

Automated Analysis of Performance Variations in Folk Song Recordings

Meinard Müller
Saarland University and
MPI Informatik
Campus E1.4
Saarbrücken, Germany
meinard@mpi-inf.mpg.de

Peter Grosche
Saarland University and
MPI Informatik
Campus E1.4
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

Performance analysis of recorded music material has become increasingly important in musicological research and music psychology. In this paper, we present various techniques for extracting performance aspects from field recordings folk songs. Main challenges arise from the fact that the recorded songs are performed by non-professional singers, who deviate significantly from the expected pitches and timings even within a single recording of a song. Based on a multimodal approach, we exploit the existence of a symbolic transcription of an idealized stanza in order to analyze a given audio recording of the song that comprises a large number of stanzas. As the main contribution of this paper, we introduce the concept of chroma templates by which consistent and inconsistent aspects across the various stanzas of a recorded song are captured in the form of an explicit and semantically interpretable matrix representation. Altogether, our framework allows for capturing differences in various musical dimension such as tempo, key, tuning, and melody.

Categories and Subject Descriptors

H.5.5 [Information Interfaces and Presentation]: Sound and Music Computing—*Signal analysis, synthesis, and processing*; J.5 [Arts and Humanities]: *Music*

Keywords

Folk songs, performance analysis, music information retrieval, chroma feature, music synchronization

1. INTRODUCTION

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 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. These transcriptions are often idealized and tend to represent the presumed intention of the singer rather than the actual performance. 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. 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.

In this paper, we present various techniques for analyzing the variations within the recorded folk song material, where each song consists of a large number of different stanzas. Main challenges arise from the fact that the recorded 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, from a technical point of view, their voices often fluctuate by several semitones downwards or upwards across the various stanzas of the same recording. Finally, there are also significant temporal and melodic variations between the stanzas belonging to the same folk song recording. It is important to realize that variabilities and inconsistencies may be, to a significant extent, properties of the repertoire and not necessarily errors of the singers. To measure such deviations and variations within the acoustic audio material, we use a multimodal approach by exploiting the existence of a symbolically given transcription of an idealized stanza.

As the main contribution of this paper, we propose a novel method for capturing temporal and melodic characteristics of the various stanzas of a recorded song in a compact matrix representation, which we refer to as *chroma template* (CT). The computation of such a chroma template involves several steps. First, we convert the symbolic transcription as well as each stanza of a recorded song into a suitable chroma representation. On the basis of this feature representation, we determine and compensate for the tuning differences between the recorded stanzas using the transcription as reference. To account for temporal variations, we use time warping techniques to balance out the timing differences between the stanzas. Finally, we derive a chroma template by averaging the suitably transposed and warped chroma representations of all recorded stanzas and the reference. The key property of a chroma template is that it reveals consistent and inconsistent melodic performance aspects across the various

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MIR'10, March 29–31, 2010, Philadelphia, Pennsylvania, USA.
Copyright 2010 ACM 978-1-60558-815-5/10/03 ...\$10.00.

stanzas. Here, one advantage of our concept is its simplicity, where the information is given in form of an explicit and semantically interpretable matrix representation. We show how our framework can be used to automatically measure variabilities in various musical dimensions including tempo, pitch, and melody. Extracting such information constitutes an important step for making the audio material accessible to performance analysis and to folk song research.

The remainder of this paper is structured as follows. First, in Sect. 2, we outline current directions in folk song research and in Sect. 3 we describe the Dutch folk song collection used in our experiments. In Sect. 4, we summarize the concept of chroma features, which are used as common mid-level representation for comparing the symbolic transcriptions and the audio material. In particular, we present various strategies that capture and compensate for variations in intonation and tuning. In Sect. 5, we introduce and discuss in detail our concept of chroma templates. Finally, in Sect. 6, we describe various experiments on performance analysis while discussing our concept by means of a number of representative examples. Conclusions and prospects on future work are given in Sect. 7. Related work is discussed in the respective sections.

2. FOLK SONG RESEARCH

Folk songs are typically performed by common people of a region or culture during work or recreation. These songs are generally not fixed by written scores but are learned and transmitted by listening to and participating in performance. Systematic research on folk song traditions started in the 19th century. At first researchers wrote down folk songs in music notation at performance time, but from an early date onwards performances were recorded using available technologies. Over more than a century of research, enormous amounts of folk song data have been assembled. Since the late 1990s, digitization of folk song holdings has become a matter of course. An overview of European collections is given in [2]. Digitized folk songs offer interesting challenges for computational research, and the availability of extensive folk song material requires computational methods for large-scale musicological investigation of this data. Much interdisciplinary research into such methods has been carried out within the context of music information retrieval (MIR). An important challenge is to create computational methods that contribute to a better musical understanding of the repertoire [21].

Folk songs can be studied from a number of viewpoints: text, music, performance and social context. The musical viewpoint is often concerned with the identification of relationships between folk song melodies at various levels. For example, using computational methods, motivic relationships between different folk song repertoires are studied in [10]. Within individual traditions, the notion of tune family is important. Tune families consist of melodies that are considered to be historically related through the process of oral transmission. In the WITCHCRAFT project, computational models for tune families are investigated in order to create a melody search engine for Dutch folk songs [22, 26]. In the creation of such models aspects from music cognition play an important role. The representation of a song in human memory is not literal. During performance, the actual appearance of the song is recreated. Melodies thus tend to change over time and between performers. But even within

a single performance of a strophic song interesting variations of the melody may be found.

Even though folk songs are typically orally transmitted in performance, much of the research is conducted on the basis of notated musical material and leaves potentially valuable performance aspects enclosed in the recorded audio material out of consideration. Performance analysis has become increasingly important in musicological research and in music psychology. In folk song research (or more widely, in ethnomusicological research) computational methods are beginning to be applied to audio recordings as well. Examples are the study of African tone scales [12] and Turkish rhythms [8]. In [14], the availability of MIDI transcriptions has been exploited to automatically segment audio recordings of strophic folk songs into constituent stanzas. The present paper continues this research by comparing the various stanzas to study performance and melodic variation within a single performance of a folk song.

3. OGL FOLK SONG COLLECTION

In the Netherlands, folk song ballads (strophic, narrative songs) have been extensively collected and studied. A long-term effort to record these songs was started by Will Scheepers in the early 1950s, and it was continued by Ate Doornbosch until the 1990s [7]. 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, in the following referred to as *OGL 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.

At the time of the recording, ballad singing had already largely disappeared from popular culture. Ballads were widely sung during manual work until the first decades of the 20th century. The tradition came to an end as a consequence of two innovations: the radio and the mechanization of manual labor. Decades later, when the recordings were made, the mostly female, elderly singers often had to delve deeply in their memories to retrieve the melodies. The effect is often audible in the recordings: there are numerous false starts, and it is evident that singers regularly began to feel comfortable about their performance only after a few strophes.

The OGL collection, which is currently hosted at the Meertens Institute in Amsterdam, is available through the *Nederlandse Liederenbank* (NLB)¹. The database also gives access to very rich metadata, including date and location of recording, information about the singer, and classification by tune family and (textual) topic. The OGL collection contains 7277 audio recordings, which have been digitized as MP3 files (stereo, 160 kbit/s, 44.1 kHz). Nearly all of the field recordings are monophonic and comprise a large number of stanzas (often more than 10 stanzas). When the collection was assembled, melodies were transcribed on paper by experts. Usually only one stanza is given in music notation, but variants from other stanzas are regularly included. The transcriptions are often idealized and tend to represent the presumed intention of the singer rather than the actual performance. For a large number of melodies,

¹Dutch Song Database, <http://www.liederenbank.nl>

transcribed stanzas are available in various symbolic formats including LilyPond² and Humdrum [19], from which MIDI representations have been generated (with a tempo set at 120 BPM for the quarter note). At this date (November 2009) around 2500 folk songs from OGL have been encoded. In addition, the encoded corpus contains 1400 folk songs from written sources, and 1900 instrumental melodies from written, historical sources, bringing the total number of encoded melodies at approximately 5800. A detailed description of the encoded corpus is provided in [23].

4. CHROMA REPRESENTATION

In the following, we assume that, for a given folk song, we have an audio recording consisting of a various stanzas as well as a transcription of a representative stanza in form of a MIDI file, which will act as a reference. Recall from Sect. 3 that this is exactly the situation we have with the songs of the OGL collection. In order to compare the MIDI reference with the stanzas of the audio recording, we use the well-known chroma features as a common mid-level representation, see [1, 9, 13, 20]. Here, the chroma refer to the 12 traditional pitch classes of the equal-tempered scale encoded by the attributes C, C[#], 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 [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 [9]. Disregarding information on dynamics, we derive a binary chromagram assuming only the values 0 and 1. Furthermore, dealing with monophonic tunes, one has for each frame at most one non-zero chroma entry that is equal to 1. Fig. 1 (b) shows a chromagram of a MIDI reference corresponding to the score shown in Fig. 1 (a). In the following, the chromagram of the transcription is referred to as *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 [1] or based on suitable multirate filter banks [13]. Fig. 1 (c) shows a chromagram of a field recording of a single stanza. In the following, we refer to the chromagram of an audio recording as *audio chromagram*. In our implementation, all chromagrams are computed at a feature resolution of 10 Hz (10 features per second). For technical details, we refer to the cited literature.

As mentioned above, 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. To account for poor recording conditions, intonation problems, and pitch fluctuations we apply various enhancement strategies similar to [14]. First, we enhance the audio chromagram by exploiting the fact that we are dealing with monophonic music. To this end,

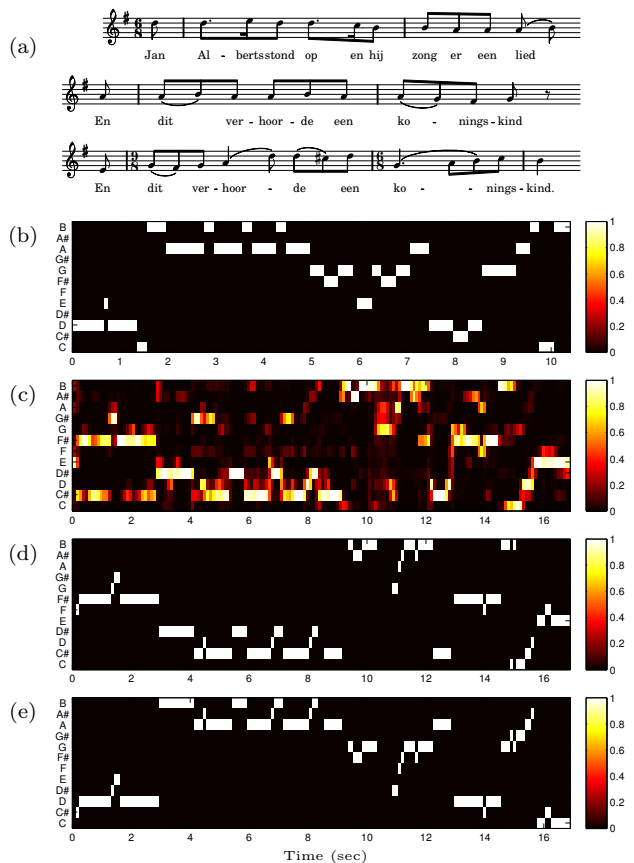


Figure 1: Multimodal representation of a stanza of the folk song NLB72246. (a) Idealized transcription given in form of a score. (b) Reference chromagram of transcription. (c) Audio chromagram of a field recording of a single stanza. (d) F0-enhanced audio chromagram. (e) Transposed F0-enhanced audio chromagram cyclically shifted by eight semitones upwards ($l = 8$).

we use a modified autocorrelation method as suggested in [3] to estimate the fundamental frequency (F0) for each audio frame. Then, we determine the MIDI pitch $p \in [1 : 120]$ having center frequency

$$f(p) = 2^{\frac{p-69}{12}} \cdot 440 \text{ Hz} \quad (1)$$

that is closest to the estimated fundamental frequency. Finally, for each frame, we compute a binary chroma vector having exactly one non-zero entry that corresponds to the determined MIDI pitch projected onto the chroma scale. The resulting binary chromagram is referred to *F0-enhanced audio chromagram*, see Fig. 1 (d). By using an F0-based pitch quantization, most of the noise resulting from poor recording conditions is suppressed. Also local pitch deviations caused by the singers' intonation problems as well as vibrato are compensated to a substantial degree. Furthermore, octave errors as typical in F0 estimations become irrelevant when using chroma representations.

To account for global differences in key between the MIDI reference and the recorded stanzas, we revert to the observation by Goto [6] that the twelve cyclic shifts of a

²www.lilypond.org

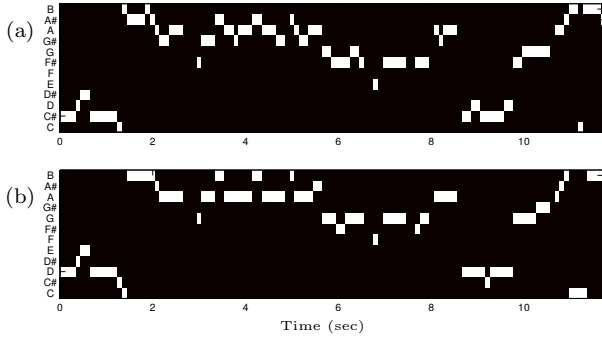


Figure 2: Tuned audio chromagrams of a recorded stanza of the folk song NLB72246. (a) Audio chromagram with respect to tuning parameter $\tau = 6$. (b) Audio chromagram with respect to tuning parameter $\tau = 6.5$.

12-dimensional chroma vector naturally correspond to the twelve possible transpositions. Therefore, it suffices to determine the cyclic shift index $\iota \in [0 : 11]$ (where shifts are considered upwards in the direction of increasing pitch) that minimizes the distance between a stanza’s audio and reference chromagram and then to cyclically shift the audio chromagram according to this index, see Fig. 1. Here, the distance measure between the reference chromagram and the audio chromagram is based on dynamic time warping as described in Sect. 5.

So far, we have accounted for transpositions that correspond to integer semitones of the equal-tempered pitch scale. However, the above mentioned voice fluctuations are fluent in frequency and do not stick to a strict pitch grid. To cope with pitch deviations that are fractions of a semitone, we consider different shifts $\sigma \in [0, 1]$ in the assignment of MIDI pitches and center frequencies as given by (1). More precisely, for a MIDI pitch p , the σ -shifted center frequency $f^\sigma(p)$ is given by

$$f^\sigma(p) = 2^{\frac{p-69-\sigma}{12}} \cdot 440 \text{ Hz.} \quad (2)$$

Now, in the F0-based pitch quantization as described above, one can use σ -shifted center frequencies for different values σ to account for tuning nuances. In our context, we use four different values $\sigma \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$ in combination with the 12 cyclic chroma shifts to obtain 48 different audio chromagrams. Actually, a similar strategy is suggested in [5, 20] where generalized chroma representations with 24 or 36 bins (instead of the usual 12 bins) are derived from a short-time Fourier transform. We then determine the cyclic shift index ι and the shift σ that minimize the distance between the reference chromagram and the resulting audio chromagram. These two minimizing numbers can be expressed by a single rational number

$$\tau := \iota + \sigma \in [0, 12), \quad (3)$$

which we refer to as *tuning parameter*. The audio chromagram obtained by applying a tuning parameter is also referred to as *tuned audio chromagram*. Fig. 2 illustrates the importance of introducing the additional rational shift parameter σ . Here, slight fluctuations around a frequency that lies between the center frequencies of two neighboring

pitches leads to oscillations between the two corresponding chroma bands in the resulting audio chromagram, see Fig. 2 (a). By applying an additional half-semitone shift ($\sigma = 0.5$) in the pitch quantization step, these oscillations are removed, see Fig. 2 (b).

5. CHROMA TEMPLATES

In the last section, we have shown how to handle differences in intonation and tuning by comparing F0-enhanced boolean audio chromagrams with corresponding reference chromagrams. We now show how one can account for temporal and melodic differences by introducing the concept of chroma templates, which reveal consistent and inconsistent performance aspects across the various stanzas. Our concept of chroma templates is similar to the concept of motion templates proposed in [16], which were applied in the context of content-based retrieval of motion capture data.

For a fixed folk song, let $Y \in \{0, 1\}^{d \times L}$ denote the boolean reference chromagram of dimension $d = 12$ and of length (number of columns) $L \in \mathbb{N}$. Furthermore, we assume that for a given field recording of the song we know the segmentation boundaries of its constituent stanzas. Such a segmentation may be derived manually or, with some minor degradation, automatically as described in [14]. We will comment on this in more detail at the end of this section. In the following, let N be the number of stanzas and let $X_n \in \{0, 1\}^{d \times K_n}$, $n \in [1 : N]$, be the F0-enhanced and suitably tuned boolean audio chromagrams, where $K_n \in \mathbb{N}$ denotes the length of X_n . To account for temporal differences, we temporally warp the audio chromagrams to correspond to the reference chromagram Y . Let $X = X_n$ be one of the audio chromagrams of length $K = K_n$. To align X and Y , we employ classical dynamic time warping (DTW) using the Euclidean distance as local cost measure $c : \mathbb{R}^{12} \times \mathbb{R}^{12} \rightarrow \mathbb{R}$ to compare two chroma vectors. (Note that when dealing with binary chroma vectors that have at most one non-zero entry, the Euclidean distance equals the Hamming distance.) Recall that a *warping path* is a sequence $p = (p_1, \dots, p_M)$ with $p_m = (k_m, \ell_m) \in [1 : K] \times [1 : L]$ for $m \in [1 : M]$ satisfying the boundary condition

$$p_1 = (1, 1) \text{ and } p_M = (K, L)$$

as well as the step size condition

$$p_{m+1} - p_m \in \{(1, 0), (0, 1), (1, 1)\}$$

for $m \in [1 : M - 1]$. The total cost of p is defined as $\sum_{m=1}^M c(X(k_m), Y(\ell_m))$. Now, let p^* denote a warping path having minimal total cost among all possible warping paths. Then, the DTW distance $\text{DTW}(X, Y)$ between X and Y is defined to be the total cost of p^* . It is well-known that p^* and $\text{DTW}(X, Y)$ can be computed in $O(KL)$ using dynamic programming, see [13, 17] for details. Next, we locally stretch and contract the audio chromagram X according to the warping information supplied by p^* . Here, we have to consider two cases. In the first case, p^* contains a subsequence of the form

$$(k, \ell), (k, \ell + 1), \dots, (k, \ell + n - 1)$$

for some $n \in \mathbb{N}$, i. e., the column $X(k)$ is aligned to the n columns $Y(\ell), \dots, Y(\ell + n - 1)$ of the reference. In this case, we duplicate the column $X(k)$ by taking n copies of it. In

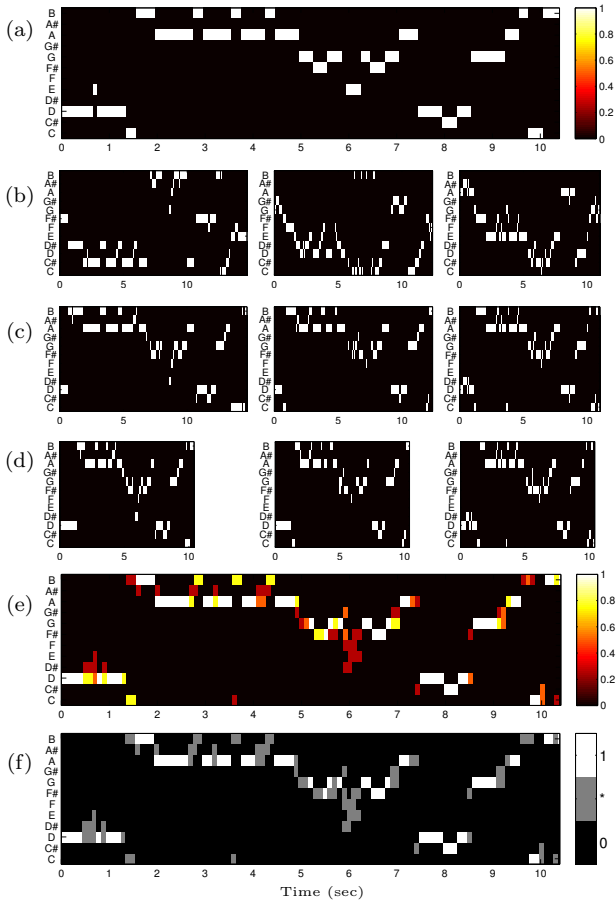


Figure 3: Chroma template computation for the folk song NLB72246. (a) Reference chromagram. (b) Three audio chromagrams. (c) Tuned audio chromagrams. (d) Warped audio chromagrams. (e) Average chromagram obtained by averaging the three audio chromagrams of (d) and the reference of (a). (f) Chroma template.

the second case, p^* contains a subsequence of the form

$$(k, \ell), (k + 1, \ell), \dots, (k + n - 1, \ell)$$

for some $n \in \mathbb{N}$, i. e., the n columns $X(k), \dots, X(k+n-1)$ are aligned to the single column $Y(\ell)$. In this case, we replace the n columns by a single column by taking the component-wise AND-conjunction $X(k) \wedge \dots \wedge X(k+n-1)$. For example, one obtains

$$\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \wedge \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \wedge \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

The resulting warped chromagram is denoted by \bar{X} . Note that \bar{X} is still a boolean chromagram and the length of \bar{X} equals the length L of the reference Y , see Fig. 3 (d) for an example.

After the temporal warping we obtain an optimally tuned and warped audio chromagram for each stanza. Now, we simply average the reference chromagram Y with the warped audio chromagrams $\bar{X}_1, \dots, \bar{X}_N$ to yield an average chroma-

gram

$$Z := \frac{1}{N+1} \left(Y + \sum_{n \in [1:N]} \bar{X}_n \right). \quad (4)$$

Note that the average chromagram Z has real-valued entries between zero and one and has the same length L as the reference chromagram. Fig. 3 (e) shows such an average chromagram obtained from three audio chromagrams and the reference chromagram.

The important observation is that black/white regions of Z indicate periods in time (horizontal axis) where certain chroma bands (vertical axis) consistently assume the same values zero/one in all chromagrams, respectively. By contrast, colored regions indicate inconsistencies mainly resulting from variations in the audio chromagrams (and partly from inappropriate alignments). In other words, the black and white regions encode characteristic aspects that are shared by all chromagrams, whereas the colored regions represent the variations coming from different performances. To make inconsistent aspects more explicit, we further quantize the matrix Z by replacing each entry of Z that is below a threshold δ by zero, each entry that is above $1 - \delta$ by one, and all remaining entries by a *wildcard character* * indicating that the corresponding value is left unspecified, see Fig. 3 (f). The resulting quantized matrix is referred to as *chroma template* for the audio chromagrams X_1, \dots, X_N with respect to the reference chromagram Y . In the following section, we discuss the properties of such chroma templates in detail by means of several representative examples.

As mentioned above, the necessary segmentation of the field recording into its stanzas may be computed automatically. Using a combination of robust audio features along with various cleaning and audio matching strategies, the automated approach as described in [14] yields a segmentation accuracy of over 90 percent for the OGL field recordings, even in the presence of strong deviations. Small segmentation deviations, as our experiments show, do not have a significant impact on the final chroma templates. However, severe segmentation errors that are mainly caused by structural differences between the various stanzas may distort the final results, as is also illustrated by Fig. 6 (c).

6. PERFORMANCE ANALYSIS

The analysis of different interpretations, also referred to as *performance analysis*, has become an active research field [4, 11, 18, 24, 25]. Here, one objective is to extract expressive performance aspects such as tempo, dynamics, and articulation from audio recordings. To this end, one needs accurate annotations of the audio material by means of suitable musical parameters including onset times, note duration, sound intensity, or fundamental frequency. To ensure such a high accuracy, annotation is often done manually, which is infeasible in view of analyzing large audio collections. For the folk song scenario discussed in this paper, we now sketch how various performance aspects can be derived in a fully automated fashion by using the techniques discussed in the previous sections. In particular, we discuss how one can capture performance aspects and variations regarding tuning, tempo, as well as melody across the various stanzas of a field recording.

For the sake of concreteness, we explain these concepts by means of our running example NLB72246 shown in Fig. 1 (a). As discussed in Sect. 4, we first compensate

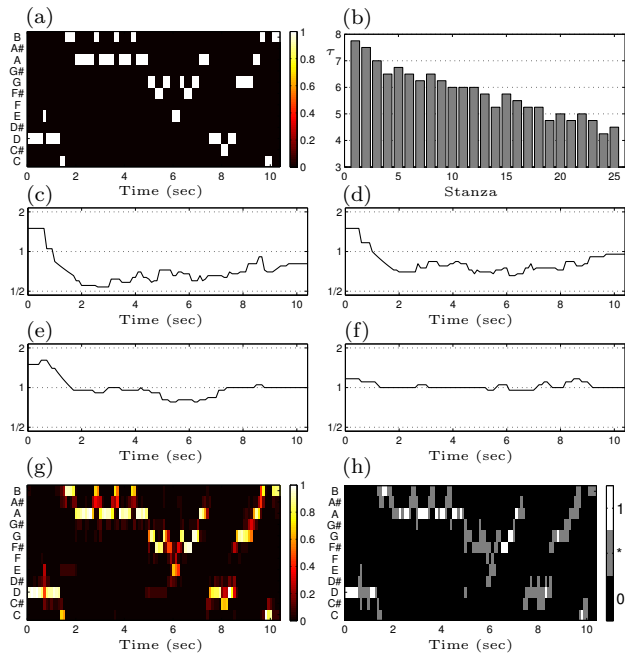


Figure 4: Various performance aspects for a field recording of NLB72246 comprising 25 stanzas. (a) Reference chromagram. (b) Tuning parameter τ for each stanza. (c) - (f) Tempo curves for the stanzas 1, 7, 19, and 25. (g) Average chromagram. (h) Chroma template.

for difference in key and tuning by estimating a tuning parameter τ for each individual stanza of the field recording. This parameter indicates to which extent the stanza's audio chromagram needs to be shifted upwards to optimally agree with the reference chromagram. Fig. 4 (b) shows the tuning parameter τ for each of the 25 stanzas of the field recording. As can be seen, the tuning parameter almost constantly decreases from stanza to stanza, thus indicating a constant rise of the singer's voice. The singer starts the performance by singing the first stanza roughly $\tau = 7.75$ semitones lower than indicated by the reference transcription. Continuously going up with the voice, the singer finishes the song with the last stanza only $\tau = 4.5$ semitones below the transcription, thus differing by more than three semitones from the beginning. Note that in our processing pipeline, we compute tuning parameters on the stanza level. In other words, significant shifts in tuning within a stanza cannot yet be captured by our methods. This may be one unwanted reason when obtaining many inconsistencies in our chroma templates. For the future, we think of methods on how to handle such detuning artifacts within stanzas.

After compensating for tuning differences, we apply DTW-based warping techniques in order to compensate for temporal differences between the recorded stanzas, see Sect. 5. Actually, an optimal warping path p^* encodes the relative tempo difference between the two sequences to be aligned. In our case, one sequence corresponds to one of the performed stanzas of the field recording and the other sequence corresponds to the idealized transcription, which was converted into a MIDI representation using a constant

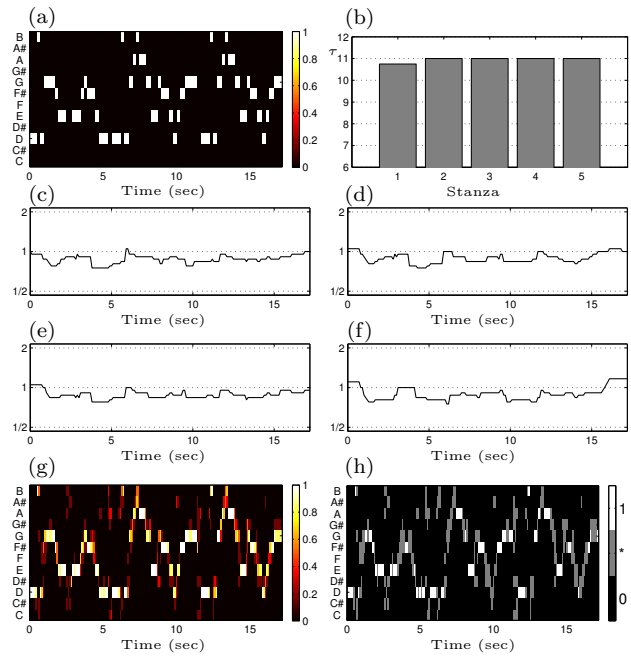


Figure 5: Various performance aspects for a field recording of NLB73626 comprising 5 stanzas. (a) Reference chromagram. (b) Tuning parameter τ for each stanza. (c) - (f) Tempo curves for the first 4 stanzas. (g) Average chromagram. (h) Chroma template.

tempo of 120 BPM. Now, by aligning the performed stanza with the reference stanza (on the level of chromagram representations), one can derive the relative tempo deviations between these two versions [15]. These tempo deviations can be described through a tempo curve that, for each position of the reference, indicates the relative tempo difference between the performance and the reference. In Fig. 4 (c) to (f), the tempo curves for the first four recorded stanzas of NLB72246 are shown. The horizontal axis encodes the time axis of the MIDI reference (rendered at 120 BPM), whereas the vertical encodes the relative tempo difference in form of a factor. For example, a value of 1 indicates that the performance has the same tempo as the reference (in our case 120 BPM). Furthermore, the value 1/2 indicates half the tempo (in our case 60 BPM) and the value 2 indicates twice the tempo relative to the reference (in our case 240 BPM). As can be seen from Fig. 4 (c), the singer performs the first stanza at an average tempo of roughly 85 BPM (factor 0.7). However, the tempo is not constant throughout the stanza. Actually, the singer starts with a fast tempo, then slows down significantly, and accelerates again towards the end of the stanza. Similar tendencies can be observed in the performances of the other stanzas. As an interesting observation, the average tempo of the stanzas continuously increases throughout the performance. Starting with an average tempo of roughly 85 BPM in the first stanza, the tempo averages to 99 BPM in stanza 7, 120 BPM in stanza 19, and reaches 124 BPM in stanza 25. Also, in contrast to stanzas at the beginning of the performance, the tempo is nearly constant for the stanzas towards the end

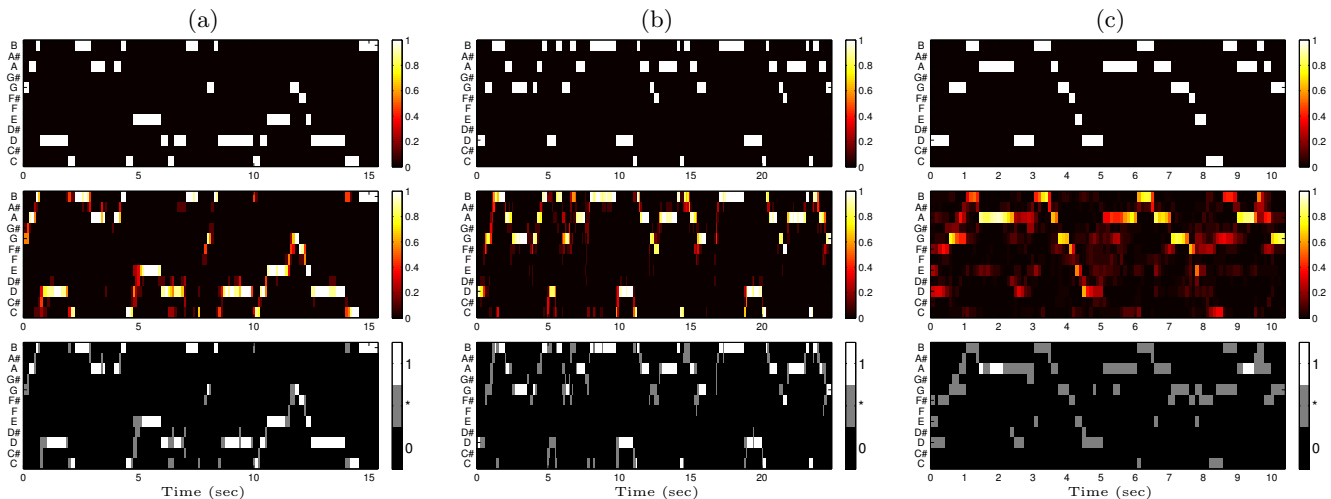


Figure 6: Reference chromagram (top), average chromagram (middle) and chroma template (bottom) for 3 folk song recordings: (a) NLB74437 comprising 8 stanzas. (b) NLB73287 comprising 11 stanzas. (c) NLB72395 comprising 12 stanzas.

of the recording. This may be an indicator that the singer becomes more confident in her singing capabilities as well as in her capabilities of remembering the song.

Finally, after tuning and temporally warping the audio chromagrams, we compute an average chromagram and a chroma template, see Sect. 5. In the quantization step, we use a threshold δ . In our experiments, we set $\delta = 0.1$, thus disregarding inconsistencies that occur in less than 10% of the stanzas. This introduces some robustness towards outliers. The average chromagram and a chroma template for NLB72246 are shown (g) and (h) of Fig. 4, respectively. Here, in contrast to Fig. 3, all 25 stanzas of the field recording were considered in the averaging process. As explained above, the wildcard character * (gray color) of a chroma template indicates inconsistent performance aspects across the various stanzas of the field recording. Since we already compensated for tuning and tempo differences before averaging, the inconsistencies indicated by the chroma templates tend to reflect local melodic inconsistencies and inaccuracies. We illustrate this by our running example, where the inconsistencies particularly occur in the third phrase of the stanza (starting with the fifth second of the MIDI reference). One possible explanation for these inconsistencies may be as follows. In the first two phrases of the stanza, the melody is relatively simple in the sense that neighboring notes differ only either by a unison interval or by a second interval. Also the repeating note A4 plays the role of a stabilizing anchor within the melody. In contrast, the third phrase of the stanza is more involved. Here, the melody contains several larger intervals as well as a meter change. Therefore, because of the higher complexity, the singer may have problems in accurately and consistently performing the third phrase of the stanza.

As a second example, we consider the folk song NLB73626, see Fig. 5. The corresponding field recording comprises 5 stanzas, which are sung in a relatively clean and consistent way. Firstly, the singer keeps the pitch more or less on the same level throughout the performance. This is also indicated by Fig. 5 (b), where one has a tuning parameter

of $\tau = 4$ for all, except for the first stanza where one has $\tau = 3.75$. Secondly, as shown by (c)-(f) of Fig. 5, the average tempo is consistent over all stanzas. Also, the shapes of all the tempo curves are highly correlated. This temporal consistency may be an indicator that the local tempo deviations are a sign of artistic intention rather than a random and unwanted imprecision. Thirdly, the chroma template shown in Fig. 5 (h) exhibits many white regions, thus indicating that many notes of the melody have been performed in a consistent way. The gray areas, in turn, which correspond to the inconsistencies, appear mostly in transition periods between consecutive notes. Furthermore, they tend to have an ascending or descending course while smoothly combining the pitches of consecutive notes. Here, one reason is that the singer tends to slide between two consecutive pitches, which has the effect of some kind of portamento. All of these performance aspects indicate that the singer seems to be quite familiar with the song and confident in her singing capabilities.

We close our discussion on performance analysis by having a look at the chroma templates of another three representative examples. Fig. 6 (a) shows the chroma template of the folk song NLB74437, the field recording of which comprises 8 stanzas. The template shows that the performance is very consistent, with almost all notes remaining unmasked. Actually, this is rather surprising since NLB74437 is one of the few recordings, where several singers perform together. Even though, in comparison to other recordings, the performers do not seem to be particularly good singers and even differ in tuning and melody, singing together seems to mutually stabilize the singers thus resulting in a rather consistent overall performance. Also the chroma template shown in Fig. 6 (b) is relatively consistent. Similarly to the example shown in Fig. 5, there are inconsistencies that are caused by portamento effects. As a last example, we consider the chroma template of the folk song NLB72395, where nearly all notes have been marked as inconsistent, see Fig. 6 (c). This is a kind of negative result, which indicates the limitations of our concept. A manual inspection showed that some of the

stanzas of the field recording exhibit significant structural differences, which are neither reflected by the transcription nor in accordance with most of the other stanzas. For example, in at least two recorded stanzas one entire phrase is omitted by the singer. In such cases, using a global DTW-based approach for aligning the stanzas inevitably leads to poor and semantically meaningless alignments that cause many inconsistencies. The handling of such structural differences constitutes an interesting research problem, which we plan to approach in our future work.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a multimodal approach for extracting performance parameters from folk song recordings by comparing the audio material with symbolically given reference transcriptions. As the main contribution, we introduced the concept of chroma templates that reveal the consistent and inconsistent melodic aspects across the various stanzas of a given recording. In computing these templates, we used tuning and time warping strategies to deal with local variation in melody, tuning and tempo.

The variabilities revealed and observed in this research may have various causes, which need to be further explored in future research. Often these causes are related to questions in the area of music cognition. A first hypothesis is that stable notes are structurally more important than variable notes. The stable notes may be the ones that form part of the singer's mental model of the song, whereas the variable ones are added to the model at performance time. Variations may also be caused by problems in remembering the song. It has been observed that often melodies stabilize after a few iterations. Such variation may offer insight in the working of the musical memory. If the aim is to approach an *accurate* version of the melody, it may be better to discard initial variations. Furthermore, melodic variabilities caused by ornamentations can also be interpreted as a creative aspect of performance. Such variations may be motivated by musical reasons, but also by the lyrics of a song. Sometimes song lines have an irregular length, necessitating the insertion or deletion of notes. Variations may also be made to emphasize key words in the text or, more general, to express the meaning of the song. One would expect such variations to be more or less evenly distributed over the song and not be concentrated at the beginning. Finally one may study details on tempo, timing, pitch, and loudness in relation to performance, as a way of characterizing performance styles of individuals or regions.

As can be seen from these issues, the techniques introduced in this paper constitute only a first step towards making field recordings more accessible to performance analysis and folk song research. Only by using automated methods, one can deal with vast amounts of audio material, which would be infeasible otherwise. Here, our techniques can be considered as a kind of preprocessing to automatically 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 cognition.

Acknowledgement. The first two authors are supported by the Cluster of Excellence on Multimodal Computing and Interaction at Saarland University. Furthermore, the au-

thors thank Anja Volk and Peter van Kranenburg for preparing part of the ground truth segmentations.

8. REFERENCES

- [1] M. A. Bartsch and G. H. Wakefield. Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on Multimedia*, 7(1):96–104, 2005.
- [2] O. Cornelis, M. Lesaffre, D. Moelants, and M. Leman. Access to ethnic music: advances and perspectives in content-based music information retrieval. *Signal Processing*, In Press, 2009.
- [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(4):1917–1930, 2002.
- [4] S. Dixon. Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, 30:39–58, 2001.
- [5] E. Gómez. *Tonal Description of Music Audio Signals*. PhD thesis, UPF Barcelona, 2006.
- [6] M. Goto. A chorus-section detecting method for musical audio signals. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 437–440, Hong Kong, China, 2003.
- [7] L. P. Grijp and H. Roodenburg. *Blues en Balladen. Alan Lomax en Ate Doornbosch, twee muzikale veldwerkers*. AUP, Amsterdam, 2005.
- [8] A. Holzapfel and Y. Stylianou. Rhythmic similarity in traditional Turkish music. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 99–104, Kobe, Japan, 2009.
- [9] N. Hu, R. Dannenberg, and G. Tzanetakis. Polyphonic audio matching and alignment for music retrieval. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2003.
- [10] Z. Juhász. Motive identification in 22 folksong corpora using dynamic time warping and self organizing maps. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 171–176, Kobe, Japan, 2009.
- [11] J. Langner and W. Goebel. Visualizing expressive performance in tempo-loudness space. *Computer Music Journal*, 27(4):69–83, 2003.
- [12] D. Moelants, O. Cornelis, and M. Leman. Exploring African tone scales. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 489–494, Kobe, Japan, 2009.
- [13] M. Müller. *Information Retrieval for Music and Motion*. Springer, 2007.
- [14] M. Müller, P. Grosche, and F. Wiering. Robust segmentation and annotation of folk song recordings. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 735–740, Kobe, Japan, 2009.
- [15] M. Müller, V. Konz, A. Scharfstein, S. Ewert, and M. Clausen. Towards automated extraction of tempo parameters from expressive music recordings. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 69–74, Kobe, Japan, 2009.

- [16] M. Müller and T. Röder. Motion templates for automatic classification and retrieval of motion capture data. In *Proc. ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, pages 137–146, Vienna, Austria, 2006.
- [17] L. R. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall Signal Processing Series, 1993.
- [18] C. S. Sapp. Comparative analysis of multiple musical performances. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 497–500, Vienna, Austria, 2007.
- [19] E. Selfridge-Field, editor. *Beyond MIDI: the handbook of musical codes*. MIT Press, Cambridge, MA, USA, 1997.
- [20] 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*, 16:1138–1151, 2008.
- [21] P. van Kranenburg, J. Garbers, A. Volk, F. Wiering, L. P. Grijp, and R. C. Veltkamp. 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. Forthcoming in *Journal of Interdisciplinary Music Studies* (2010).
- [22] P. van Kranenburg, A. Volk, F. Wiering, and R. C. Veltkamp. Musical models for folk-song melody alignment. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 507–512, Kobe, Japan, 2009.
- [23] A. Volk, P. van Kranenburg, J. Garbers, F. Wiering, R. C. Veltkamp, and L. P. Grijp. 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.
- [24] G. Widmer. Machine discoveries: A few simple, robust local expression principles. *Journal of New Music Research*, 31(1):37–50, 2002.
- [25] G. Widmer, S. Dixon, W. Goebel, E. Pampalk, and A. Tobudic. In search of the Horowitz factor. *AI Magazine*, 24(3):111–130, 2003.
- [26] F. Wiering, L. P. Grijp, R. C. Veltkamp, J. Garbers, A. Volk, and P. van Kranenburg. Modelling folksong melodies. *Interdisciplinary Science Reviews*, 34(2-3):154–171, 2009.