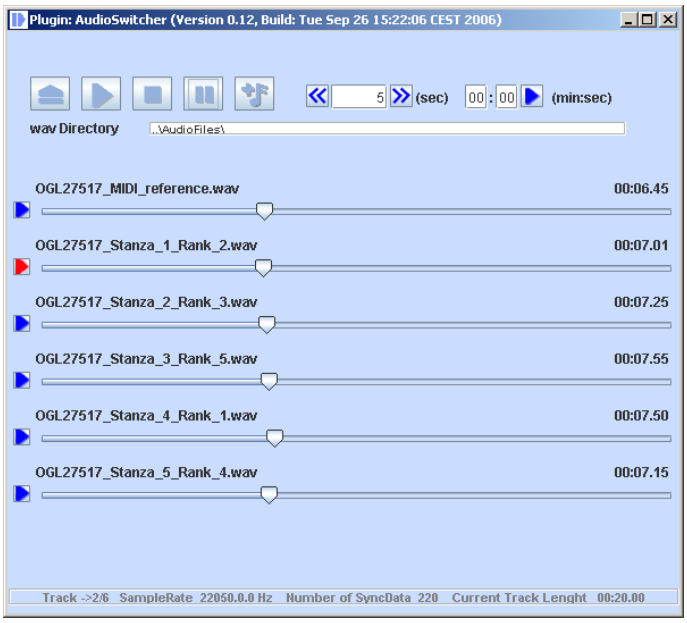

Fig. 4. Instance of the Audio Switcher plug-in of the SyncPlayer showing the synthesized version of the MIDI reference and the five different stanzas of the audio recording of OGL27517.

alignment path, one can then derive the temporal correspondences between the MIDI and the audio representation, see [3] for details. Altogether, one obtains an annotation of the entire audio recording.

Such annotations facilitate multimodal browsing and retrieval in MIDI and audio data, thus opening new ways of experiencing and researching music. For example, most successful algorithms for melody-based retrieval work in the domain of symbolic or MIDI music. On the other hand, retrieval results may be most naturally presented by playing back the original recording of the melody, while a musical score or a piano-roll representation may be the most appropriate form for visually displaying the query results. For a description of such functionalities, we refer to [3, 5, 25]

4.2 Audio Switcher

Aligning each stanza of the audio recording to the MIDI reference yields a multi-alignment between all stanzas. Exploiting the availability of such links, one can implement interfaces that allows a user to seamlessly switch between the various stanzas of the recording thus facilitating a direct access and comparison of the audio material [25, 26, 7].

The *Audio Switcher* [25] constitutes such a user interface, which allows the user to open in parallel a synthesized version of the MIDI reference as well as all stanzas of the folk song recording, see Fig. 4. Each of the stanzas is represented by a slider bar indicating the current playback position with respect to the stanza's particular time scale. The stanza that is currently used for audio playback, in the following referred to as active stanza, is indicated by a red marker located to the left of the slider bar. The slider knob of the active stanza moves at constant speed while the slider knobs of the other stanzas move accordingly to the relative tempo variations with respect to the active stanza. The active stanza may be changed at any time simply by clicking on the respective playback symbol located to the left of each slider bar. The playback of the new active stanza then starts at the time position that musically corresponds to the last playback position of the former active stanza. This has the effect of seamlessly crossfading from one stanza to another while preserving the current playback position in a musical sense. One can also jump to any position within any of the stanzas by directly selecting a position of the respective slider. Such functionalities assists the user in detecting and analyzing the differences between several recorded stanzas of a single folk song.

The Audio Switcher is realized as plug-in of the SyncPlayer system [5, 25], which is an an advanced software audio player with a plug-in interface for MIR applications and provides tools for navigating within audio recordings and browsing in music collections. For further details and functionalities, we refer to the literature.

4.3 Performance Analysis

As a final application, we sketch how the segmentation and synchronization techniques can be used for automatically extracting expressive aspects referring to tempo, dynamics, and articulation from the audio recording. The automated analysis of such expressive aspects, often referred to as *performance analysis*, has become an active research field [27]. Most algorithms for automated performance analysis rely on accurate annotations of the audio material by means of suitable musical parameters. Here, the annotation process is often done manually, which is prohibitive in view of large audio collections. For the case of the folk songs, we present a fully automatic approach for computing tempo curves that reveal the relative tempo difference between two performed stanzas.

As described in Sect. 2.1, the melodies of the OGL songs were manually transcribed based on expert knowledge and then encoded in LilyPond. As a result, one has neutral and idealized representations that do not contain any expressive information concerning tempo or dynamics. The MIDI references were obtained by exporting the LilyPond encodings using a constant tempo of 120 BPM. Now, by comparing a given stanza of a folk song recording with the corresponding MIDI reference, one can derive the local tempo deviations of the respective performance. These tempo deviations can be encoded by means of a *tempo curve*, which yields for each position of the MIDI reference (given in seconds) the deviating factor from the reference tempo at the corresponding position in the

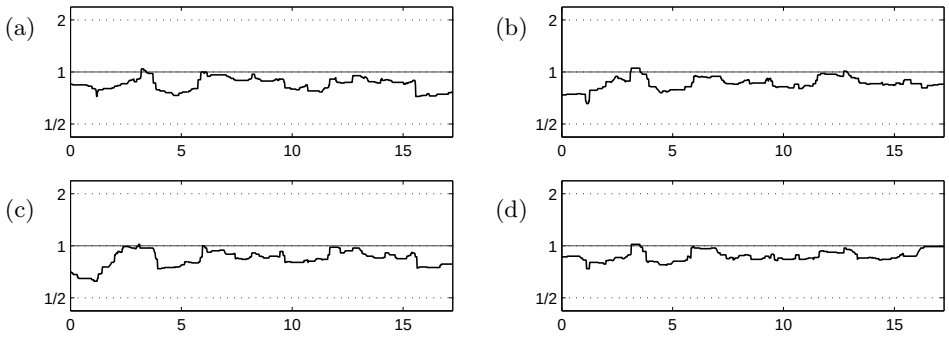

Fig. 5. Tempo curves for the first four stanzas of the song OGL27517. The horizontal axis describes the time scale of the MIDI reference (measured in seconds), while the vertical axis indicates the tempo of the respective stanza relative to the reference (given as factor).

respective performance. As an example, Fig. 5 shows the tempo curves for the first four stanzas of the song OGL27517. Here, a value 1 of the tempo curve indicates that the performance has the same tempo as the MIDI reference (in our case 120 BPM). Similarly, a value 1/2 indicates half the tempo and a value 2 twice the tempo relative to the reference. As the curves of Fig. 5 indicate, the singer starts each stanza with some hesitation (slow tempo), then accelerates before slowing down again towards reference position 5, and so on. Actually, in this example, the four tempo curves reveal similar overall characteristics thus indicating a homogeneous performance with respect to tempo of the singer over the four stanzas.

Similarly, one can extract other important expressive parameters. For example, based on a note-level annotation of a recorded stanza, it is possible to extract the loudness of each sung note within the performance and to derive a dynamic curve. Another interesting aspect would be to capture the actual deviation in frequency of the singer's voice from the expected fundamental frequency given by the reference. Such information would not only reveal expressive elements such as vibrato or glissando but also the inaccuracies such as pitch fluctuations that particularly occur in performances of non-professional singers.

5 Conclusions and Future Work

In this paper, we have introduced various methods from automated music processing with the goal to make recorded folk song material more easily accessible for research and retrieval purposes. In particular, we showed how synchronization techniques can be used for segmenting and annotating folk song recordings performed by elderly non-professional solo singers. Our assumption is that by looking at the original audio recordings, one may derive new insights that can not be derived simply by looking at the transcribed melodies. This assumption is fos-

tered by the fact that folk songs are part of oral culture. Therefore, performance aspects that are enclosed in the recorded audio material but no longer contained in the transcriptions should be an important source in folk song research.

In the next step of our research, we need to systematically evaluate our segmentation algorithm on a larger corpus of folk songs. To this end, we need to establish an evaluation database with manually generated ground truth segmentations. First experiments show that the segmentation procedure can be made more robust to fluctuations by introducing an additional correction step based on previously extracted fundamental frequencies [28, 29]. Such information is also important in view of an automated transcription of the folk song recordings. For the future, we also 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 recorded folk song material constitutes a new challenging and interdisciplinary field of research with many practical implications to folk song research.

References

1. van Kranenburg, P., Garbers, J., Volk, A., Wiering, F., Grijp, L., Veltkamp, R.: Towards integration of MIR and folk song research. In: Proc. ISMIR, Vienna, AT. (2007) 505–508
2. Juhász, Z.: A systematic comparison of different European folk music traditions using self-organizing maps. Journal of New Music Research **35** (June 2006) 95–112(18)
3. Müller, M.: Information Retrieval for Music and Motion. Springer (2007)
4. Arifi, V., Clausen, M., Kurth, F., Müller, M.: Synchronization of music data in score-, MIDI- and PCM-format. Computing in Musicology **13** (2004)
5. Kurth, F., Müller, M., Damm, D., Fremerey, C., Ribbrock, A., Clausen, M.: SyncPlayer – an advanced system for content-based audio access. In: Proc. ISMIR, London, GB. (2005)
6. Fremerey, C., Kurth, F., Müller, M., Clausen, M.: A demonstration of the SyncPlayer system. In: Proc. ISMIR, Vienna, Austria. (2007)
7. Dixon, S., Widmer, G.: Match: A music alignment tool chest. In: Proc. ISMIR, London, GB. (2005)
8. Wiora, W.: Systematik der musikalischen Erscheinungen des Umsingens. Jahrbuch für Volksliedforschung **7** (1941) 128–195
9. Bayard, S.P.: Prolegomena to a study of the principal melodic families of British-American folk song. Journal of American Folklore **63** (1950) 1–44
10. Bronson, B.H.: Some observations about melodic variation in British-American folk tunes. Journal of the American Musicological Society **3** (1950) 120–134
11. van Kranenburg, P., Garbers, J., Volk, A., Wiering, F., Grijp, L., Veltkamp, R.: Towards integration of music information retrieval and folk song research. Technical Report UU-CS-2007-016, Department of Information and Computing Sciences, Utrecht University (2007)

12. Grijp, L.P., Roodenburg, H.: Blues en Balladen. Alan Lomax en Ate Doornbosch, twee muzikale veldwerkers. AUP (2005)
13. Selfridge-Field, E., ed.: Beyond MIDI: the handbook of musical codes. MIT Press, Cambridge, MA, USA (1997)
14. Volk, A., Kranenburg, P.v., Garbers, J., Wiering, F., Veltkamp, R., Grijp, L.: The study of melodic similarity using manual annotation and melody feature sets. Technical Report UU-CS-2008-013, Department of Information and Computing Sciences, Utrecht University (2008)
15. Bartsch, M.A., Wakefield, G.H.: Audio thumbnailing of popular music using chroma-based representations. IEEE Trans. on Multimedia **7** (2005) 96–104
16. Hu, N., Dannenberg, R., Tzanetakis, G.: Polyphonic audio matching and alignment for music retrieval. In: Proc. IEEE WASPAA, New Paltz, NY. (2003)
17. Goto, M.: A chorus-section detecting method for musical audio signals. In: Proc. IEEE ICASSP, Hong Kong, China. (2003) 437–440
18. Serrà, J., Gómez, E., Herrera, P., Serra, X.: Chroma binary similarity and local alignment applied to cover song identification. IEEE Transactions on Audio, Speech and Language Processing **16** (2008) 1138–1151
19. Müller, M., Clausen, M.: Transposition-invariant self-similarity matrices. In: Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007). (2007) 47–50
20. Gómez, E.: Tonal Description of Music Audio Signals. PhD thesis, Ph.D. Dissertation. UPF (2006)
21. Dannenberg, R., Hu, N.: Polyphonic audio matching for score following and intelligent audio editors. In: Proc. ICMC, San Francisco, USA. (2003) 27–34
22. Müller, M., Kurth, F., Röder, T.: Towards an efficient algorithm for automatic score-to-audio synchronization. In: Proc. ISMIR, Barcelona, Spain. (2004)
23. Raphael, C.: A hybrid graphical model for aligning polyphonic audio with musical scores. In: Proc. ISMIR, Barcelona, Spain. (2004)
24. Soulez, F., Rodet, X., Schwarz, D.: Improving polyphonic and poly-instrumental music to score alignment. In: Proc. ISMIR, Baltimore, USA. (2003)
25. Fremerey, C., Kurth, F., Müller, M., Clausen, M.: A Demonstration of the SyncPlayer System. In: Proc. ISMIR, Vienna, AT. (2007)
26. Damm, D., Fremerey, C., Kurth, F., Müller, M., Clausen, M.: Multimodal presentation and browsing of music. In: Proceedings of the 10th International Conference on Multimodal Interfaces (ICMI 2008). (2008)
27. Widmer, G., Dixon, S., Goebl, W., Pampalk, E., Tobudic, A.: In search of the Horowitz factor. AI Mag. **24** (2003) 111–130
28. Klapuri, A.: Multipitch analysis of polyphonic music and speech signals using an auditory model. IEEE Transactions on Audio, Speech, and Language Processing **16** (2008) 255–266
29. de Cheveigné, A., Kawahara, H.: YIN, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America **111** (2002) 1917–1930