ENHANCING SIMILARITY MATRICES FOR MUSIC AUDIO ANALYSIS

Meinard Müller, Frank Kurth

Department of Computer Science III, University of Bonn
Römerstr. 164, D-53117 Bonn, Germany
{meinard, frank}@cs.uni-bonn.de

ABSTRACT

Similarity matrices have become an important tool in music audio analysis. However, the quadratic time and space complexity as well as the intricacy of extracting the desired structural information from these matrices are often prohibitive with regard to real-world applications. In this paper, we describe an approach for enhancing the structural properties of similarity matrices based on two concepts: first, we introduce a new class of robust and scalable audio features which absorb local temporal variations. As a second contribution, we then incorporate contextual information into the local similarity measure. The resulting enhancement leads to significant reduction in matrix size and also eases the structure extraction step. As an example, we sketch the application of our techniques to the problems of audio summarization and audio synchronization, obtaining effective and computationally feasible algorithms.

1. INTRODUCTION

The concept of similarity matrices has been introduced to the music context by Foote in order to visualize the time structure of audio and music [1]. The general idea is as follows: given two audio data streams, one first transforms them into sequences $\vec{V} := (\vec{v}^1, \vec{v}^2, \ldots, \vec{v}^N)$ and $\vec{W} := (\vec{w}^1, \vec{w}^2, \ldots, \vec{w}^M)$ of feature vectors $\vec{v}^n \in \mathcal{F}$, $1 \leq n \leq N$, and $\vec{w}^m \in \mathcal{F}$, $1 \leq m \leq M$. Here, $\mathcal{F}$ denotes a suitable feature space, e.g., a space of spectral, MFCC, or chroma vectors. Based on a suitable similarity measure $d : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$, one can form a *similarity matrix* $\mathcal{S} = (d(\vec{v}^n, \vec{w}^m))_{nm}$ by pairwise comparison of the features $\vec{v}^n$ and $\vec{w}^m$. In case that $\vec{V} = \vec{W}$, the resulting matrix is also referred to as *self-similarity matrix*.

Similarity matrices have proven to be a valuable tool in audio analysis. In Sect. 3, we address two such analysis tasks: *audio summarization* and *audio synchronization*. The underlying principle is that similar segments are revealed as paths along diagonals in the corresponding similarity matrix. As an example, we consider the first 94 seconds of an Ormandy interpretation of Brahms' Hungarian Dance No. 5, having the musical form $A_1A_2B_1B_2$ (segment A_2 being a repetition of A_1 and B_2 being a repetition of B_1). The self-similarity matrix (with respect to some suitable audio features), shown in Fig. 1, reveals this structure: the path in the lower left corner indicates that the segment between 1 and 22 is similar to the segment between 22 and 42 (measured in seconds), whereas the curved path in the upper right corner indicates that the segment between 42 and 69 is similar to the segment between 69 and 89. Note that in the Ormandy interpretation, the tempo of B_2 is much faster than that of B_1, which is revealed by the gradient of the path encoding the relative tempo difference between the two segments.

There are two major problems in music audio analysis based on similarity matrices: the first problem concerns the robust extraction of suitable paths revealing the structural similarity relations between the underlying audio streams. So far, this problem has been studied under the constant tempo assumption, which typically holds for pop music, see Sect. 3.1 for references. For the case, however, that musically similar segments exhibit significant local tempo variations—as often holds for Western classical music—there are yet no effective and efficient solutions. The second problem concerns the high time and space complexity $O(NM)$ to compute and store the similarity matrices, which makes the usage of similarity matrices infeasible for large N and M. Here, reducing the number N and M by simply increasing the feature analysis window often destroys the structural properties of the similarity matrices, see Fig. 5.

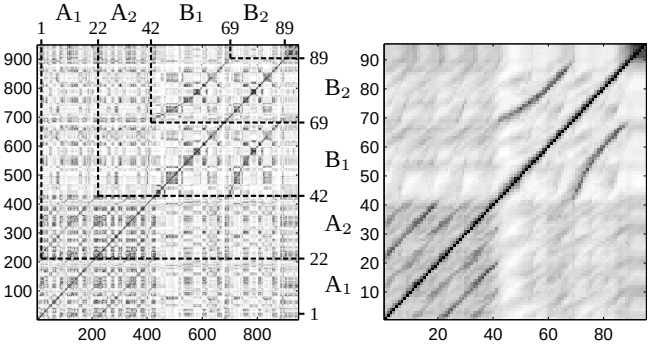

Fig. 1. Self-similarity matrices of the first 94 seconds of an Ormandy interpretation of Brahm's Hungarian Dances No. 5. The musical form $A_1A_2B_1B_2$ is revealed by the path structure. The left side shows a matrix with a feature sampling rate of 10 Hz. The right side shows an enhanced similarity matrix ($\mathcal{S}^{\min}_{10,2}(21, 5)$) with a feature sampling rate of 1 Hz.

In this paper, we suggest an approach for enhancing the path structure of similarity matrices, which constitutes an important step towards a solution of the above mentioned problems. In particular, we cope with the delicate tradeoff between needing coarse and robust features on the one hand and requiring sufficient flexibility to deal with local tempo variations on the other hand. Our basic idea towards finding a good tradeoff can be summarized as follows. Instead of relying on one single mechanism, we take care of the temporal variations on various levels simultaneously: on the "feature level" (using statistical features to absorb micro-variations), on the "local distance measure level" (including flexible contextual information to account for local variations) as well as on the "path extraction level" (accounting for coarse global time variations). In Sect. 2, we describe this approach in detail and apply the techniques to the class of chroma features. In Sect. 3, we then sketch the impact of our matrix enhancement techniques to the problems of music summarization

and (multiresolution-based) audio synchronization. Suitable references to the related work are given in each section. Further results and examples can be found at `www-mmdb.iai.uni-bonn.de/projects/simmat`.

2. ENHANCEMENT TECHNIQUES

The properties of a similarity matrix $\mathcal{S}$ depend on the kind of audio features extracted from the audio data streams as well as on the (local) similarity measure d. In this section, we describe some techniques for the design of robust and scalable features (Sect. 2.2) and for the enhancement of similarity measures (Sect. 2.3 and Sect. 2.4) in order to amplify the path structure of $\mathcal{S}$. As an example, these techniques are applied to the class of chroma features (Sect. 2.1).

2.1. Chroma Features

In the first stage, each audio signal is converted into a sequence of acoustic features, e.g., spectral, MFCC, or chroma features. In the following, we consider the case of *chroma features* as suggested by [2], which represent the spectral energy contained in each of the 12 traditional pitch classes of the equal-tempered scale. More specifically, we decompose the audio signal into 88 frequency bands corresponding to the musical notes A0 to C8 (MIDI pitches 21 to 108) using a suitable multirate filter bank. We then take the short-time mean-square power (STMSP) for each of the 88 subbands by convolving the squared subband signals with a rectangular window corresponding to 200 ms with a 50% overlap. Adding up the corresponding STMSPs of all pitches belonging to the same chroma class yields a real 12-dimensional vector $\vec{v} = (v_1, \ldots, v_{12})$ for each analysis window. Finally, we normalize each chroma vector by replacing $\vec{v}$ by $\vec{v}/(\sum_{i=1}^{12} v_i)$. The resulting sequence of 12-dimensional feature vectors expresses the local energy distribution in the 12 chroma classes and strongly correlates to the harmonic progression of the audio signal, see [3]. The resulting feature sampling rate of 10 Hz will constitute the finest resolution level, chosen to be sufficient in view of our intended applications.

2.2. Designing Robust and Scalable Features

For enhancing the similarity matrix, a flexible and computationally inexpensive procedure is needed to adjust the feature resolution. Instead of simply modifying the analysis window during the above feature computation, we introduce a second, much larger statistics window and consider *short-time statistics* concerning the chroma energy distribution over this window. More specifically, let $Q : [0,1] \to \{0,1,2,3,4\}$ be a quantization function defined by $Q(a) := 0$ for $a \in [0, 0.05)$, $Q(a) := 1$ for $a \in [0.05, 0.1)$, $Q(a) := 2$ for $a \in [0.1, 0.2)$, $Q(a) := 3$ for $a \in [0.2, 0.4)$, and $Q(a) := 4$ for $a \in [0.4, 1]$. Then, we quantize each chroma energy distribution vector $\vec{v}^n = (v_1^n, \ldots, v_{12}^n) \in [0,1]^{12}$ by applying Q to each component of $\vec{v}^n$, yielding $Q(\vec{v}^n) := (Q(v_1^n), \ldots, Q(v_{12}^n))$. Intuitively, this quantization assigns the value 4 to a chroma component v_i^n if the corresponding chroma class contains more than 40 percent of the signal's total energy and so on. The thresholds are chosen in a logarithmic fashion. Furthermore, chroma components below a 5 percent threshold are excluded from further considerations. In a subsequent step, we convolve the sequence $(Q(\vec{v}^1), \ldots, Q(\vec{v}^N))$ component-wise with a Hann window of length $w \in \mathbb{N}$. This again results in a sequence of 12-dimensional vectors with non-negative entries, representing a kind of weighted statistics of the energy distribution over a window of w consecutive vectors. In a last step, this sequence is downsampled by a factor of q. The resulting vectors are normalized with respect to the Euclidean norm. For example, if $w = 41$ and $q = 10$, one obtains one feature vector per second each corresponding to roughly 4100 ms of audio. The resulting feature sequence will be referred to as CENS(w, q) (**C**hroma **E**nergy distribution **N**ormalized **S**tatistics) sequence. Similar features have been applied in the audio matching scenario, see [3].

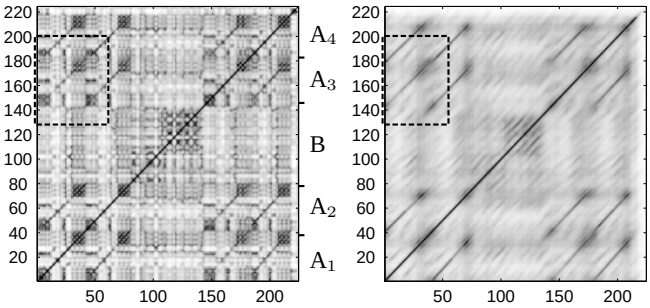

Fig. 2. Similarity matrices $\mathcal{S}(41, 10)$ (left) and $\mathcal{S}_{10}(41, 10)$ (right) of Shostakovich's Waltz 2, Jazz Suite No. 2, interpreted by Chailly.

By modifying the parameters w and q, we may adjust the feature granularity and sampling rate without repeating the cost-intensive computations in Sect. 2.1. Furthermore, by changing the thresholds and values of the quantization function Q one can enhance or mask out certain aspects of the audio signal, e.g., making the CENS features insensitive to noise components that may arise during note attacks. Finally, taking statistics over relatively large windows not only smooths out micro-temporal deviations, as may occur for articulatory reasons, but also compensates for different realizations of note groups such as trills or arpeggios.

In the following discussion, we use the similarity measure d defined by $d(\vec{v}, \vec{w}) := 1 - \langle \vec{v}, \vec{w} \rangle$ for CENS feature vectors $\vec{v}, \vec{w} \in [0,1]^{12}$ and for fixed parameters w and q. Since $\vec{v}$ and $\vec{w}$ are normalized, the inner product $\langle \vec{v}, \vec{w} \rangle$ coincides with the cosine of the angle between $\vec{v}$ and $\vec{w}$. For short, the resulting similarity matrix will also be denoted by $\mathcal{S}(w, q)$.

2.3. Including Contextual Information

In order to keep the size of the similarity matrix manageable, one often has to drastically reduce the feature sampling rate. This can lead to a heavily deteriorated similarity matrix when simply enlarging the feature analysis window (with a fixed overlap ratio), see Fig. 5. To alleviate the loss in quality, we incorporate some contextual information into the local similarity measure. A similar approach has been suggested in [4], where HMM-based "dynamic" features are used, which model the temporal evolution of the spectral shape over a fixed time duration. For the case of CENS features, the following simple procedure has proven to be a flexible and effective method. (For the moment, we assume constant tempo and then describe in Sect. 2.4 how to get rid of this assumption.) We define the *contextual similarity measure* d_L by

$$d_L(n, m) := \frac{1}{L} \sum_{\ell=0}^{L-1} d(\vec{v}^{n+\ell}, \vec{w}^{m+\ell})$$

for some length parameter $L \in \mathbb{N}$, $1 \leq n \leq N - L + 1$, $1 \leq m \leq M - L + 1$. By suitably extending the CENS sequences $(\vec{v}^1, \ldots, \vec{v}^N)$ and $(\vec{w}^1, \ldots, \vec{w}^M)$, e.g., via zero-padding, one may extend the definition to $1 \leq n \leq N$ and $1 \leq m \leq M$. Then, the

contextual similarity matrix $\mathcal{S}_L$ is defined by $\mathcal{S}_L = (d_L(n,m))_{nm}$. In this matrix, a value $d_L(n,m) \in [0,1]$ close to zero implies that the entire L-sequence $(\vec{v}_n, \ldots, \vec{v}_{n+L-1})$ is similar to the L-sequence $(\vec{w}_n, \ldots, \vec{w}_{n+L-1})$, resulting in an enhancement of the diagonal path structure in the similarity matrix. This is also illustrated by Fig. 2, showing $\mathcal{S}$ and $\mathcal{S}_{10}$ for a Chailly interpretation of Shostakovich's Waltz 2, Jazz Suite No. 2. In this piece, the theme appears four times (A_1, A_2, A_3, A_4) each time in a modified form, e.g., concerning instrumentation and articulation. Here, the diagonal path structure of $\mathcal{S}_{10}$—opposed to the one of $\mathcal{S}$—is prominent which not only facilitates the extraction of structural information but also the reduction of the feature sampling rate.

Finally, note that the contextual similarity matrix $\mathcal{S}_L$ can be efficiently computed from $\mathcal{S}$ by applying an averaging filter along the diagonals. More precisely, $\mathcal{S}_L(n,m) = \frac{1}{L}\sum_{\ell=0}^{L-1} \mathcal{S}(n+\ell, m+\ell)$ (with a suitable zero-padding of $\mathcal{S}$).

2.4. Incorporating Flexibility towards Local Tempo Variations

So far, we have enhanced similarity matrices by regarding the context of L consecutive features vectors. This procedure is problematic when similar segments do not have the same tempo. Such a situation frequently occurs in classical music—even within the same interpretation—as is shown by the Brahms example of Fig. 1, where the theme in A_1 (seconds 42 to 69) is repeated in a much faster and increasing tempo in A_2 (seconds 69 and 89). To account for such variations we, intuitively spoken, create several versions of one of the audio data streams corresponding to different tempi, which are then incorporated into one single similarity matrix. We now describe the procedure in detail.

Recall from Sect. 2.2 that the two audio data streams are transformed into CENS(w,q) sequences $\vec{V}(w,q)$ of length $N(w,q)$ and $\vec{W}(w,q)$ of length $M(w,q)$, where the dependency on the window size w and the downsampling factor q is indicated in terms of the arguments. For the sake of concreteness we chose $w = 41$ and $q = 10$, resulting in a feature sampling rate of 1 Hz. We now simulate a tempo change of the second data stream by changing the values of w and q. For example, using a window size of $w = 53$ (instead of 41) and a downsampling factor of $q = 13$ (instead of 10) simulates a tempo change of the original data stream by a factor of $10/13 \approx 0.77$. In our experiments, we used 8 different tempi as indicated by Table 1, covering tempo variations of roughly -40 to $+40$ percent. We then define a new similarity measure $d_L^{\min}$ by

$$d_L^{\min}(n,m) = \min_{(w,q)} \frac{1}{L} \sum_{\ell=0}^{L-1} d\Big(\vec{v}(41,10)^{n+\ell}, \vec{w}(w,q)^{\lceil m\cdot 10/q\rceil+\ell}\Big),$$

where the minimum is taken over the pairs (w,q) from Table 1. In other words, at position (n,m), the L-subsequence of $\vec{V}(41,10)$ starting at absolute time n (note that the feature sampling rate is 1 Hz) is compared with the L-subsequence of $\vec{W}(w,q)$ (simulating a tempo change of $10/q$) starting at absolute time m (corresponding to feature position $\lceil m \cdot 10/q \rceil$). From this we obtain the modified contextual similarity matrix $\mathcal{S}_L^{\min} = (d_L^{\min}(\vec{v}^n, \vec{w}^m))_{nm}$. Fig. 3 shows that incorporating local tempo variations into contextual similarity matrices significantly improves the quality of the path structure, in particular for the case that similar audio segments exhibit different local relative tempi.

Altogether, we have introduced a combination of techniques that enhance the structural properties of similarity matrices. Introducing contextual information, expressed by the parameter L, may allow to further decrease the matrix size by a subsequent downsampling operation without a significant loss of structural quality, see Fig. 5. We express this operation by an additional downsampling parameter $p \in \mathbb{N}$ and denote the resulting similarity matrix by $\mathcal{S}_{L,p}^{\min}(w,q)$.

w	29	33	37	**41**	45	49	53	57
q	7	8	9	**10**	11	12	13	14
tc	1.43	1.25	1.1	**1.0**	0.9	0.83	0.77	0.7

Table 1. Tempo changes (tc) simulated by changing the statistics window size w and the downsampling factor q.

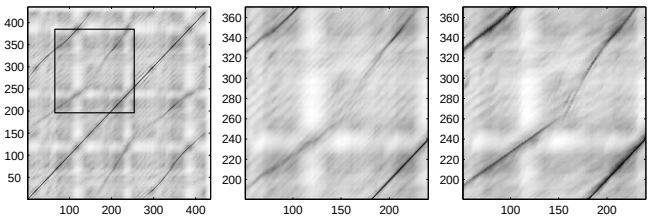

Fig. 3. Left: Similarity matrix $\mathcal{S}_{15}(41,10)$ of three concatenations (normal tempo/150 percent of normal tempo/accelerating from 70 to 150 percent and then decelerating back to 70 percent of normal tempo) of Bach's Toccata BWV 565. Center: Enlargement of the region of $\mathcal{S}_{15}(41,10)$ indicated by the dashed square. Right: Corresponding region of $\mathcal{S}_{15}^{\min}(41,10)$.

3. APPLICATIONS

Our matrix enhancement strategy enables certain audio analysis tasks such as audio summarization or audio synchronization on music audio data, where musically similar segments may exhibit large variations in dynamics, timbre, articulation, and local tempo. In the following, we sketch how problems which typically arise in previous approaches can be overcome by using the proposed techniques.

3.1. Audio summarization

One major goal of *audio summarization* is, given a particular audio signal, to automatically extract the significant repetitions, from which musical thumbnails may be derived, see, e.g., [5, 6, 4] and the references therein. Most of these approaches are based on the constant tempo assumption (dealing with pop music) and develop refined methods for extracting suitable (straight) diagonal paths from the self-similarity matrix. Here, the main difficulties arise from the fact that due to spectral and temporal variations, actual repetitions may correspond to a number of disconnected path fragments, see, e.g., the fragmentary path representing the repetition A_3 of A_1 shown in the dashed box of the left part of Fig. 2. Instead of relying on complicated and delicate path extraction algorithms, we suggest a different approach by taking care of the variations also at the feature and similarity measure levels. By improving the path structure of the matrix, one can then extract the paths in an automatic and robust fashion. As an illustration, compare the left and right part of Fig. 2. Similarly, the enhanced matrix $\mathcal{S}_{10,2}^{\min}(21,5)$ of the Brahms example shown in the right part of Fig. 1 clearly reveals the path structure, which makes a robust extraction even of the curved path possible. Here, the main observation is that such an amplified and smoothed path can typically be identified by looking at local vertical and horizontal minima.

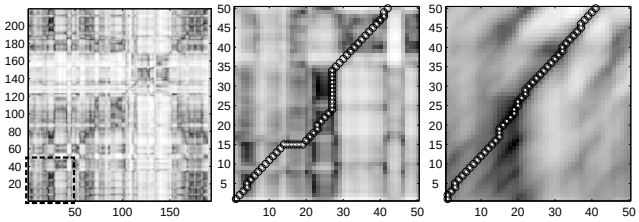

Fig. 4. Left: $\mathcal{S}(41, 10)$ of a Zukerman (vertical) and Mae (horizontal) interpretation of Vivaldi's Spring RV 269, No. 1. Center: Enlargement of the region of $\mathcal{S}(41, 10)$ indicated by the dashed square with (incorrect) alignment path indicated by white dots. Right: Corresponding region of $\mathcal{S}_{10}^{\min}(41, 10)$ with (correct) alignment path.

3.2. Multiresolution Audio Synchronization

The goal of *audio synchronization* is to time-align two given audio versions of the same underlying piece of music. Most approaches, see, e.g., [7, 8], rely on some variant of dynamic time warping (DTW): first, a suitable similarity (or cost) matrix is computed with respect to the two audio versions. Then, the similarity-maximizing (cost-minimizing) alignment path is determined from this matrix via dynamic programming. One major problem (as typical for, e.g., classical music) arises from the fact that two interpretations of the same piece of music may differ considerably in some sections due to different realizations of note groups such as trills or arpeggios. For example, the Mae interpretation of the Vivaldi example of Fig. 4 includes many additional ornamentations, which can not be found in the Zukerman interpretation. This may lead to distorted and incorrect alignment paths as illustrated by the center part of Fig. 4. In this case, the inclusion of contextual information into the similarity measure helps to absorb such local inconsistencies, leading to robuster and more reliable alignment estimations, see the right part of Fig. 4.

Another problem concerns the high time and space complexity of $O(NM)$ to compute and store similarity matrices. To this end, Salvador et al. [9] propose a multiresolution DTW approach that recursively projects an alignment solution from a coarse resolution level to the next higher level and then refines the projected solution. One hazard with this approach is that an incorrect alignment on a low resolution level propagates to higher levels resulting in erroneous alignment results. This hazard is fostered by the fact that coarsening the features can lead to heavily deteriorated similarity matrices, see the first row of Fig. 5. In this context, by our enhancement strategy one can achieve a good compromise in reducing the feature sampling rate without sacrificing too much of the structural properties of the similarity matrix; the second row of Fig. 5 gives an example.

4. CONCLUSIONS

We have described an approach for enhancing the structural properties of similarity matrices by taking care of temporal variations in the audio data on the three levels of feature design, similarity measure, and structure extraction. As our experiments indicate, this strategy serves as an important step towards solving various audio analysis tasks for genres such as classical music, which exhibit the above mentioned variations to a significant degree even for musically similar excerpts. (See also `www-mmdb.iai.uni-bonn.de/projects/simmat` for a more comprehensive account on our experiments.) In the future, a more detailed investigation of the application of our enhancement techniques to music analysis tasks as well as a quantitative evaluation of the experimental results will be necessary. Here, an important point will be the automatic adjustment of the enhancement parameters. We furthermore plan to transfer our general enhancement techniques—in this paper only applied to the class of chroma features—to other features classes.

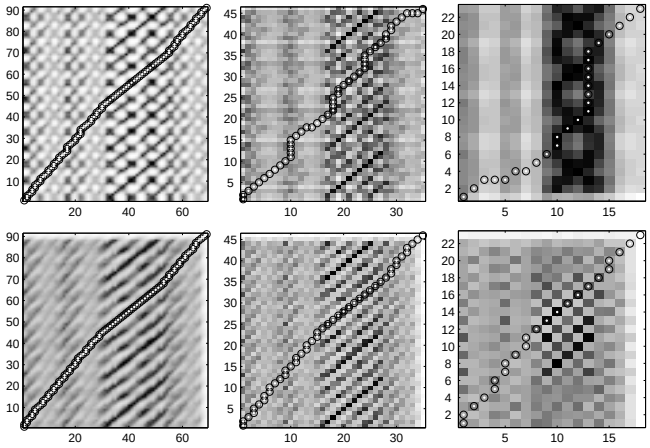

Fig. 5. First row: $\mathcal{S}(41, 10)$, $\mathcal{S}(81, 20)$ and $\mathcal{S}(161, 40)$ of an audio file and a temporally distorted version. The alignment paths indicated by the white dots are incorrect for the two lower resolution levels. Second row: $\mathcal{S}_{4,1}^{\min}(41, 10)$, $\mathcal{S}_{4,2}^{\min}(41, 10)$ and $\mathcal{S}_{8,4}^{\min}(41, 10)$ of the same audio files leading to correct alignment paths even on the lower resolution levels.

5. REFERENCES

[1] Jonathan Foote, "Visualizing music and audio using self-similarity," in *ACM Multimedia (1)*, 1999, pp. 77–80.

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

[3] Meinard Müller, Frank Kurth, and Michael Clausen, "Audio matching via chroma-based statistical features," in *Proc. ISMIR, London, GB*, 2005.

[4] Geoffroy Peeters, Amaury La Burthe, and Xavier Rodet, "Toward automatic music audio summary generation from signal analysis," in *Proc. ISMIR, Paris, France*, 2002.

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

[6] Lie Lu, Muyuan Wang, and Hong-Jiang Zhang, "Repeating pattern discovery and structure analysis from acoustic music data," in *Workshop on Multimedia Information Retrieval, ACM Multimedia*, 2004.

[7] Ning Hu, Roger Dannenberg, and George Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *Proc. IEEE WASPAA, New Paltz, NY*, October 2003.

[8] Robert J. Turetsky and Daniel P.W. Ellis, "Force-Aligning MIDI Syntheses for Polyphonic Music Transcription Generation," in *Proc. ISMIR, Baltimore, USA*, 2003.

[9] Stan Salvador and Philip Chan, "FastDTW: Towards accurate dynamic time warping in linear time and space," in *Proc. KDD Workshop on Mining Temporal and Sequential Data*, 2004.