TOWARD MUSICALLY-MOTIVATED AUDIO FINGERPRINTS

Peter Grosche and Meinard Müller

Saarland University and MPI Informatik

{pgrosche,meinard}@mpi-inf.mpg.de

ABSTRACT

In this paper, we investigate to which extent well-known audio fingerprinting techniques, which aim at identifying a specific audio recording, can be modified to also deal with more musical variations. To this end, we replace the standard peak fingerprints based on a spectrogram by peak fingerprints based on other more "musical" feature representations. Our systematic experiments show that such modified peak fingerprints allow for a robust identification of different versions and performances of the same piece of music if the query length is at least 15 seconds. This indicates that highly efficient audio fingerprinting techniques can also be applied to accelerate tasks such as audio matching or cover song identification.

Index Terms— Fingerprinting, spectral peaks, music representations, audio matching, cover song identification

1. INTRODUCTION

The task of *audio fingerprinting* or *audio identification* constitutes an important research topic of commercial relevance [1, 2, 3]. Here, given a short fragment of an audio signal, the goal is to retrieve the original audio recording of this fragment from a huge music database without relying on manually created metadata. Recent identification algorithms can handle background noise and signal degradations and are highly efficient. However, such systems are not capable of retrieving different performances of the same piece of music. The reason for this is that existing audio fingerprinting algorithms are not designed for dealing with musical variations such as strong nonlinear temporal distortions, variations that concern the articulation, instrumentation, or ornamentation. Opposed to traditional audio fingerprinting, the goal of *audio matching* [4] and *cover song identification* [5, 6] is to retrieve all audio clips and versions that are musically (semantically) related to the query fragment. First index-based approaches to these tasks have been suggested in [4, 7].

In this paper, we investigate to which extent well-established audio fingerprints as introduced in [3] can be modified to allow for retrieving musically related recordings. To this end, we replace the traditional fingerprints based on spectral peaks by fingerprints based on peaks of more musically oriented feature representations including log-frequency and chroma representations. Our motivation for adopting this approach is that such peak structures, according to [3], are temporally localized, reproducible, and robust against many, even significant distortions of the signal. Furthermore, the spectral peaks allow for applying efficient hash-based indexing techniques. The main contribution of this paper is to systematically analyze the resulting peak structures in view of robustness and discriminative power. Finding a good trade-off between these two principles is a non-trivial task. On the one hand, using fine-grained feature representations (such as a spectrogram) results in fingerprints that are too specific, thus not facilitating cross-version retrieval. On the other hand, using coarse feature representations (such as a chromagram) results in peak fingerprints that are too unspecific and noisy, thus not having the required discriminative power. Our experimental results in the context of a music retrieval scenario indicate that, using suitably modified peak fingerprints, one can transfer traditional audio fingerprinting techniques to other tasks such as audio matching and cover song identification.

The remainder of the paper is organized as follows. In Section 2 we introduce various peak fingerprints based on different feature representations. In Section 3, as our main contribution, we systematically investigate the trade-off between robustness and discriminative power of the various audio fingerprints. Finally, conclusions and an outlook on future work towards a modified audio fingerprinting system can be found in Section 4. Further related work is discussed in the respective sections.

2. FINGERPRINTS

In this section, we introduce the various peak fingerprints used in our investigations. Our approach is based on the concept of spectral peaks originally introduced by Wang [3] and now widely used in commercial products.[1] In this approach, characteristic time-frequency peaks extracted from a spectrogram are used as fingerprints, thus reducing a complex spectrogram to a sparse peak representation of high robustness against signal distortions. Such peak representations allows for applying efficient hash-based indexing techniques. We transfer this approach to a more flexible retrieval scenario by considering various feature representations that are obtained by partitioning the frequency axis of the original spectrogram, while the temporal axis of all representations is fixed to yield a feature rate of 20 Hz (20 feature per second), see Figure 1 for an illustration.

The first feature representation is a magnitude spectrogram as employed in the original approach. Following [3], the audio signal is sampled at $f_s = 8000$ Hz and Discrete Fourier Transforms are calculated over windows of 1024 samples. In the following, the resulting feature representation is referred to as `SPEC`, see Figure 1b. The second feature representation is a log-frequency spectrogram [2]. Using a suitable binning strategy, we group the Fourier coefficients of the original spectrogram into 33 non-overlapping frequency bands covering the frequency range from 300 Hz to 2000 Hz. Exhibiting a logarithmic spacing, the bands roughly represent the Bark scale. In the following, this feature representation is referred to as `LOGF`, see Figure 1c. As third feature representation, we consider a constant-Q transform where the frequency bins are logarithmically spaced

The authors are supported by the Cluster of Excellence on Multimodal Computing and Interaction at Saarland University.

[1] www.shazam.com

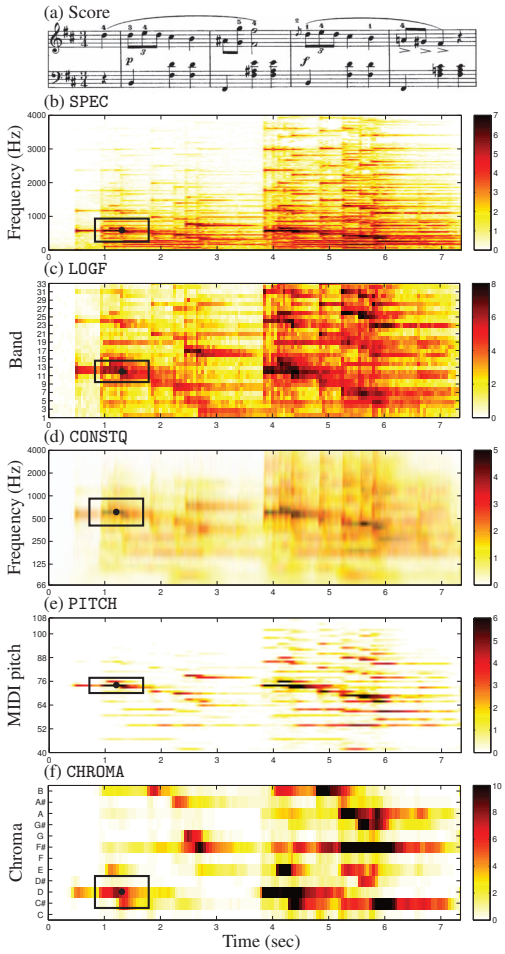

Fig. 1: Score and various feature representations for the first 7.35 seconds of a Hatto (2006) performance of the first 5 bars of Chopin's Mazurka Op. 30 No. 2. One peak and the corresponding neighborhood is shown for each of the feature representations.

and the ratios of the center frequencies to bandwidths of all bins are equal (Q factor). In our investigation, we employ the efficient implementation provided by the *Constant-Q Transform Toolbox for Music Processing*[2], see [8]. Here, we set the number of frequency bins per octave to 12 (each bin corresponds to one semitone of the equal-tempered scale) and consider the frequency range from 80 Hz to 4000 Hz. In the following, this feature is referred to as `CONSTQ`, see Figure 1d. To obtain the fourth feature representation, we decompose the audio signal into 88 frequency bands with center frequencies corresponding to the pitches of the equal-tempered scale and compute the short-time energy in windows of length 100 ms. For deriving this decomposition, we use a multirate filter bank as described in [9] and denote the resulting feature as `PITCH`, see Figure 1e. The fifth feature representation is a chroma representation which is obtained from `PITCH` by adding up the corresponding values that belong to the same chroma. In the following, this feature is referred to as `CHROMA`, see Figure 1e. Implementations for `PITCH` and `CHROMA` are provided by the *Chroma Toolbox*[3], see [10].

In the second step, we employ a similar strategy as proposed in [3] to extract characteristic peaks from the various feature representations. Given a feature representation $\mathcal{F} \in \mathbb{R}^{T \times K}$ where $\mathcal{F}(t,k)$ denotes the feature value at frame $t \in [1:T] := \{1,2,\ldots,T\}$ for some $T \in \mathbb{N}$ and frequency bin $k \in [1:K]$ for some $K \in \mathbb{N}$, we select a point (t_0,k_0) as a peak if $\mathcal{F}(t_0,k_0) \geq \mathcal{F}(t,k)$ for all $(t,k) \in \left[t-\Delta^{\text{time}} : t+\Delta^{\text{time}}\right] \times \left[k-\Delta^{\text{freq}} : k+\Delta^{\text{freq}}\right]$ in a local neighborhood defined by Δ^{time} and Δ^{freq}. The size of this neighborhood allows for adjusting the peak density. In our implementation, we use an additional absolute threshold on the values $\mathcal{F}(t_0,k_0)$ to prevent the selection of more or less random peaks in regions of very low dynamics. The selected peaks are represented in the form of a binary matrix $\mathcal{P} \in \{0,1\}^{T \times K}$ by setting $\mathcal{P}(t_0,k_0) = 1$ for (t_0,k_0) being a peak and zero elsewhere. This peak selection strategy reduces a complex time-frequency representation $\mathcal{F}$ to a sparse set $\mathcal{P}$ of time-frequency points. Note that the values of $\mathcal{F}(t,k)$ are no longer considered in the fingerprints.

In our experiments, we fix $\Delta^{\text{time}} = 20$ corresponding to one second for all five feature representations. The range of the frequency neighborhood Δ^{freq}, however, was experimentally determined for each feature representation. For `SPEC` we set $\Delta^{\text{freq}} = 25$ (corresponding to 200 Hz), for `LOGF` we set $\Delta^{\text{freq}} = 2$, for `CONSTQ` we set $\Delta^{\text{freq}} = 3$, for `PITCH` we set $\Delta^{\text{freq}} = 3$, and for `CHROMA` we set $\Delta^{\text{freq}} = 1$, see Figure 1 for an illustration of the neighborhood for each of the feature representations.

3. EXPERIMENTS

We now investigate the musical expressiveness of the various peak fingerprints. In Section 3.1, we start with introducing the datasets used in our experiments. Then, in Section 3.2, we sketch how the peaks of different performances are warped to a common time line. In Section 3.3, we discuss an experiment that indicates the degree of peak consistency across different performances depending on the underlying feature representation. Finally, in Section 3.4, we describe a document-based retrieval experiment.

3.1. Dataset

For our subsequent experiments, we use three different groups of audio recordings corresponding to pieces of classical music by three different composers, see Table 1. The first group **Chop** consists of 298 piano recordings of five Mazurkas by Frédéric Chopin collected in the Mazurka Project.[4] The second group **Beet** consists of ten recorded performances of Beethoven's *Symphony No. 5*. This collection contains orchestral as well as piano performances. The third group **Viva** contains seven orchestral performances of the *Summer* from Vivaldi's *Four Seasons*. Table 1 lists the number of performances as well as the total duration of each movement/piece. The union of all groups is referred to as **All** and contains 359 recordings with an overall length of 19 hours. In view of extracting peak fingerprints, these three groups are of increasing complexity. While for the piano recordings of **Chop**, one expects relatively clear peak structures, peak picking becomes much more problematic for general orchestral music (group **Beet**) and music dominated by strings (group **Viva**).

3.2. Synchronization

In our retrieval scenario, there typically are tempo differences between the different interpretations of a piece. In our initial experi-

[2] http://www.elec.qmul.ac.uk/people/anssik/cqt/

[3] http://www.mpi-inf.mpg.de/resources/MIR/chromatoolbox/

[4] http://mazurka.org.uk/

Groups	Composer	Piece	Description	#(Perf.)	Dur. (min)
Chop	Chopin	Op. 17, No. 4	Mazurka	62	269
	Chopin	Op. 24, No. 2	Mazurka	64	147
	Chopin	Op. 30, No. 2	Mazurka	34	48
	Chopin	Op. 63, No. 3	Mazurka	88	189
	Chopin	Op. 68, No. 3	Mazurka	50	84
Beet	Beethoven	Op. 67, 1. Mov.	Fifth	10	75
	Beethoven	Op. 67, 2. Mov.	Fifth	10	98
	Beethoven	Op. 67, 3. Mov.	Fifth	10	52
	Beethoven	Op. 67, 4. Mov.	Fifth	10	105
Viva	Vivaldi	RV 315, 1. Mov.	Summer	7	38
	Vivaldi	RV 315, 2. Mov.	Summer	7	17
	Vivaldi	RV 315, 3. Mov.	Summer	7	20
All				359	1145

Table 1: The groups of audio recordings used in our experiments. The last two columns denote the number of different performances and the overall duration in minutes.

ments, we do not want to deal with this issue and compensate for tempo differences in the performances by temporally warping the peak representations onto a common time line. To this end, we, in a preprocessing step, use a music synchronization technique [11] to temporally align the different performances of a given piece of music. More precisely, suppose we are given N different performances of the same piece yielding the peak representations $\mathcal{P}_n$, $n \in [1:N]$. Then, we take the first performance as reference and compute alignments between the reference and the remaining $N-1$ performances. The alignments are then used to temporally warp the peak representations $\mathcal{P}_n$ for $n \in [2:N]$ onto the time axis of the peak representation $\mathcal{P}_1$. The resulting warped peak fingerprints are denoted by $\tilde{\mathcal{P}}_n$ and we set $\tilde{\mathcal{P}}_1 = \mathcal{P}_1$.

3.3. Experiment: Peak Consistency

In a first experiment, we investigate to which extent the various peak fingerprints coincide across different performances of a piece. Here, the degree of peak consistency serves as an indicator for the robustness of the respective feature representation towards musical variations. We express the consistency of the fingerprints of two performances in terms of pairwise precision P, recall R, and F-measure F. More precisely, given two performances $n, m \in [1:N]$ of a piece, we obtain the aligned peak fingerprints $\tilde{\mathcal{P}}_n$ and $\tilde{\mathcal{P}}_m$ as explained in Section 3.2. Then, a peak (t_0, k_0) of $\tilde{\mathcal{P}}_m$ is called *consistent* relative to $\tilde{\mathcal{P}}_n$ if there is a peak (t, k_0) of $\tilde{\mathcal{P}}_n$ with $t \in [t_0 - \tau : t_0 + \tau]$. Here, the parameter $\tau \geq 0$ specifies a small temporal tolerance window. Otherwise, the peak is called *non-consistent*. The number of consistent fingerprints is denoted by $\#(\tilde{\mathcal{P}}_n \cap \tilde{\mathcal{P}}_m)$, the overall number of peaks in $\tilde{\mathcal{P}}_n$ and $\tilde{\mathcal{P}}_m$ is denoted $\#(\tilde{\mathcal{P}}_n)$ and $\#(\tilde{\mathcal{P}}_m)$, respectively. Then, pairwise precision $\mathrm{P}_{n,m}$, recall $\mathrm{R}_{n,m}$, and F-measure $\mathrm{F}_{n,m}$ are defined as

$$\mathrm{P}_{n,m} = \frac{\#(\tilde{\mathcal{P}}_n \cap \tilde{\mathcal{P}}_m)}{\#(\tilde{\mathcal{P}}_m)},\ \mathrm{R}_{n,m} = \frac{\#(\tilde{\mathcal{P}}_n \cap \tilde{\mathcal{P}}_m)}{\#(\tilde{\mathcal{P}}_n)}, \quad (1)$$

$$\mathrm{F}_{n,m} = \frac{2 \cdot \mathrm{P}_{n,m} \cdot \mathrm{R}_{n,m}}{\mathrm{P}_{n,m} + \mathrm{R}_{n,m}}. \quad (2)$$

Note, that $\mathrm{P}_{n,m} = \mathrm{R}_{m,n}$, $\mathrm{R}_{n,m} = \mathrm{P}_{m,n}$, and therefore $\mathrm{F}_{n,m} = \mathrm{F}_{m,n}$. F-measure values are computed for all N performances of a group yielding an $(N \times N)$-matrix of pairwise F values. Mean values for the groups are obtained by averaging over the respective F-measures. Here, as $\mathrm{F}_{n,n} = 1$ and $\mathrm{F}_{n,m} = \mathrm{F}_{m,n}$, we only consider the values of the upper triangular part of the matrix excluding the main diagonal.

Table 2 shows the mean of pairwise F-measure values for the different groups of our dataset. In this experiment, we use the tolerance parameter $\tau = 1$ (corresponding to ± 50 ms), which turned out to be a suitable threshold for compensating inaccuracies introduced by the synchronization procedure, see [11]. First note that the originally used spectrogram peaks do not work well across different performances. For example, in the case of **Chop**, one obtains $\mathrm{F} = 0.081$ for `SPEC` indicating that only few of the peak fingerprints consistently occur across different performances. The peaks extracted from the other four feature representations show a higher degree of consistency across performances e.g., in the case of **Chop**, $\mathrm{F} = 0.205$ for `LOGF`, $\mathrm{F} = 0.157$ for `CONSTQ`, $\mathrm{F} = 0.185$ for `PITCH`, and $\mathrm{F} = 0.375$ for `CHROMA`. This improvement is achieved by the coarser and musically more meaningful partition of the frequency axis. Furthermore, our results show a dependency on the characteristics of the audio material. In particular, the peaks are more consistent for **Chop** (e.g. $\mathrm{F} = 0.375$ for `CHROMA`) than for **Beet** ($\mathrm{F} = 0.288$) and **Viva** ($\mathrm{F} = 0.262$). The reason for this effect is twofold. Firstly, the piano pieces as contained in **Chop** exhibit pronounced note onsets leading to consistent peaks. For complex orchestral and string music as in **Beet** and **Viva**, however, the peaks are less dominant leading to a lower consistency. Secondly, the consistency results are also influenced by the accuracy of the peak synchronization as introduced in Section 3.2. Typically, the synchronization technique [11] yields very precise alignments for piano music as contained in **Chop**. For orchestral and string pieces as in **Beet** and **Viva**, however, there are more synchronization inaccuracies leading to lower F-measure values.

Groups	`SPEC`	`LOGF`	`CONSTQ`	`PITCH`	`CHROMA`
Chop	0.081	0.205	0.157	0.185	0.375
Beet	0.051	0.139	0.126	0.137	0.288
Viva	0.059	0.143	0.124	0.132	0.262
All	0.080	0.203	0.156	0.184	0.373

Table 2: Mean of pairwise F-measure values expressing peak consistencies for the different groups.

3.4. Experiment: Document-based Retrieval

In the second experiment, we investigate the identification rate of the modified peak fingerprints in a document-based retrieval scenario. Given a short query extracted from one performance, the goal is to correctly retrieve all performances of the same piece from a larger dataset. Exploiting the warped peak fingerprints $\tilde{\mathcal{P}}$ (see Section 3.2), we define a query Q and a database collection $\mathcal{D}$. The database consists of $N_{\mathcal{D}}$ performances (documents) of different groups. For a query Q and a document $D \in \mathcal{D}$, we compute the peak consistency F-measure as in (2) between Q and all subsegments of D having the same length as Q. High F-values indicate high degrees of peak consistency between Q and subsegments of D. Considering document-level retrieval, the similarity between Q and D is defined as the maximum F-measure over all subsegments of D.

In the evaluation, the N_Q interpretations of the piece underlying the query are considered *relevant*, all other *irrelevant*. Following [6], we express the retrieval accuracy using the *mean of average precision measure* (MAP) denoted as $\langle\overline{\psi}\rangle$.[5] To this end, we sort the documents $D \in \mathcal{D}$ in descending order with respect to the similarity between D and Q and obtain the precision ψ_Q at rank $r \in [1:N_{\mathcal{D}}]$ as

$$\psi_Q(r) = \tfrac{1}{r} \textstyle\sum_{i=1}^{r} \Gamma_Q(i)\,, \quad (3)$$

[5] The same measure is used in the MIREX Cover Song Identification, see www.music-ir.org/mirex/wiki/2010:Audio_Cover_Song_Identification

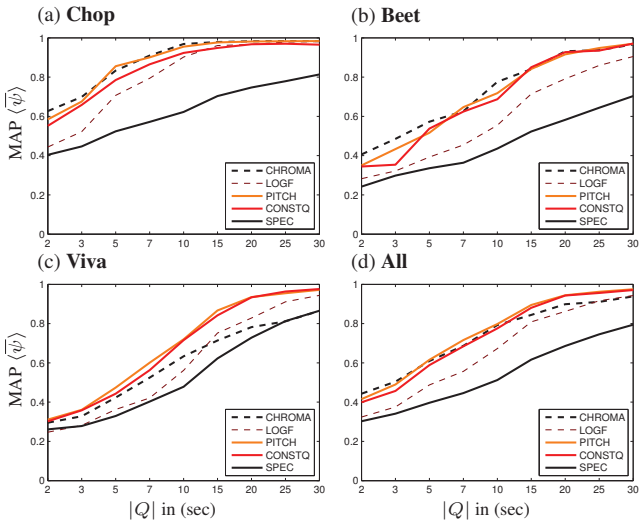

Fig. 2: Results for the retrieval experiment showing the dependency of MAP values $\langle\overline{\psi}\rangle$ on the query length $|Q|$ using queries from **(a) Chop** ($\langle\overline{\psi}\rangle_{\text{null}} = 0.190$), **(b) Beet** ($\langle\overline{\psi}\rangle_{\text{null}} = 0.040$), **(c) Viva** ($\langle\overline{\psi}\rangle_{\text{null}} = 0.032$), and **(d)** average over all queries.

where $\Gamma_Q(r) \in \{0, 1\}$ indicates whether the document at rank r is relevant for Q. Then, the average precision $\overline{\psi}_Q$ is defined as

$$\overline{\psi}_Q = \tfrac{1}{N_Q} \sum_{r=1}^{N_{\mathcal{D}}} \psi_Q(r)\Gamma_Q(r) \; . \tag{4}$$

Finally, given C different queries we compute $\overline{\psi}_Q$ for each Q and average over all C values to obtain the mean of average precision measure $\langle\overline{\psi}\rangle$. In our experiments, for a fixed query length $|Q|$, we randomly select $C = 100$ queries from each group. Additionally, we estimate the accuracy level $\langle\overline{\psi}\rangle_{\text{null}}$ expected under the null hypothesis of a randomly created sorted list, see [6] for details.

Figure 2 shows the resulting MAP $\langle\overline{\psi}\rangle$ values as a function of the query length $|Q|$ for the five features. The queries are taken from the different groups, the database $\mathcal{D}$ contains all performances of **All**. As the results show, the retrieval accuracy using modified peak fingerprints is much higher than using the original spectrogram peaks. In particular, using the musically motivated features `PITCH`, `CONSTQ`, and `CHROMA` results in the highest MAP $\langle\overline{\psi}\rangle$ for **All**, see Figure 2d. For **Viva** (Figure 2c), the retrieval accuracy for `PITCH` and `CONSTQ` is significantly higher than for `CHROMA`. Here, a manual inspection revealed that the peaks of `CHROMA`, although more consistent across performances than peaks of `PITCH` and `CONSTQ` (see Section 3.3), exhibit less discriminative power. In the case of the less pronounced peaks of **Viva**, this frequently results in undesired high consistency for unrelated fingerprints when using `CHROMA`. Contrary, the higher discriminative power of peaks from `PITCH` and `CONSTQ` (although of lower overall consistency) results in higher retrieval accuracies.

Furthermore, the results show a great dependency of the retrieval accuracy on the query length $|Q|$. Surprisingly, in the case of **Chop** (Figure 2a), even $|Q| = 2$ sec leads to already relatively high MAP values. Increasing the query length, the MAP values increase for all feature representations and groups of audio recordings. For all groups, using a query length of 20 sec in combination with peak fingerprints extracted from `PITCH` or `CONSTQ` results in MAP values $\langle\overline{\psi}\rangle > 0.9$. In particular for the more complex data contained in **Beet** (Figure 2b) and **Viva** (Figure 2c) using longer queries further improves the identification rate across performances.

4. CONCLUSIONS

In this paper, we studied the robustness and discriminative power of modified audio fingerprints by considering peak consistencies across different versions of the same piece of music. As our experiments reveal, modified peak fingerprints based on musically motivated time-pitch or time-chroma representations allow for an identification across different performances of the same piece of music. We also showed that, in contrast to 3-5 sec long queries considered for traditional audio fingerprinting, 15-25 sec long queries are necessary for obtaining a robust and accurate cross-performance identification procedure.

Our results indicate that, using more musical feature representations, it is possible to employ similar techniques as used by Shazam for other music retrieval tasks such as audio matching or cover song retrieval. In our investigation, temporal differences between performances were compensated in a preprocessing step using an offline music synchronization technique. Future work concerns the issue on how the temporal differences between performances can be considered in the actual retrieval step. In particular, for designing an efficient and scalable system, indexing techniques based on robust and discriminative hashes that can cope with temporal differences between performances need to be investigated.

5. REFERENCES

[1] Eric Allamanche, Jürgen Herre, Oliver Hellmuth, Bernhard Fröba, and Markus Cremer, "AudioID: Towards content-based identification of audio material," in *Proc. 110th AES Convention*, Amsterdam, NL, 2001.

[2] Pedro Cano, Eloi Batlle, Ton Kalker, and Jaap Haitsma, "A review of algorithms for audio fingerprinting," in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*, St. Thomas, Virgin Islands, USA, 2002, pp. 169–173.

[3] Avery Wang, "An industrial strength audio search algorithm," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Baltimore, USA, 2003, pp. 7–13.

[4] Frank Kurth and Meinard Müller, "Efficient index-based audio matching," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 382–395, Feb. 2008.

[5] Daniel P. W. Ellis and Graham. E. Poliner, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, Hawaii, USA, Apr. 2007, vol. 4.

[6] Joan Serrà, Xavier Serra, and Ralph G. Andrzejak, "Cross recurrence quantification for cover song identification," *New Journal of Physics*, vol. 11, no. 9, pp. 093017, 2009.

[7] Michael Casey, Christophe Rhodes, and Malcolm Slaney, "Analysis of minimum distances in high-dimensional musical spaces," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 16, no. 5, 2008.

[8] Christian Schörkhuber and Anssi Klapuri, "Constant-Q transform toolbox for music processing," in *Sound and Music Computing Conference (SMC)*, Barcelona, 2010.

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

[10] Meinard Müller and Sebastian Ewert, "Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), to appear*, Miami, USA, 2011.

[11] Sebastian Ewert, Meinard Müller, and Peter Grosche, "High resolution audio synchronization using chroma onset features," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 1869–1872.