Discovery of repeated melodic phrases in folk singing recordings

Nadine Kroher, Aggelos Pikrakis, *Member, IEEE* and José-Miguel Díaz-Báñez

***Abstract*—In music, repetition is a fundamental concept to establish structure and create temporal relationships. Previous approaches to detecting repetition in music recordings have mainly focused on discovering repeated patterns of variable length and instrumentation at arbitrary locations. In this paper, we present a novel method for the discovery of repeated sung phrases in folk music recordings and in particular in oral music traditions, where written scores are usually unavailable. At a first stage, a segmentation algorithm partitions automatically generated note-level transcriptions of the singing melody into sections which correspond to the structural unit of a phrase. A clustering algorithm is then used to form clusters of phrases, where each cluster contains instances of the same melodic content. The clustering algorithm operates on the basis of a distance measure between melodic sequences and, to this end, various melodic distance measures are investigated. A detailed evaluation procedure is used to assess the performance of the algorithm on three different European music traditions and the influence of transcription and segmentation errors is investigated. The proposed system is shown to outperform the state-of-the-art in audio-based approaches to repeated phrase discovery for this task.**

***Index Terms*—Music Information Retrieval, Melodic pattern discovery, Melody segmentation**

I. Introduction

In recent years, the trend towards constantly growing collections of digitally available audio recordings has expanded beyond the scope of commercial popular and classical music. Digital folk music libraries, like the *Meertens Tune Collection* [1], the *Irish Traditional Music Archive*[1] and the *Andalusian Centre for Documentation of Flamenco Music*[2], provide thousands of recordings for researchers, students and folk music enthusiasts. To avoid time consuming manual annotations and enrich available metadata in ways that advance standard metadata schemes, computational analysis tools for the automatic indexing of folk music traditions can be of major importance. Reliable, efficient and genre-tailored audio indexing technologies can play a key role in the development of powerful tools for large-scale musicological studies and can furthermore enable users to search large collections using high-level content-driven queries and explore the musical content of recordings in intuitively meaningful ways.

Repetition is a fundamental concept in music, which allows the listener to perceive structure via the creation of temporal relationships within a composition. In particular, the structure of European folk music is heavily based on repeated melodic sections at various hierarchical levels, e.g. *motifs*, *phrases* and *stanzas* [2]. The perception of repetition in music, its influence on listener preference, and the related concept of perceived music similarity are frequently addressed topics in the literature (see, for example, [3], [4], [5], [6], [7]). In this study, we aim to automatically detect phrase-level repetition of the singing voice in folk music recordings, so as to generate an automatic structural annotation of a piece, as shown in Figure 1. While in the context of classical, jazz and popular music, the term *phrase* usually refers to the rather loose concept of a "complete" or "resolved" musical segment, in folk music, a *phrase* is a clearly defined, organisational unit [8]. A *stanza* or *verse* can be subdivided into several perceptually identifiable *phrases*, which can again be decomposed to smaller units such as *motifs* and *melodic cells*. Consequently, the task at hand can be formulated as a pattern detection scenario, where the endpoints of the patterns correspond to phrase beginnings

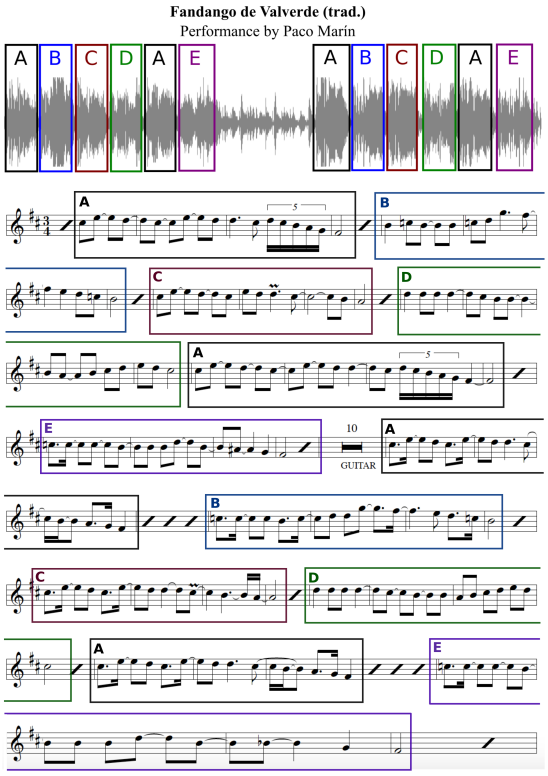

Fig. 1. Structural annotation of a *Fandango de Huelva* recording with respect to sung phrases: (top) Audio annotation (bottom) Corresponding phrases marked on a manual note transcription.

N. Kroher and J. M. Díaz-Bañez are with the Department of Applied Mathematics II, University of Seville, Spain e-mail: nkroher@us.es, dbanez@us.es.

Aggelos Pikrakis is with the Department of Informatics University of Piraeus, Greece, email: pikrakis@unipi.gr.

[1]http://www.itma.ie

[2]http://www.centroandaluzdeflamenco.es

and endings. Furthermore, our analysis is constrained on repetitions in the singing voice melody, which implies the additional challenge of detecting the relevant vocal segments in the possible presence of accompaniment.

We claim that structural analysis at the level of sung phrases provides a comprehensive summary that is suitable for exploring the formal structure of a folk music recording. Furthermore, this type of representation sets the basis for solving various related tasks. For example, stanzas (higher level repetition) can be extracted by detecting repeating sequences of phrases. Repeated phrases can also be used for data compression, making tasks like tune family classification [9] easier to deal with. Moreover, phrase-wise alignment of melodies facilitates similarity studies involving a set of melodies that exhibit structural differences. As another application, query-by-humming tasks can benefit from song representations at the phrase level, because, user-defined queries are likely to contain one or more complete phrases, given that phrases are perceived as closed, coherent units [10]. An example of annotated phrase-level repetition is shown in Figure 1 for a raw audio recording, along with the corresponding ground-truth note transcription.

Several previous methods have focused on the analysis of symbolic music representations. For a review and taxonomy, the reader is referred to [11]. In contrast to our approach, which targets the detection of repeated phrases, most methods operating on symbolic transcriptions have either addressed the detection of smaller structural units (motifs), or the detection of repetition in general, without setting any constraints on pattern length and boundaries. Those studies have revealed that only a fraction of all detected repetitions are actually perceptually relevant [11] and filtering out irrelevant patterns has turned out to be an important problem in such systems. The work in [12] has explored phrase-level repetition in the context of sheet music in order to achieve a repetition-based segmentation of melodies. However, the automatic clustering and respective annotation of the repeated patterns as groups was not addressed by that method. Furthermore, in the symbolic scenario, the patterns can be located anywhere in the score, whereas in our task, only the singing voice melody is relevant and our algorithm needs to be capable of omitting repetition during instrumental interludes.

Approaches targeting the extraction of repetition directly from the audio signal have been mainly developed in the context of music structure analysis [13] and audio-based pattern discovery [14]. Music structure analysis refers to the task of segmenting a recording into adjacent, non-overlapping segments and detecting repetitions among them. In that case, every frame of the audio file belongs to a segment, whereas in our scenario only vocal sections are relevant. In the context of structural annotation of folk music, prior work has targeted the higher level musical unit of stanzas [15], [16], [17] in singing recordings without accompaniment. Audio-based pattern discovery methods attempt to detect repeated segments of variable length. In [18], a large corpus of Indian Art Music is mined by comparing fixed-length segments using dynamic time warping. In an attempt to bridge the semantic gap between audio signals and symbolic representations, [19] applied a symbolic pattern detection approach to automatic polyphonic transcriptions. In [20], repeated themes are detected using techniques from the field of information dynamics. In [21], a silence-based segmentation scheme was used to identify cognitively meaningful units that can be efficiently used for tune family recognition. In contrast to the task of detecting repeated sung phrases, audio-based pattern discovery methods do not demand that the pattern endpoints are located at phrase boundaries and, in general, they do not differentiate between patterns located in the lead voice (in our case the singing melody) and patterns in the accompaniment. A preliminary audio-based attempt to discover phrase-level repetition was proposed by the authors in [22] in the context of accompanied flamenco singing: at a first stage, sung phrases were detected using a vocal detection algorithm and then, at a second stage, pair-wise alignment of chroma-based representations of the detected vocal segments was performed and groups of similar phrases were formed using a frame-centric clustering scheme.

In this paper, we introduce the new task of detecting repeated sung phrases directly from audio recordings that contain a cappella or accompanied singing performances. In addition, we propose a novel method that operates on automatically extracted, most likely noisy, note-level transcriptions of the singing voice melody from such recordings. The symbolic representation contains only the melody of the singing voice, whereas instrumental interludes are encoded as silence. The note transcription stage employs off-the-shelf, state-of-the-art algorithms, which are expected to yield a varying percentage of transcription errors. Instead of focusing on reducing such errors at a post-processing stage, we deal with them during the subsequent stages of our approach. Furthermore, instances of a repeated pattern may exhibit melodic variation introduced by the performer, either in the form of intonation errors or consciously, as an expressive asset. Therefore, we design segmentation and clustering stages that exhibit a certain robustness to transcription errors and melodic variation among repetitions. Specifically, with a simple and efficient procedure, we first segment an automatically generated transcription into phrases by exploiting basic musicological characteristics of phrase boundaries in folk music. We then investigate a number of commonly used distance metrics to compute pair-wise melodic distances among the detected segments. Finally, a clustering scheme is employed to identify clusters (categories) of repeated phrases. The output of our system is an annotation of the audio recording (Figure 2, top), where phrase boundaries and repetitions have been automatically labelled.

Overall, the novelty of our method lies in the combination of the following elements: a new problem formulation and a new approach for solving the problem as a pipeline of building blocks, which operate on the intermediate layer of noisy transcriptions and not on the low-level layer of audio descriptors.

We provide a detailed evaluation of the proposed method on different European folk music traditions and investigate the influence of transcription and phrase segmentation errors on the system performance. Our method yields convincing results and outperforms the approach in [22]. Furthermore, we demonstrate, via a comparative performance analysis and

by analysing the behaviour on selected examples, that representative methods for the more generic tasks of structural analysis and pattern discovery fail to address the current task of phrase-level annotation.

The rest of the paper is structured as follows: the datasets of this study are presented in Section II, the proposed method and the adopted evaluation procedure are described in Section III, the experimental setup and obtained results are presented in Section IV and, finally, conclusions are drawn in Section V.

II. Music datasets

We investigated the automatic discovery of repeated sung phrases in three collections with distinct musical characteristics: a subset of the *Onder de groene linde* dataset (DFS) containing amateur recordings of dutch folk songs, a collection of *Fandangos de Huelva* (FH), a sub-genre of *flamenco* music [23], and a set of Greek *Rebetiko* recordings (REB). A musically trained individual annotated manually all three datasets regarding phrase boundaries and phrase repetitions. In addition, for the FH dataset, a flamenco expert with formal music education provided manual transcriptions of the singing voice melody. For the sake of reproducibility of results, the respective automatic transcriptions, manual annotations and links to the original audio files are made publicly available[3]. The adopted datasets are described below in more detail. In the following, the abbreviation ALL stands for the union of all three datasets. A summary of the number of phrases and repeated patterns and their distributions in the three datasets is given in Table I. It is important to note, that throughout the datasets, all phrases are repeated at least once.

TABLE I
STATISTICS OF THE THREE DATASETS WITH RESPECT TO ANNOTATED PHRASES AND PATTERN REPETITION

	DFS	FH	REB
number of recordings	50	11	30
number of phrases per song (min. / median / max.)	8/22/64	12/12/30	12/24/33
total number of phrases	1235	176	711
number of repetitions per phrase (min. / median / max.)	2/6/22	2/2/9	2/4/9
total number of unique patterns	194	52	178

A. Dutch folk songs (DFS)

The *Onder de groene linde* audio dataset is part of the publicly available *Meertens Tune Collection* [1] which contains more than 7000 recordings of amateur singers performing traditional Dutch folk tunes without instrumental accompaniment. Recordings contain one or more verses of a tune. We selected a subset of 50 songs, for which a total of 1235 phrases were manually annotated to serve as ground-truth data. The annotated phrases are instances of 194 patterns, with each phrase appearing approximately 6 times on the average.

B. Fandangos de Huelva (FH)

The *Fandangos de Huelva* collection contains 11 commercial recordings and has been previously used to evaluate the algorithm in [22], which serves as a baseline method in this study. In all recordings, the singing voice is accompanied by guitar playing. We manually annotated 176 phrases, from which 52 distinct phrases can be extracted (on the average, approximately 3 repetitions per distinct phrase). For all songs, a flamenco expert additionally provided manual corrections of the automatically generated transcriptions, where pitch, segmentation and vocal detection errors were corrected. Such errors include melodic lines stemming from the accompaniment that were mistakenly transcribed as singing voice melody or singing voice segments which were mistakenly classified as accompaniment and therefore not transcribed. These ground truth transcriptions were used in the glass ceiling analysis that we performed (Section IV-D).

[3] This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. This includes automatic transcriptions and manual ground truth annotations of the three datasets and a Matlab implementation of the proposed system.

C. Rebetiko (REB)

In the scope of this study, we also gathered 30 recordings of Greek *Rebetiko* music, available on a video sharing platform, performed by the renown singer *Rita Abatzi*. Rebetiko is an art form of Greek urban songs which appeared towards the end of the 19th century and was shaped to the style that we know today until the third decade of the 20th century. It is a blend of elements from Greek music traditions and influences stemming from the Greeks of Asia Minor. In the recordings that we studied, the singing voice is accompanied by instrumentation including the violin, the bouzouki and the guitar. A total of 711 phrases were manually annotated, stemming from 178 distinct patterns, with each distinct pattern appearing 4 times on the average.

III. Methodology

An overview of the proposed system is shown in Figure 2. Starting from a raw audio recording, we first employ an automatic singing transcription algorithm to estimate a symbolic representation of the vocal melody, acknowledging that the resulting transcription will inevitably contain a percentage of errors. At a second stage, a novel, computationally efficient, phrase segmentation method, which exploits certain musical properties typically encountered in folk songs, segments the previously extracted transcription into a set of M subsequences corresponding to sung phrases. Then, all pairwise melodic distances among the detected subsequences are computed, resulting in a $M \times M$ distance matrix, D. To this end, we investigate various distance measures commonly used in the context of music similarity. Finally, a standard clustering algorithm receives the computed distance matrix and groups the phrases into k categories, where k is automatically estimated on a song basis. As a result, each cluster corresponds to a prototypical melodic pattern and the members of the cluster to the occurrences of the pattern.

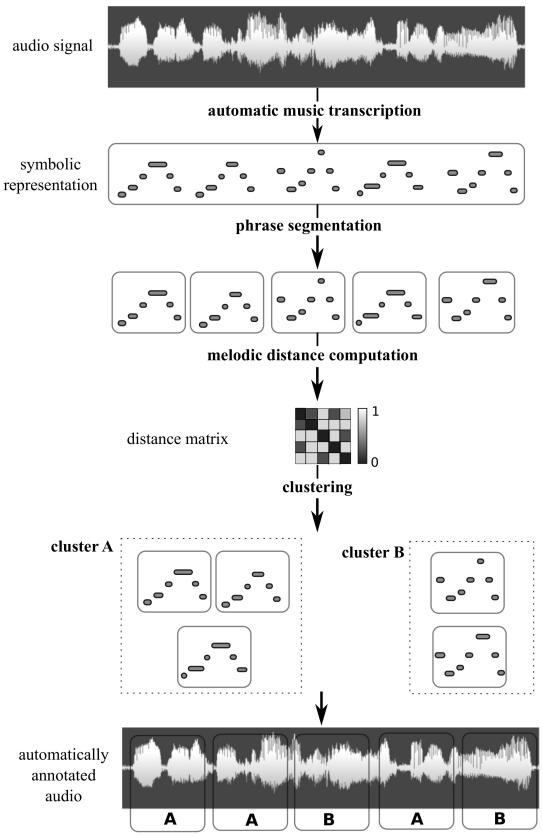

Fig. 2. Discovery of repeated melodic phrases: system overview.

A. Automatic transcription

The first stage of our method employs an Automatic Music Transcription (AMT) algorithm to extract a symbolic representation of the melodic content of the audio recording. We aim to transcribe the melody of the singing voice from monophonic and polyphonic recordings. In our study, the term polyphonic refers to the case of a dominant singing voice in the presence of accompanying instruments. As it is mentioned in [24], best results are achieved with AMT systems which are specifically developed for the genre and instrumentation under study. We therefore employ different state of the art singing transcription systems which are selected based on their suitability for the music corpora at hand.

More specifically, the DFS dataset is transcribed using the *TONY* [25] transcription system, which has been developed for the transcription of monophonic sources. For the *flamenco* recordings of the FH dataset, which contain a large amount of melismatic, micro-tonal ornamentation and guitar accompaniment, we employ the *CANTE* [26] transcription system, which specifically targets singing transcription from accompanied flamenco recordings. The Greek *Rebetiko* corpus exhibits similar characteristics to the *flamenco* case with respect to ornamentation and pitch instability. Therefore, we apply again the system used for the FH dataset, but we replace the vocal detection stage (which targets the particular case of guitar accompaniment) with a generic vocal detector based on machine learning [27]. This detector has given convincing results for this type of instrumentation [28].

In all cases, the output of the AMT system is a symbolic representation in the form of a sequence of note events, where each note is described by its *onset time*, *duration* and *pitch value*. The *CANTE* method provides semitone-quantised MIDI values, while the *TONY* algorithm does not apply quantisation on the extracted frequencies. In the latter case, we quantise to the closest semitone.

B. Structural properties

In this section, we demonstrate three structural properties which serve as basic assumptions for the proposed phrase segmentation algorithm and, to this end, we analyse the manually annotated data with respect to phrase durations and occurrence of silences. To proceed, we first introduce some notation and definitions. Specifically, let us represent an automatic transcription as an ordered sequence, X, of N notes,

$$X = (x_1, ..., x_N)$$

The i-th note, $x_i = (x_i^o, x_i^d, x_i^p, x_i^s)$, is described by its onset time, x_i^o, note duration, x_i^d, pitch, x_i^p and silence, x_i^s, where the term silence stands for the time duration between the note offset and the onset of the next note. Our goal is to split a given transcription, X, into a set of M non-overlapping continuous segments corresponding to melodic phrases. Let $\mathcal{S} = \{S_1, ..., S_M\}$ be a segmentation of X, where $S_j = (x_{L_j}, ..., x_{K_j}), 1 \leq j \leq M, S_j \in \mathcal{S}$, is a continuous subsequence of X (Figure 4 (b)), starting at the L_j-th note and ending at the K_j-th note.

We denote the sum of note durations of the j-th segment, S_j, with τ_{S_j},

$$\tau_{S_j} = \sum_{n=L_j}^{K_j} x_n^d \tag{1}$$

For the sake of simplicity, we will refer to τ_{S_j} with the term *phrase duration*, even though the silence between notes, although part of the phrase, has been ignored in this definition.

Furthermore, let $\hat{\tau}$ be the phrase duration, if a transcription, X (N notes long), is segmented into M segments of equal length, i.e.,

$$\hat{\tau} = \frac{\sum_i^N x_i^d}{M} \tag{2}$$

Figure 3(a) displays the histogram of τ_{S_j}, for the case of manually annotated phrases over all datasets under study. It can be seen that for the majority of phrases (92%), τ_{S_j} ranges between $\tau_{\min} = 2.0$ and $\tau_{\max} = 5.5$ seconds. Figure 3(b) shows the histogram of the absolute duration difference, $\Delta\hat{\tau}_j$, of manually annotated phrases from $\hat{\tau}$. It can be seen that $\Delta\hat{\tau}_j$ is smaller than 1 s for the majority of phrases (90%). It is worth noting that similar observations have been reported for Irish traditional tunes which contain two-bar phrases that are only occasionally fused or split as a means of expressive performance [29]. Further support to our analysis can be found in [4], where a phrase length rule, that favours phrases containing between 6 and 10 notes, was proposed in the context of folk music segmentation. These values were determined based on an analysis of the *Essen folk song collection*.

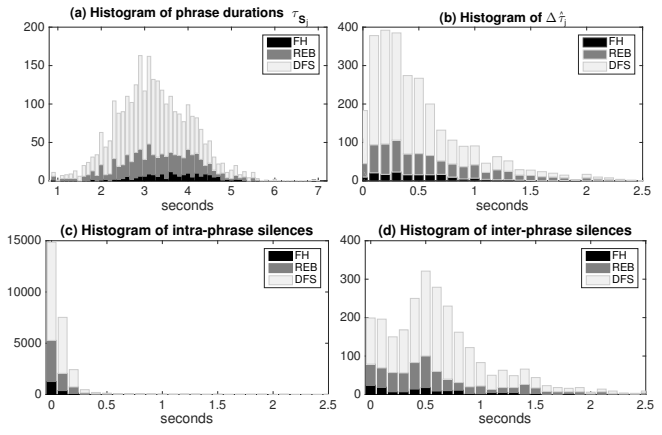

Fig. 3. (a) Histograms of phrase durations (τ_{S_j}), (b) histogram of absolute duration differences, $\Delta\hat{\tau}_j$, (c) intra-phrase silence durations, and (d) inter-phrase silence durations.

Figures 3(c) and 3(d) demonstrate that inter-phrase silence (Figure 3(d)) tends to be longer than intra-phrase silence (Figure 3(c)), indicating that phrase boundaries tend to coincide with vocal rests. This observation is also in line with one of the grouping preference rules of *Generative Theory of Tonal Music* [30], which states that phrase boundaries are likely to occur at rests that are longer than intra-phrase silences. Similar observations have been made in listening experiments [31] and data-driven studies [32], [33].

Consequently, we base the design of our phrase segmentation algorithm on the following three observations:

(I) Phrases cannot be of arbitrary length and global lower and upper phrase duration bounds can be assumed for a given corpus.
(II) Within a song, phrase duration exhibits a relatively small deviation.
(III) The end of a phrase is likely to be followed by a vocal rest that is longer than the average inner-phrase silence.

In other words, our goal is to split a note transcription into a number of phrases which are of similar duration, assuming that the phrase length lies within a pre-defined duration range. Furthermore, long silences should more likely occur at phrase boundaries and less likely within a phrase.

C. Phrase segmentation

Based on the three structural properties described above, we aim for a segmentation, $\mathcal{S} = \{S_1, ..., S_M\}$, where:

(i) the τ_{S_j}'s exhibit a low deviation from the expected $\hat{\tau}$ value
(ii) intra-phrase silences are short compared to inter-phrase silences

This is achieved with a computationally simple algorithm that combines the above objectives to produce the desired segmentation of a given transcription sequence. As the number of phrases, M, in a transcription, is not known in advance, we first define a segmentation algorithm given the number of phrases. Then, we compute a segmentation for each candidate value of M, assess the quality of each segmentation based on the criteria established above and, finally, select the highest quality segmentation.

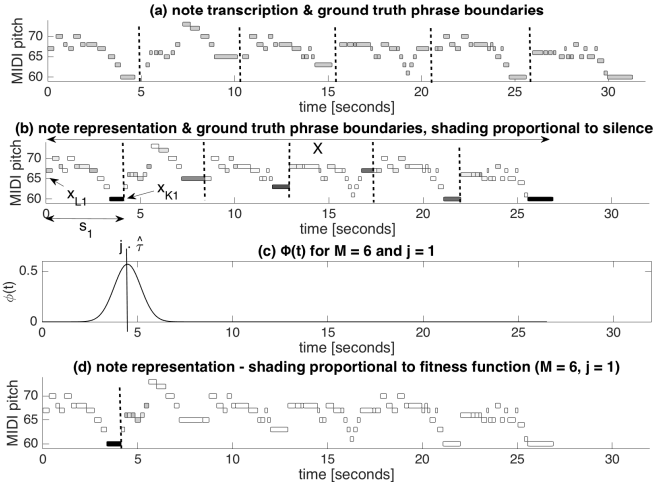

Fig. 4. Phrase segmentation: (a) automatic transcription, (b) note representation with marked ground truth phrase boundaries ($M = 6$) with shading proportional to x^s, (c) $\Phi(t)$ for $j = 1$ and $M = 6$, and (d) note representation and first estimated phrase boundary, shading proportional to $\lambda(x)$;

We now describe the segmentation algorithm when the number of phrases, M, is assumed to be known (Figure 4). In order to identify the j-th phrase boundary, $j = 1, \cdots, M-1$, we assign a segmentation score, $\lambda_j(x_i)$, to each note, x_i, that quantifies its fitness as a cut-off candidate (Figure 4 (b)). The score of x_i is defined as

$$\lambda_j(x_i) = x_i^s \cdot \phi(X, i, j) \tag{3}$$

where ϕ is a Gaussian distribution with the mean corresponding to the optimal temporal segmentation position, $j \cdot \hat{\tau}$, and a standard deviation, $\sigma_{\hat{\tau}}$, which is estimated from the dataset (Figure 4 (c)), i.e.,

$$\phi(X, i, j) = \frac{1}{\sqrt{2\pi}\sigma_{\hat{\tau}}} e^{-\frac{\left(\sum_{n=1}^{i} x_n^d - j\cdot\hat{\tau}\right)^2}{2\sigma_{\hat{\tau}}^2}} \tag{4}$$

The Gaussian is evaluated at the accumulated duration of the first i notes. It can be seen from Eq. (3) that the segmentation score is related directly to the objectives of equal phrase length and long inter-phrase silence. Specifically, function $\phi(X, i, j)$ takes its maximum value at $j \cdot \hat{\tau}$. Segmenting at these positions for $j = 1, \cdots, M-1$ would yield phrases of equal length.

However, in practice, a segmentation at this exact location is usually not possible, since a boundary can only be located after a note. In fact, several possible phrase boundaries may be located in close proximity to $j \cdot \hat{\tau}$. Therefore, factor x_i^s, which denotes the silence duration after note x_i, ensures that high segmentation scores will be assigned to notes that end close to $j \cdot \hat{\tau}$ and are followed by a long silence. It can be inferred, that $\phi(X, i, j)$ is evaluated on the accumulated note duration to which silences and instrumental interludes do not contribute. Furthermore, the contribution of factor x_i^s will force a phrase to end before long instrumental sections or silences. Thus, for each $j = 1, \cdots, M-1$, the j-th phrase boundary is placed

after the note, $x_{i_{\max}}$, that maximises the segmentation score, i.e., $x_{i_{\max}} = \arg\max_i\{\lambda_j(x_i)\}$ (Figure 4 (d)).

We now elaborate on the computation of the estimate of the number of phrase, M. First, by using the aforementioned upper and lower phrase duration boundaries ($\tau_{\min}$ and $\tau_{\max}$) and $\hat{\tau}$, we reduce the candidate values for M to a small set of integers, $\mathcal{M}$. In other words, we only consider positive integer values for M which yield phrase durations that fall into a predefined phrase duration range.

For each $M \in \mathcal{M}$, we compute (using the algorithm above) an optimal segmentation $\mathcal{S}_M$ ($\mathcal{S}$ for short), where each segment is denoted, as before, with $S_j = (x_{L_j}, ..., x_{K_j}), 1 \leq j \leq M, S_j \in \mathcal{S}_M$. We then assign a cost, $c(M)$, to the extracted segmentation for M

$$c(M) = r_{\mathcal{S}} \cdot \sigma_{\tau,\mathcal{S}} \tag{5}$$

where $r_{\mathcal{S}}$ is the ratio of mean inner-phrase to mean inter-phrase silences, given by

$$r_{\mathcal{S}} = \frac{\frac{1}{N-M+1}\sum_{j=1}^{M}\sum_{i=L_j}^{K_j-1} x_i^s}{\frac{1}{M-1}\sum_{j=1}^{M-1} x_{K_j}^s} \tag{6}$$

and $\sigma_{\tau,\mathcal{S}}$ is the standard deviation of durations of the estimated phrases from their mean value. In line with the criteria established above, the cost $c(M)$ will be low if the durations of the resulting phrases exhibit low variation and silences within phrases tend to be shorter than silences at phrase boundaries. Consequently, the value in $\mathcal{M}$ with the lowest computed cost corresponds to the best phrase segmentation.

In our experimental evaluation, parameters $\tau_{\min}$ and $\tau_{\max}$ are estimated directly from the training data and $\sigma_{\hat{\tau}}$ is estimated via an exhaustive search procedure on the training dataset (see Section III-G)

D. Melodic distance computation

As our method is targeting recordings of performances and not written scores, we expect that repetitions of phrases will not to be exact; they are likely to exhibit melodic and/or rhythmic variation because professional performers use variation as an expressive resource. It is also worth noting that amateur singers may reproduce a phrase incorrectly due to intonation and timing inaccuracies. Furthermore, the stage of automatic transcription may introduce pitch, note segmentation and vocal detection errors. However, irrespective of performance variation and errors, our goal remains to identify clusters of similar phrases. Therefore, after partitioning a transcription X into M phrases, $\mathcal{S} = \{S_1, ..., S_M\}$, we proceed to the computation of an $M \times M$ distance matrix, D, such that $D(i,j) = d(S_i, S_j)$, where $S_i, S_j \in \mathcal{S}$ and $i, j = 1, \ldots, M$. The distance function $d(S_i, S_j)$ assigns a dissimilarity value to the alignment of S_i against S_j and, therefore, matrix D holds all pair-wise distances among the detected phrases.

The computation of melodic (dis)similarity has been a popular research topic for many years, especially for melodies that are represented in symbolic (MIDI-like) formats. For a review of related work, the reader is referred to [34] and [35]. The respective problem for the case of audio signals has mainly been studied in the context of query-by-humming [36] systems. In this paper, we investigate four measures that can be used for the alignment of automatically generated transcriptions and which are commonly used in MIR systems.

1) Dynamic time warping (DTW): DTW [37] techniques compute an optimal alignment path and respective matching cost between two sequences of feature vectors. In our study, we first convert the note sequence of each phrase to a one-dimensional representation. Specifically, assuming a short-frame length equal to 0.01 s, the i-th note of the phrase is converted to a sequence of (equal) pitch values, T_i frames long, where $T_i = \text{round}(\frac{x_i^d}{0.01s})$. This type of conversion ignores intra-phrase silences.

Subsequently, at the first stage, if $S_i, S_j \in \mathcal{S}$ are the one-dimensional representations of two phrases with lengths I and J, respectively, then a dissimilarity matrix (cost grid), $D_{s,I\times J}$, is formed, where $D_s(m,n) = 0$, $m = 1, 2, \ldots, I, n = 1, 2, \ldots, J$. D_s is then scanned in a zig-zag mode, from left to right, starting from the first row, and each element (node) accumulates the cost to reach it from its allowable predecessors. In our case, the allowable predecessors are defined by the well known local path constraints which were proposed in [38] and specify that node (m,n) can only be reached from nodes $(m-1,n)$ (vertical transition), $(m,n-1)$ (horizontal transition) and $(m-1,n-1)$ (diagonal transition). Therefore, the cost, $D_s(m,n)$, to reach (m,n) is computed as: $D_s(m,n) = l_c(m,n) + \min\{D_s(m-1,n), D_s(m,n-1), D_s(m-1,n-1)\}$, where $l_c(m,n)$ is the Euclidean distance between the m-th pitch value of S_i and the n-th pitch value of S_j. Besides the Euclidean distance, there exist other measures in the literature [39], which are motivated by musicological assumptions. Those measures require at least a temporal quantisation of the symbolic representation with respect to meter and some even demand for higher-level information, such as the tonality of the piece. However, the automatic transcriptions on which our method operates, describe onset time and note duration as absolute time values and do not provide this type of information. At the end of this processing stage, $D_s(I,J)$ is the global alignment cost between S_i and S_j.

2) Edit distance (EDIT): The edit (or Levenstein) distance [40] computes the dissimilarity between two symbol sequences as the cost of the optimal way to transform one sequence to the other by means of insertions, deletions and substitutions. The edit distance has been frequently applied to estimate melodic dissimilarity between note sequences since the early days of MIR ([41], [42]). In order to apply the edit distance in this paper, we again convert the note representation of a phrase to a sequence of pitch values, as it was done for the DTW algorithm.

3) Earth mover's distance (EMD): The earth mover's distance was originally introduced in the field of image analysis [43] but it has also been used over the years to estimate melodic similarity in the context of various content-based music retrieval methods [44], [45]. To use the EMD measure in this paper, we follow the approach in [44]: melodies are represented as weighted point sets, where pitch x_i^p and onset x_i^o form the coordinates of points on a two-dimensional plane

and note duration, x_i^d, becomes the weight of the point. Note onsets and note durations are normalised to the interval $[0, 1]$. EMD is a transportation distance which estimates the minimum amount of effort needed to transform one weighted point set to another.

4) Shape similarity (SHAPE): The *shape-time* algorithm [46] is a geometric approach which represents melodies as interpolated curves on a pitch-time plane. Similarity among melodies is estimated through the shape similarity of their corresponding curve representations. This approach has been the best performing algorithm in the MIREX[4] symbolic melodic similarity evaluation framework throughout all editions from 2010 to 2015. Due to the fact that the output of the SHAPE algorithm is a similarity value, we map it to the interval $[0, 1]$ using min-max normalization and then we subtract the resulting value from 1, to produce a dissimilarity value.

E. Clustering

Starting from the phrase dissimilarity matrix D, we now aim to identify groups (clusters) of similar phrases. Each cluster will contain instances (repetitions) of the same melodic sequence. Formulating this problem as a clustering task, our goal is to assign a label l_j to each phrase S_j in $\mathcal{S} = \{S_1, ..., S_M\}$, where $l_i \in \{1, ..., k\}$, and k is the number of clusters.

Here, we apply the well-known *k-medoids* [47] clustering algorithm, which estimates a partitioning of a distance matrix into k clusters by minimizing the sum of pairwise intra-cluster distances. Since the number of clusters, k, is unknown, we employ the *silhouette* validation method [48] to estimate the value of k.

We remove clusters containing a single instance, because in the data of our study each phrase is repeated at least once. Therefore, such segments either originate from phrase segmentation errors or from dominant instrumental lines that are mistakenly transcribed as vocals. We thus obtain a set, Ξ, of phrase-level clusters, $\Xi = \{\mathcal{Q}_1, ...\mathcal{Q}_k\}$, where the k-th cluster, $\mathcal{Q}_k = \{Q_{k_1}, ...Q_{k_m}\}$, contains m occurrences of the respective prototypical pattern.

F. Evaluation

In this section we describe the evaluation metrics that we applied in order to assess the quality of the proposed phrase segmenter and the system as a whole. As it was stated in Section I, the task at hand differs from audio-based pattern detection and music structure analysis. Therefore, we modified standard evaluation metrics accordingly. The adopted metrics are based on the pattern discovery metrics that have been used over the past few years in the "*Discovery of repeated themes and sections*" task of the MIREX[5] evaluation framework. The MIREX metrics account for the fact that an algorithm may detect pattern occurrences that overlap, but not necessarily coincide, with the annotated ground truth phrases. The MIREX metrics define the following *cardinality score*, $s_c(P, Q)$, to measure the melodic similarity between two note sequences, P and Q

$$s_c(P, Q) = \frac{P \cap Q}{\max\{|P|, |Q|\}} \tag{7}$$

where the intersection $P \cap Q$ is the set of the note symbols which both sequences have in common and $|.|$ stands for sequence length. While this score is suitable for algorithms that analyse symbolic data or audio-based scenarios where the written score is available, the task at hand requires a minor modification because ground truth transcriptions are not available for all datasets under study. Therefore, instead of evaluating the intersection of two note-set representations, we assess their temporal intersection, where a segment $P = \{t_P^s, t_P^e\}$ is defined by its left and right endpoints, t_P^s and t_P^e, respectively (both measured in time units):

$$\frac{P \cap Q}{\max\{|P|, |Q|\}} \equiv \frac{\min\{t_P^e, t_Q^e\} - \max\{t_P^s, t_Q^s\}}{\max\{t_P^e - t_P^s, t_Q^e - t_Q^s\}} \tag{8}$$

In this way, mistakenly detected patterns in instrumental sections will yield a zero score and detected patterns with endpoints that do not coincide with phrase beginnings and endings will result in a reduced score. This metric was also used to assess the performance of the system in [22]. Having defined the modified score, we can now proceed to the description of the procedures that evaluate the phrase segmentation stage and the system as a whole.

1) Phrase segmentation: To evaluate the phrase segmenter, we ignore the clustering output and only take into account the endpoints of the detected and manually annotated phrases. In the related MIREX structural segmentation task, the output of a segmentation algorithm is evaluated based on precision and recall of detected boundaries which are within a tolerance range of a ground truth boundary. However, although structural segmentation aims at segmenting a stream into adjoining sections, sung phrases are not necessarily adjoining and might alternate with instrumental sections, which are not considered phrases. Consequently, evaluating boundary retrieval can be misleading in the context of phrase segmentation. An example is shown in Figure 5, where mistakenly estimating all instrumental interludes as phrases yields a boundary recall of 0.67 and a boundary precision of 1.0.

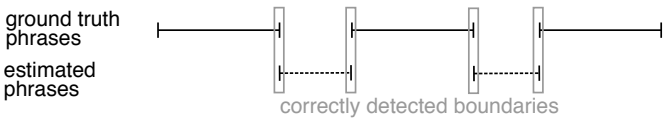

Fig. 5. Example of a misleading result when applying boundary detection evaluation to phrase segmentation.

We therefore adapt the MIREX boundary retrieval metrics to the task at hand, by distinguishing between phrase start and end points. A detected phrase onset is accepted as a hit, if it is within a tolerance range around a ground truth phrase onset. Similarly, a phrase offset is accepted, if it is within a tolerance range around an annotated phrase offset. Here, we use a tolerance of 0.5 seconds, as suggested in [49]. Based on the correctly identified phrase on- and offsets, we compute

[4]http://www.music-ir.org/mirex/wiki/MIREX_HOME
[5]http://www.music-ir.org/mirex/wiki/

the *phrase precision* ($\mathrm{Phr_P}$), *phrase recall* ($\mathrm{Phr_R}$) and *phrase f-measure* ($\mathrm{Phr_F}$).

In addition, we compute the median deviation in seconds between detected and estimated phrase boundaries as proposed in [49], while respecting again the distinction between phrase onsets and offsets. Specifically, we compute the *median true-to-guess deviation* ($\mathrm{Phr_{TTG}}$) as the median deviation between true phrase boundaries and the closest estimated boundary, and the *median guess-to-true deviation* ($\mathrm{Phr_{GTT}}$)) as the median deviation between estimated phrase boundaries and the closest ground truth boundary.

2) System as a whole: To evaluate the system as a whole, i.e., the joint performance of the segmenter, the melodic distance measure and the clustering algorithm, we adopt the metrics from the MIREX task on "*Discovery of repeated themes and sections*", but each metric is modified based on Eq. (8). The *establishment* metrics, i.e., establishment precision Est_P, recall Est_R and F-measure Est_F, measure the algorithm's capability of detecting that a given ground truth pattern appears at least once as a repetition throughout the recording. In the context of our formulation of the task at hand, the establishment metrics refer to the fact that a phrase is a member of a cluster containing more than one instances. The more detailed *occurrence* measures, i.e., occurrence precision Occ_P, recall Occ_R and F-measure Occ_F, evaluate how many repetitions of a given pattern are successfully detected. An occurrence is considered to be correctly detected if it overlaps with a ground truth pattern instance by more than 75%. For more details regarding the MIREX definition, the reader is referred to the respective MIREX task description[6].

G. Cross-fold validation

All experiments described in Section IV are performed in a 5-fold cross-validation scheme. Care is taken so that the instances of each dataset are uniformly distributed over the folds. At each run, we estimate the parameters $\tau_{\min}$ and $\tau_{\max}$ from the train partition as the 5th and 95th percentile of annotated phrase durations, respectively. The parameter $\sigma_{\hat{\tau}}$ is determined based on the highest achieved Est_F value on the train set in an exhaustive search over values ranging from 0.1 to 2.0 in steps of 0.1. The reported performance metrics refer to track-wise values which are first averaged within each run and then across runs. For the generation of genre-specific reports, the same procedure is applied (train and test folds contain data from all databases), but metrics are reported separately for each genre.

IV. Results

In this paper we provide a detailed experimental evaluation of the proposed method. First, we evaluate the phrase segmentation stage by means of phrase precision, recall and f-measure and then, we compute the pattern discovery metrics for different melodic distance measures. We evaluate the system as a whole across datasets and folds, and investigate the influence of phrase segmentation and automatic transcription errors from a glass ceiling analysis perspective. In addition, we provide some examples of automatically generated annotations and give an overview of common errors. Finally, we show the relation of our repeated phrase discovery approach to neighbouring MIR tasks by analysing the output of state of the art algorithms for selected audio examples. We demonstrate that these algorithms are not suitable for the task of repeated sung phrase discovery via a comparative evaluation on the three datasets.

[6]http://www.music-ir.org/mirex/wiki/2016:Discovery_of_Repeated_Themes_%26_Sections

A. Phrase segmentation evaluation

Figure 6 shows the phrase evaluation metrics described in Section III-F1 for all four datasets. We can observe certain performance differences across genres. We obtain a phrase f-measure of 0.87 on the DFS dataset, and only 0.55 on the FH dataset and 0.57 on the REB collection. Through manual inspection we identified vocal detection errors in the two polyphonic datasets, FH and REB, as the main source of error. Specifically, dominant melodic lines from the accompaniment, e.g. from the violin in REB, are mistakenly transcribed, causing the segmentation into sections of similar length to fail. Furthermore, for these two datasets, we observe much larger values for $\mathrm{Phr_{GTT}}$ compared to $\mathrm{Phr_{TTG}}$ ($\mathrm{Phr_{GTT}} = 0.76$ vs. $\mathrm{Phr_{TTG}} = 0.34$ for FH and $\mathrm{Phr_{GTT}} = 0.74$ vs. $\mathrm{Phr_{TTG}} = 0.21$ for REB). These values indicate a tendency towards over- rather than under-segmentation. On the other hand, for the DFS dataset, these two metrics are more balanced ($\mathrm{Phr_{GTT}} = 0.13$ vs. $\mathrm{Phr_{TTG}} = 0.12$).

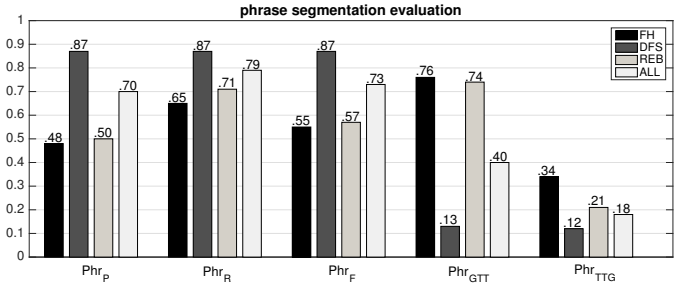

Fig. 6. Phrase evaluation metrics across datasets.

B. Comparison of melodic distance measures

We now proceed to the comparison of the four melodic measures described in Section III-C and we investigate how they affect the clustering results. Specifically, we assess the overall system performance by means of the establishment and occurrence f-measures (Figure 7). Note, that the phrase segmentation stage is not affected by the choice of melodic distance measure and we therefore do not report the phrase evaluation metrics here.

First of all, it can be seen that all methods give satisfactory performance regarding establishment; they only differ at the second decimal digit. This is explained by the "mild" nature of the establishment measure: if a cluster is detected, even if it only contains one correct pattern and all the other patterns

are wrong, the cluster will still contribute successfully to the establishment f-measure.

Secondly, the DTW, EMD and EDIT methods appear to be competitive, also in the case of the occurrence f-measure, which is a "harder"' measure compared to establishment. Again, the performance of these three methods only differs at the second decimal digit. The SHAPE method, on the other hand, has inferior performance from this perspective. It is worth noting that the SHAPE method was proposed in the context of perfect transcriptions, but we are using it in a noisy environment due to the transcription errors that our system has to deal with. Therefore, these errors affect the quality of the interpolated curve that is synthesized by the SHAPE method and this in turn affects its performance. In the light of these findings and due to space restrictions, we select the EMD method for presenting further performance results, because of its slightly better performance, over all datasets, regarding the occurrence f-measure.

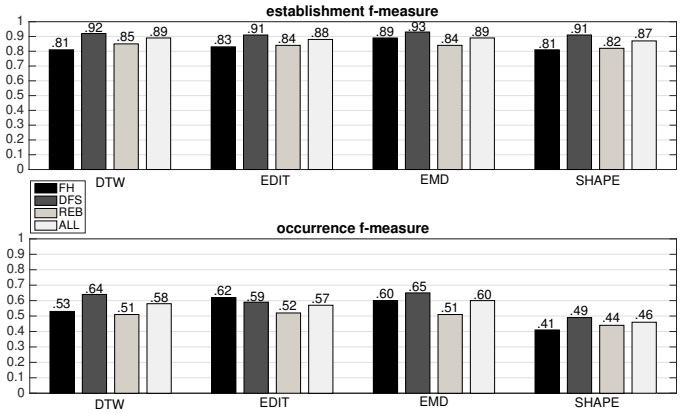

Fig. 7. Pattern evaluation metrics for different similarity measures.

C. Pattern discovery results across genres and folds

Table II shows the establishment and occurrence metrics for each one of the three datasets, when the system is evaluated as a whole using the EMD method, starting from automatic transcriptions and going through all the stages of the proposed pipeline. Again, it can be seen, that a higher performance is achieved for the DFS dataset compared to the FH and REB datasets. In particular with respect to the occurrence metrics, which evaluate the detection of all instances of a pattern, for the REB dataset, an F-measure of $\mathrm{Occ}_F = 0.51$ is achieved, which is low compared to the $\mathrm{Occ}_F = 0.65$ value of the DFS dataset. In a manual inspection of the results, we identified vocal detection errors and stronger melodic variation between instances of the same patterns as the main sources of error in the two polyphonic datasets. However, it has to be noted that we observe an improvement for the FH dataset ($\mathrm{Est}_F = 0.89$ and $\mathrm{Occ}_F = 0.60$), compared with the previously presented audio-based approach in [22] ($\mathrm{Est}_F = 0.60$ and $\mathrm{Occ}_F = 0.33$).

Table III shows the minimum value, maximum value, mean value and standard deviation of the three system parameters

TABLE II
ESTABLISHMENT AND OCCURRENCE MEASURES ACROSS GENRES

	FH	DFS	REB	ALL
Est-Pr	0.88	0.94	0.85	0.90
Est-Rec	0.90	0.92	0.86	0.90
Est-F	0.89	0.93	0.84	0.89
Occ-Pr	0.66	0.62	0.53	0.60
Occ-Rec	0.57	0.73	0.51	0.64
Occ-F	0.60	0.65	0.51	0.60

TABLE III
ESTIMATED PARAMETERS AND EVALUATION METRICS ACROSS FOLDS.

	min	max	mean	std
$\sigma_{\hat{\tau}}$	0.80	0.90	0.88	0.04
$\tau_{\min}$	1.82	1.95	1.86	0.05
$\tau_{\max}$	4.57	4.68	4.61	0.05
Phr_F	0.69	0.77	0.73	0.03
Est_F	0.86	0.92	0.89	0.03
Occ_F	0.54	0.66	0.60	0.05

($\tau_{\min}$, $\tau_{\max}$ and $\sigma_{\hat{\tau}}$) as well as of the three performance f-measures (Phr_F, Est_F and Occ_F) across the five folds. It can be seen that $\tau_{\min}$ and $\tau_{\max}$ vary within a small range of less than 0.2 seconds and $\sigma_{\hat{\tau}}$ was estimated between 0.8 and 0.9. Furthermore, the standard deviation of all three performance measures across folds lies within the second decimal digit.

D. Glass ceiling analysis

We now aim to discover how errors inserted by the different system components influence the overall system performance. First, in order to assess the limitations due to phrase segmentation errors, we repeat the previous experiment but use manually annotated phrase boundaries (M-Ph) instead of automatically detected phrases (A-Ph) to compute pair-wise distances with the EMD measure. In order to allow for a direct comparison, we preserve the 5-fold cross-validation scenario, even though the system parameters, which form part of the phrase segmentation stage, are not required when manually segmented phrases are used.

Figure 8 shows that the improvement when using manually annotated phrases is larger for the REB dataset ($\mathrm{Occ}_F = 0.51$ for A-Ph vs $\mathrm{Occ}_F = 0.86$ for M-Ph) than for the DFS ($\mathrm{Occ}_F = 0.65$ for A-Ph vs $\mathrm{Occ}_F = 0.84$ for M-Ph) and FH ($\mathrm{Occ}_F = 0.60$ for A-Ph vs $\mathrm{Occ}_F = 0.75$ for M-Ph) datasets. As mentioned in Section IV-A, we identified through manual inspection mistakenly transcribed melodic segments during instrumental interludes as the main cause for phrase segmentation errors. This issue occurs less frequently in the FH dataset, because the vocal detection stage of the respective transcription algorithm is tailored to the guitar accompaniment present in *flamenco* music. The vocal detection task is more complex in the REB dataset, because the instrumentation varies across recordings and no suitable genre-specific method exists for this type of music.

Furthermore, we observe that the establishment measures are close to 1.0 for M-Ph. These results are plausible, given that in folk music, all phrases are repeated at least once and

consequently, each detected phrase forms part of pattern. The only source of error in this scenario, with respect to pattern establishment, occurs if the clustering algorithm mistakenly creates clusters containing only a single instance. In this case, the respective phrase will not be considered part of any pattern, causing a decrease in Est_F. However, the phrase f-measure, which does not consider the clustering stage, results to 1.0 in this scenario.

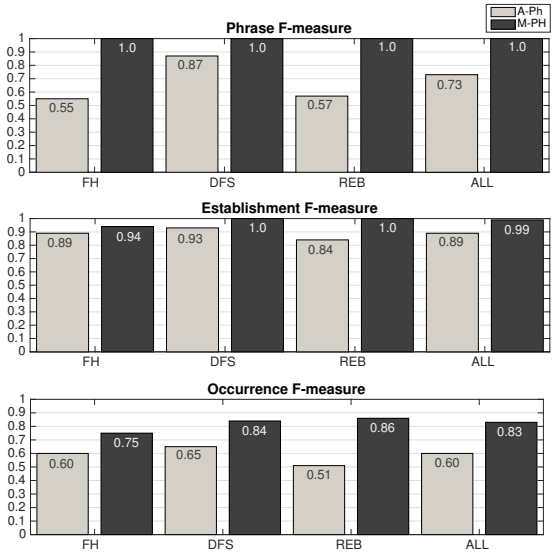

Fig. 8. Est_F (top) and Occ_F (bottom) for automatically segmented (light) and manually annotated (dark) phrases across datasets.

Finally, we assess the influence of automatic transcription errors on the performance of the FH dataset, for which flamenco experts have manually corrected the automatic transcriptions. We repeat the previous experiment, where MT denotes the use of manually corrected transcriptions and AT refers to automatic transcriptions. The results in Table IV show that when sing manual transcriptions instead of automatic ones, the phrase f-measure increases from $\mathrm{Phr}_F = 0.55$ to $\mathrm{Phr}_F = 0.78$. However, the overall performance limitation due to errors introduced in the phrase segmentation stage appears to be stronger ($\mathrm{Occ}_F = 0.66$ for MT-A-Ph vs $\mathrm{Occ}_F = 0.94$ for MT-M-Ph) than due to automatic transcription errors ($\mathrm{Occ}_F = 0.75$ for AT-M-Ph vs $\mathrm{Occ}_F = 0.94$ for MT-M-Ph). It can be furthermore seen, that with perfect transcription and phrase segmentation, the achieved occurrence F-measure is $Occ_F = 0.94$. In this scenario, the remaining error is introduced during the melodic distance computation and clustering stages.

E. Examples and qualitative error analysis

An example of an automatic annotation of repeated phrases from a *rebetiko* recording is shown in Figure 9. It can be seen that the detected phrase boundaries largely correspond to the manual annotations. Ground truth patterns 3 and 4 are correctly identified as pattern a and b respectively. Ground truth patterns 1 and 2 are covered by d and c with confusion in the cluster membership. Vocal detection errors have caused a false positive phrase to be detected at the end of the recording,

TABLE IV
ESTABLISHMENT AND OCCURRENCE F-MEASURE FOR THE FH DATASET FOR AUTOMATICALLY (A-PH) VS. MANUALLY ANNOTATED PHRASES (M-PH) AND AUTOMATIC (AT) VS. MANUAL TRANSCRIPTIONS (MT)

	Est_F	Occ_F	Phr_F
AT, A-Ph	0.89	0.60	0.55
AT, M-Ph	0.94	0.75	1.0
MT, A-Ph	0.91	0.66	0.78
MT, M-Ph	1.0	0.94	1.0

which was labeled as a. In this section of the song, the singer speaks, which caused the vocal detection algorithm to mistakenly classify this segment as voiced. Apart from inaccuracies caused by vocal detection errors, we observed a few frequently appearing error types which we briefly describe below.

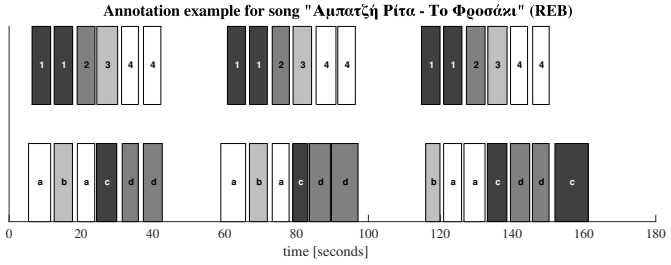

Fig. 9. First automatic annotation example, taken from REB: (top) ground truth and (bottom) annotated repetitions.

In various examples, we observed that the segmentation stage detected sub-phrases or a group of two phrases instead of the actual phrase boundary. This corresponds to a correct segmentation on a different hierarchical level in the song structure. An example is shown Figure 10. In this particular example, the low intensity of the vocals with respect to the accompaniment caused several notes inside the vocal phrases to be missed by the transcription algorithm. Consequently, the accumulated note duration was estimated at a low value, causing a higher segmentation score when two adjacent phrases were fused. Here, it can furthermore be observed that the same vocal detection errors cause missing vocal segments (around sec. 130).

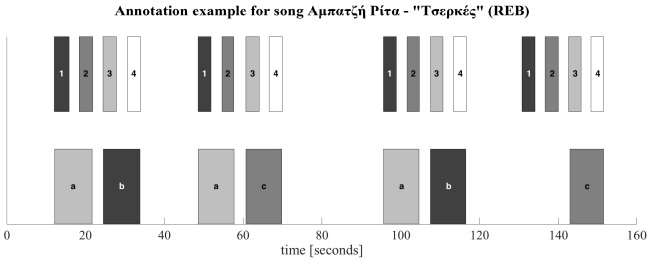

Fig. 10. Second automatic annotation example, taken from REB: (top) ground truth and (bottom) annotated repetitions.

In addition, we observed some clustering errors, where two or more patterns are mistakenly grouped together. This scenario occurs in particular when one pattern is very dis-

similar from a group of patterns which have at least a basic melodic movement in common. An example is shown in Figure 11 where ground truth patterns 1, 2 and 3 mistakenly received the common label a. Similarly, clustering errors occur when patterns show only minor melodic differences which are nevertheless perceptually significant due to different harmonic progressions in the accompaniment. This is the case for ground truth patterns 1 and 2 from the example shown in Figure 9.

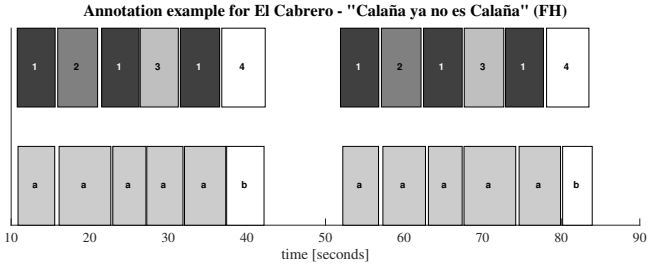

Fig. 11. Third automatic annotation example, taken from FH: (top) ground truth and (bottom) annotated repetitions.

F. Relation to pattern discovery and structural segmentation

In a final experiment, we investigate the suitability of state of the art audio-based pattern detection and structural segmentation algorithms for the task of detecting repeated sung phrases. To this end, we evaluate the algorithms listed below on our datasets and analyse their behaviour on three selected examples, after adapting available system parameters to the characteristics of this particular task. Furthermore, we compare the proposed system to a baseline method which operates on automatic transcriptions but performs a naive segmentation and label assignment.

- The audio-based pattern detection algorithm described in [50] (NF14). The tempo of each song, which is required by the algorithm, was estimated using the method described in [51]. All audio-files are resampled to 11025 Hz and the short-term spectrogram is extracted using a 290 ms long moving window, with an overlap of 50% between successive windows. This stage is followed by a constant-Q transform. As defined by the authors, the path score threshold parameter, θ, is set to the value of 0.3 and the number of eliminated diagonals, ρ, around a detected path in the similarity matrix is set to the value of 2. At the output of the system, instances which temporally overlap with other instances of the same pattern are discarded.
- The audio-based pattern detection method proposed in [20] (WD15). The method allows to specify a minimum pattern duration, which we set to τ_{min}. After downsampling the recordings to 11025 Hz, the short-term spectrogram is computed with a moving window, 743 ms long (11 ms hop size), followed by a constant-Q transform covering the frequency range from 27.5 Hz to 55125.5 Hz. The grid for determining the optimal VMO threshold, θ, ranges from 0.0 to 2.0 in steps of 0.01, as specified by the authors. We discard instances which overlap with other instances of the same pattern and extract the required tempo estimate with the method in [51].
- The structural segmentation approach proposed in [52] (WB10) which segments an audio recording into adjacent sections and assigns the same label to repeated sections. The extraction of the beat-synchronous chromagram follows the method and parameters in [53]. As recommended by the authors, the number of basis patterns is set to $K = 4$ and the minimum number of segments at the output is set to the value of 3.
- The method presented in [54] (MJG13), which detects repeated segments in audio recordings. After resampling all recordings to 22050Hz, the Chroma Energy Normalised Statistics are extracted with a feature rate of 2Hz. As recommended by the authors, the fitness tolerance parameter is set to $\delta = -2$ and the relative threshold to suppress small values in the self-similarity matrix is set to $\rho = 0.15$. The algorithm allows to specify a lower and upper bound for the allowed segment duration, which we set to τ_{min} and τ_{max}, respectively.
- A naive baseline method (BL), which segments the note transcriptions at the note offsets closest to multiples of the median phrase duration (3.11 sec, computed over ALL). The resulting segments are assigned one out of k random labels, where $k = 4$ is the median number of clusters in the annotated ground truth data.

We furthermore compared the performance of the proposed method (denoted as PR), using the generic cross-fold validation setup, to our prior approach [22] (PKDM15) which was developed in the context of accompanied flamenco singing. Figure 12 shows the output of the different algorithms for three audio examples. The manually annotated ground truth is denoted GT. All system parameters, features and preprocessing methods not mentioned above, were applied as described in the respective publications.

As described in Section I, pattern detection aims at discovering repetition in a music recording without restrictions regarding instrumentation or type of structural units. As a result, some of the repetitions detected by NF14 and WD15 are located in instrumental sections and others are located in singing sections but do not coincide with phrase start and end boundaries. Structural segmentation aims at segmenting the entire recording into labeled sections, where regions with the same labels are considered similar to each other. In the second example taken from the REB collection, the algorithm WB10 does assign a common label to two long vocal segments which are indeed instances of the same melody, but those segments span over more than two phrases. The algorithm also detects a repeated instrumental interlude between adjacent vocal segments. In the two accompanied examples taken from the FH and REB collections, the MJG13 algorithm detects repeated chord sections in the accompaniment.

In summary, we can conclude that while these four algorithms do detect repetition in the audio examples, the extracted patterns do generally not coincide with sung phrases. These findings also become apparent when analysing the establishment and occurrence f-measures for the different algorithms on all four datasets. Specifically, in Table V it can be seen

that the proposed approach outperforms the baseline method for the task of detecting repeated sung phrases. In addition, among the referenced algorithms, NF14 yields the best results with $Est_F = 0.47$ and $Occ_F = 0.15$ on ALL. It can also be observed that the proposed method yields an improvement compared to PKDM, on both the flamenco dataset for which PKDM was originally proposed ($Est_F = 0.60$ and $Occ_F = 0.33$), as well as on the union of all three datasets ($Est_F = 0.42$ and $Occ_F = 0.15$). The note-based BL method yields a high establishment score ($Est_F = 0.75$) which outperforms the four referenced algorithms in this Section, but not the method we propose. This can be explained by the fact that, despite the naive segmentation and labelling process, the BL method benefits from the transcription system: since no notes are transcribed during long instrumental sections at the beginning and end of the song, the BL method does not detect any patterns in those parts of the recording. It is however possible, that it detects patterns which span over interludes between vocal sections. This behaviour is apparent in the first two examples of Figure 12. However, it is important to note, that the random labelling scheme of the BL methods, results in a low occurrence score ($Occ - F = 0.17$).

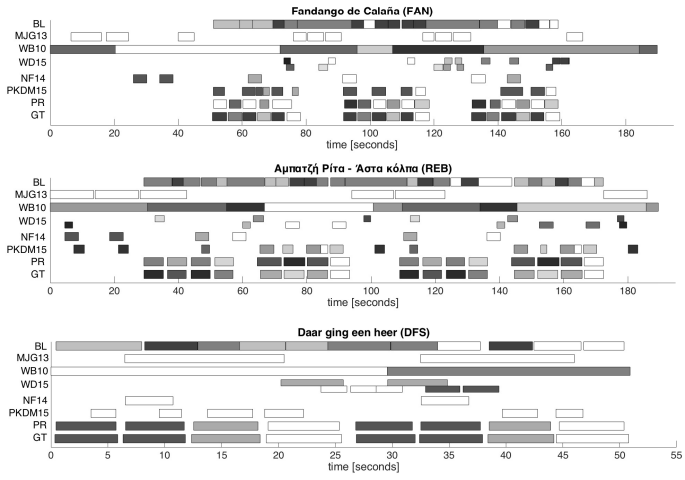

Fig. 12. Ground truth patterns (GT), output of the proposed algorithm (PR) and of three referenced systems (MJG13 [54], WHD15 [20], NF14 [50], WB10 [52] and PKDM15 [22]) for three audio examples. Shading indicates the cluster membership.

V. Conclusions

We presented a complete system for detecting repeated sung phrases in folk music recordings, starting from the audio signal. A novel phrase segmentation algorithm was proposed, which operates on automatically generated note-level transcriptions. We investigated a variety of melodic distance measures to compute pair-wise distances among phrases and, by means of a standard clustering algorithm, clusters of similar phrases were discovered. Members of a cluster were interpreted as instances of the same repeated pattern.

In a detailed evaluation procedure, we assessed the performance of the proposed method on three genres with distinct musical characteristics. In a glass ceiling analysis, we demonstrated that phrase segmentation errors cause a stronger performance limitation than transcription inaccuracies. Furthermore, we provided various examples of automatic annotations and analysed frequently occurring errors. The proposed system outperforms the only existing audio-based approach to repeated phrase discovery in folk music recordings and, we have shown that state-of-the-art methods for the related tasks of audio-based pattern detection and music segmentation are not suitable for this task, despite non-exhaustive search for the optimal settings per algorithm. Manual annotations and the source code of the system have been made available for the sake of reproducibility of research results. In the future, in the context of promoting and disseminating this method, we will include manual annotations from more experts.

TABLE V
Comparative evaluation of pattern discovery and structural segmentation: Establishment and occurrence f-measure.

	PR	PKDM15	NF14	WHD15	WB10	MJG13	BL
FH							
Est-F	**0.89**	0.60	0.47	0.25	0.33	0.22	0.73
Occ-F	**0.60**	0.33	0.11	0.02	0.03	0.05	0.16
DFS							
Est-F	**0.93**	0.38	0.50	0.42	0.17	0.15	0.75
Occ-F	**0.65**	0.13	0.16	0.07	0.01	0.00	0.16
REB							
Est-F	**0.84**	0.43	0.40	0.50	0.24	0.14	0.75
Occ-F	**0.51**	0.08	0.12	0.07	0.00	0.01	0.18
ALL							
Est-F	**0.89**	0.42	0.47	0.45	0.22	0.16	0.75
Occ-F	**0.60**	0.15	0.15	0.06	0.01	0.01	0.17

Acknowledgment

This research has been partly funded by the Junta de Andalucía and the FEDER funds of the European Union, project COFLAII under the grant identifier P12-TIC-1362. The authors thank Inmaculada Morales for providing manual ground truth transcriptions.

References

[1] P. van Kranenburg, M. de Bruin, L. P. Grijp, and F. Wiering, "The Meertens tune collections," Meertens Online Reports, Tech. Rep. 2014-1, 2014.
[2] C. W. Walton, *Basic forms in music*. Alfred Music, 2005.
[3] E. H. Margulis, "Musical repetition detection across multiple exposures," *Music Perception: An Interdisciplinary Journal*, vol. 29, no. 4, pp. 377–385, 2012. [Online]. Available: http://mp.ucpress.edu/content/29/4/377
[4] D. Temperley, *The cognition of basic musical structures*. MIT press, 2004.
[5] I. L. Bradley, "Repetition as a factor in the development of musical preferences," *Journal of Research in Music Education*, vol. 19, no. 3, pp. 295–298, 1971. [Online]. Available: https://doi.org/10.2307/3343764
[6] A. Novello, M. F. McKinney, and A. Kohlrausch, "Perceptual evaluation of music similarity." in *ISMIR*, 2006, pp. 246–249.
[7] A. Lamont and N. Dibben, "Motivic structure and the perception of similarity," *Music Perception: An Interdisciplinary Journal*, vol. 18, no. 3, pp. 245–274, 2001.
[8] A. Lomax, *Folk song style and culture*. Transaction Publishers, 1968, vol. 88.
[9] P. Boot, A. Volk, and W. B. de Haas, "Evaluating the role of repeated patterns in folk song classification and compression," *Journal of New Music Research*, vol. 45, no. 3, pp. 223–238, 2016.

[10] M. Li, Z. Zhao, and P. Shi, "Query by humming based on music phrase segmentation and matching," in *Proceedings of the12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2015, pp. 1966–1970.

[11] B. Janssen, W. B. De Haas, A. Volk, and P. van Kranenburg, "Finding repeated patterns in music: State of knowledge, challenges, perspectives," in *Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research*, 2013, pp. 277–297.

[12] M. E. Rodríguez-López and A. Volk, "Location constraints for repetition-based segmentation of melodies," in *Proceedings of the International Conference on Mathematics and Computation in Music*. Springer International Publishing, 2015, pp. 73–84.

[13] M. Müller, *Fundamentals of Music Processing*. Springer, Cham, 2015, ch. Music Structure Analysis.

[14] R. B. Dannenberg and N. Hu, "Pattern discovery techniques for music audio," *Journal of New Music Research*, vol. 32, no. 2, pp. 153–163, 2003.

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

[16] M. Müller and P. Grosche, "Automated segmentation of folk song field recordings," in *Proceedings of the 10th ITG Speech Communication Symposium*. VDE, 2012.

[17] C. Bohak and M. Marolt, "Probabilistic segmentation of folk music recordings," *Mathematical Problems in Engineering*, vol. 2016, p. 11, 2016. [Online]. Available: http://dx.doi.org/10.1155/2016/8297987%]8297987

[18] S. Gulati, J. Serra, V. Ishwar, and X. Serra, "Mining melodic patterns in large audio collections of Indian art music," in *Proceedings of the 10th International Conference on Signal-Image Technology and Internet-Based Systems (SITIS)*, 2014, pp. 264–271.

[19] T. Collins, S. Böck, F. Krebs, and G. Widmer, "Bridging the audio-symbolic gap: The discovery of repeated note content directly from polyphonic music audio," in *Proceedings of the 53rd International Audio Engineering Society Conference on Semantic Audio*. Audio Engineering Society, 2014.

[20] C. Wang and S. Dubnov, "Pattern discovery from audio recordings by variable Markov oracle: A music information dynamics approach," in *Proceedings of the IEEE Conference on Audio, Speech and Signal Processing (ICASSP)*, 2015.

[21] P. van Kranenburg and G. Tzanetakis, "A computational approach to the modeling and employment of cognitive units of folk song melodies using audio recordings," in *Proceedings of the 11th International Conference on Music Perception and Cognition*, 2010, pp. 794–797.

[22] N. Kroher, A. Pikrakis, J. Moreno, and J. M. Díaz-Báñez, "Discovery of repeated vocal patterns in polyphonic audio: a case study on flamenco music." in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Nice, France, 2016.

[23] N. Kroher, J. M. Díaz-Báñez, J. Mora, and E. Gómez, "Corpus COFLA: A research corpus for the computational study of flamenco music." *ACM Journal on Computing and Cultural Heritage.*, vol. 10:1-10:21, 2016.

[24] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: challenges and future directions," *Journal of Intelligent Information Systems*, vol. 41, pp. 407–434, 2013.

[25] M. Mauch, C. Cannam, R. Bittner, G. Fazekas, J. Salamon, J. Dai, J. Bello, and S. Dixon, "Computer-aided melody note transcription using the Tony software: Accuracy and efficiency," in *Proceedings of the First International Conference on Technologies for Music Notation and Representation*, May 2015.

[26] N. Kroher and E. Gómez, "Automatic transcription of flamenco singing from polyphonic music recordings," *IEEE Transactions on Audio, Speech and Language Processing.*, vol. 24, no. 5, pp. 901–913, 2015.

[27] L. Song, M. Li, and Y. Yan, "Automatic vocal segments detection in popular music." in *Proceedings of the Ninth International Conference on Computational Intelligence and Security*, 2013.

[28] A. Pikrakis, Y. Kopsinis, N. Kroher, and J. M. Díaz-Báñez, "Unsupervised singing voice detection using dictionary learning," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1212–1216.

[29] L. E. McCullough, "Style in traditional Irish music," *Ethnomusicology*, vol. 21, no. 1, pp. 85–97, 1977.

[30] F. Lerdahl and R. Jackendoff, *A generative theory of tonal music*. MIT press, 1985.

[31] W. J. Dowling, "Rhythmic groups and subjective chunks in memory for melodies," *Attention, Perception, & Psychophysics*, vol. 14, no. 1, pp. 37–40, 1973.

[32] M. Rodríguez-López and A. Volk, "Symbolic segmentation: a corpus-based analysis of melodic phrases," in *Proceedings of the International Symposium on Computer Music Modeling and Retrieval*. Springer, 2013, pp. 548–557.

[33] K. Frieler, "Exploring phrase form structures part 1: European folk songs." in *Proceedings of the International Workshop on Folk Music Analysis (FMA)*, 2014.

[34] D. Müllensiefen and K. Frieler, "Evaluating different approaches to measuring the similarity of melodies," in *Data Science and Classification*. Springer, 2006, pp. 299–306.

[35] V. Velardo, M. Vallati, and S. Jan, "Symbolic melodic similarity: State of the art and future challenges," *Computer Music Journal*, vol. 40, no. 2, pp. 70–83, 2016.

[36] A. Kotsifakos, P. Papapetrou, J. Hollmén, D. Gunopulos, and V. Athitsos, "A survey of query-by-humming similarity methods," in *Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA)*. ACM, 2012, pp. 5.1–4.

[37] J. B. Kruskall and M. Liberman, *Time warps, string edits and macromolecules*. Addison, 1983, ch. The symmetric time warping algorithm: From continuous to discrete.

[38] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.

[39] P. van Kranenburg, "A computational approach to content-based retrieval of folk song melodies," Ph.D. dissertation, Utrecht University, 2010.

[40] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[41] R. J. McNab, L. A. Smith, I. H. Witten, C. L. Henderson, and S. J. Cunningham, "Towards the digital music library: Tune retrieval from acoustic input," in *Proceedings of the first ACM international conference on Digital libraries*. ACM, 1996, pp. 11–18.

[42] M. Mongeau and D. Sankoff, "Comparison of musical sequences," *Computers and the Humanities*, vol. 24, no. 3, pp. 161–175, 1990.

[43] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Proceedings of the 6th International Conference on Computer Vision*, 1998, pp. 59–66.

[44] R. Typke, P. Giannopoulos, R. C. Veltkamp, F. Wiering, R. Van Oostrum *et al.*, "Using transportation distances for measuring melodic similarity." in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2003.

[45] R. Typke, F. Wiering, and R. C. Veltkamp, "Transportation distances and human perception of melodic similarity," *Musicae Scientiae*, vol. 11, no. 1, pp. 153–181, 2007.

[46] J. Urbano, J. Lloréns, J. Morato, and S. Sánchez-Cuadrado, "Melodic Similarity through Shape Similarity," in *Exploring Music Contents*, S. Ystad, M. Aramaki, R. Kronland-Martinet, and K. Jensen, Eds. Springer, 2011, pp. 338–355.

[47] L. Kaufman and P. J. Rousseeuw, *Statistical Data Analysis Based on the L1-Norm and Related Methods*. North-Holland, 1987, ch. Clustering by means of Medoids, pp. 405–416.

[48] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[49] D. Turnbull, G. Lanckriet, E. Pampalk, and M. Goto, "A supervised approach for detecting boundaries in music using difference features and boosting," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2007.

[50] O. Nieto and M. M. Farbood, "Identifying polyphonic patterns from audio recordings using music segmentation techniques," in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[51] D. P. Ellis, "Beat tracking by dynamic programming," *Journal of New Music Research*, vol. 36, no. 1, pp. 51–60, 2007.

[52] R. J. Weiss and J. P. Bello, "Identifying repeated patterns in music using sparse convolutive non-negative matrix factorization," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2010.

[53] D. P. Ellis and G. E. Poliner, "Identifying cover songs with chroma features and dynamic programming beat tracking," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–1429.

[54] M. Müller, N. Jiang, and P. Grosche, "A robust fitness measure for capturing repetitions in music recordings with applications to audio thumbnailing," *IEEE Transactions on audio, speech, and language processing*, vol. 21, no. 3, pp. 531–543, 2013.

Nadine Kroher Nadine Kroher received a MSc in Audio and Electrical Engineering from Graz University of Technology (Austria) in 2011 and a MSc in Sound and Music Computing from Universitat Pompeu Fabra (Spain). Currently, she is a PhD student at the Department of Applied Mathematics, University of Seville (Spain). Her research focuses on pattern detection and machine learning for audio signals and Music Information Retrieval for computational ethnomusicology.

Aggelos Pikrakis Aggelos Pikrakis (Eng., Ph.D.) is currently an Assistant Professor with the Department of Informatics at the University of Piraeus, Greece. His research interests lie in the areas of machine learning for audio, speech and music signals. He has co-authored two books with Academic Press in the areas of pattern recognition and audio analysis and has published his research findings in international journals and conferences, including IEEE Transactions on ASLP, IEEE Transactions Multimedia and IEEE ICASSP. He is also an Associate Editor of the EURASIP Journal on Advances in Signal Processing.

José-Miguel Díaz-Báñez José Miguel Díaz-Báñez received the Ph.D. degree in Mathematics from University of Seville in 1998 and serves as Professor of Applied Mathematics at the School of Engeniering, University of Seville. His primary research area is Computational Geometry and Optimization, mainly applied to problems in aerial robotics and music technology. He authored more than 200 publications, including 55 journal papers. He is the coordinator of the GALGO group (Geometric Algorithms) and the COFLA Project (Computational Analysis of Flamenco Music).