One Day in Half an Hour: Music Thumbnailing Incorporating Harmony- and Rhythm Structure

Björn Schuller, Florian Dibiasi, Florian Eyben, and Gerhard Rigoll

Institute for Human-Machine Communication, Technische Universität München,
Arcisstrasse 21, 80333 München, Germany
schuller@tum.de,
http://www.mmk.ei.tum.de

Abstract. A variety of approaches exist to the automatic retrieval of the key part within a musical piece - its thumbnail. Most of these however do not use adequate modeling with respect to either harmony or rhythm. In this work we therefore introduce thumbnailing that aims at adequate musical feature modeling. The rhythmic structure is extracted to obtain a segmentation based on beats and bars by an IIR comb-filter bank. Further, we extract chroma energy distribution normalized statistics features of the segmented song improving performance with dB(A) and pitch correction. Harmonic similarities are determined by construction and analysis of a similarity matrix based on the normalized scalar product of the feature vectors. Last, thumbnails are found lending techniques from image processing. Extensive test runs on roughly 24 h of music reveal the high effectiveness of our approach.

1 Introduction

A wide variety of applications uses audio thumbnails in order to provide an insight into songs such as pre-hear functions in online music stores, teaser previews on the radio, samples for deejays to create mega-mixes or efficiently browsing through large music collections, e.g. on a mobile MP3 player. In addition query by example systems highly benefit from pre-extracted thumbnails to build up the required database for similarity matching with sung or hummed queries. Nowadays these thumbnails usually have to be generated manually due to the lack of appropriate and robust methods which are capable to accomplish reliable automatic generation. Especially for acoustic formats which deal with real audio (we use the popular MPEG-1 Audio Layer 3 standard) there are no known prosperous methods or systems so far.

Since voiced, repeating sequences such as chorus sections are believed to be the most mnemonic parts of songs [Burges et al., 2004], the approach described in this work aims at extracting the chorus by successfully combining different ideas of previous works. There are several works dealing with extracting audio thumbnails or determining the musical structure of songs. They can be divided

From: Proc. AMR 2008, Berlin

into two general types regarding the feature nature they are based upon. Alternatively they can also be classified according to their kind of approach, such as building up and analyzing similarity matrices or applying a segmentation step with a subsequent clustering or classification.

Burges et al. [Burges et al., 2004] create their own features by applying a modulated complex lapped transform followed by a logarithmizing step and by reducing the information using oriented principal component analysis. A clustering algorithm is applied to merge similar sequences which are then classified by using a scaled Renyi entropie and spectral flatness. Logan et al. [Logan and Chu, 2000] extract Mel frequency cepstral coefficients (MFCC) and cluster the song using a modified Kullback Leibler distance. The second approach presented in [Logan and Chu, 2000] models each song by an ergodic hidden Markov model (HMM) in order to extract the musical structure. Aucouturier et al. use ergodic HMM with 3 states by approximating the spectral envelope with MFCC, linear predictive coefficients and discrete cepstrum coefficients [Aucouturier et al., 2005] and with Gaussian mixture models initialized by a clustering step [Aucouturier and Sandler, 2001]. Foote et al. introduce the concept of music visualization by constructing a similarity matrix based on MFCC using the scalar product [Foote, 1999] and a normalized scalar product [Cooper and Foote, 2002]. Peeters et al. [Peeters et al., 2002] use dynamic features which maximize the trans-information and build up a similarity matrix. A segmentation step splits the song into small segments which are then processed with a clustering algorithm in order to generate an initialized set for ergodic HMM. Jehan [Jehan, 2005] first divides the music signal into several segments using an event detection function and then applies dynamic time warping to extract the musical structure.

Abdallah et al. [Abdallah et al., 2005] use an unsupervised Bayesian clustering model to extract musical structure by estimating its parameters using a modified expectation maximization algorithm. Bartsch et al. [Bartsch and Wakefield, 2001] perform a beat synchronous segmentation using a beat tracker followed by a similarity matrix based on chroma features. By applying uniform moving average filtering a time-lag matrix is obtained whose maximum element is located according to constraints regarding the minimum lag and the maximum occurrence of a section. Goto [Goto, 2006] extends this work by allowing modulated repetitions and by using an adapted measure to select chorus sections. Müller et al. [Müller and Kurth, 2006] present an approach for enhancing similarity matrices by introducing a new set of features based on harmonic informations.

From these works we learn that beat-synchronous feature extraction is advantageous provided a reliable detection. Features that respect the musical background of the signal, such as chroma, are superior to e.g. MFCC, and it seems favorable to incorporate temporal information as in [Müller and Kurth, 2006]. Finally, retrieving the pre-dominant parts by similarity matrices seems most promising, and dynamic modeling, such as Dynamic Time Warp (DTW), is usually rather contra-productive provided beat-synchrony. In the next section we

describe a system built upon these considerations that lends simple but fast techniques from image processing for the similarity matching and enhances features by perceptive modeling. We also define measures for evaluation and report results on a day of real MP3 audio from diverse genres.

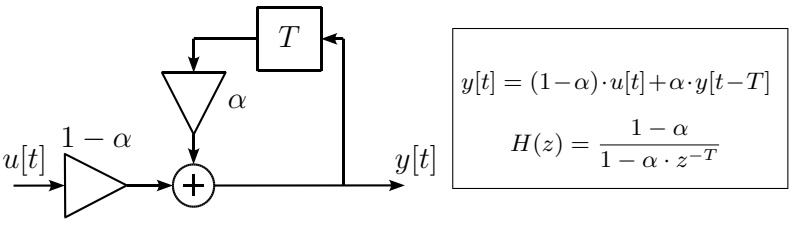

Fig. 1. Block diagram (left), difference equation (top) and transfer function (bottom) of an IIR comb filter

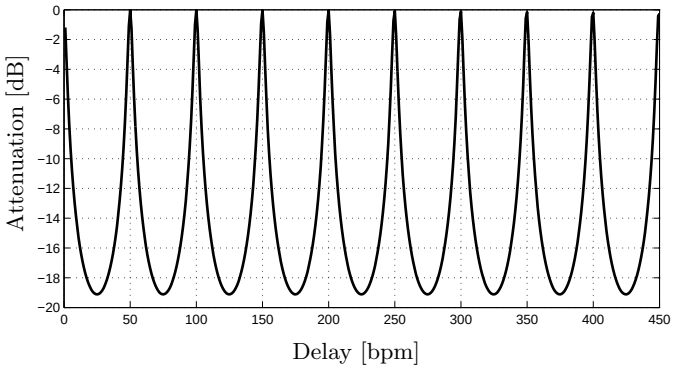

Fig. 2. Magnitude response for an IIR comb filter with gain $\alpha = 0.8$ and base tempo $50bpm$

2 Feature Extraction

2.1 Rhythm Information

We use our highly robust beat tracker introduced in [Schuller et al., 2007]to extract rhythmic structure. After a preprocessing step which involves downsampling to 11 025 kHz and transforming into the frequency domain, the signal is filtered with the A-weighting function according to the human perception of sound. In order to reduce the number of bands without loosing rhythmic information the audio signal is split into frequency bands using a bank of 24 overlapping triangular filters which are equidistant on the Mel-Frequency scale.

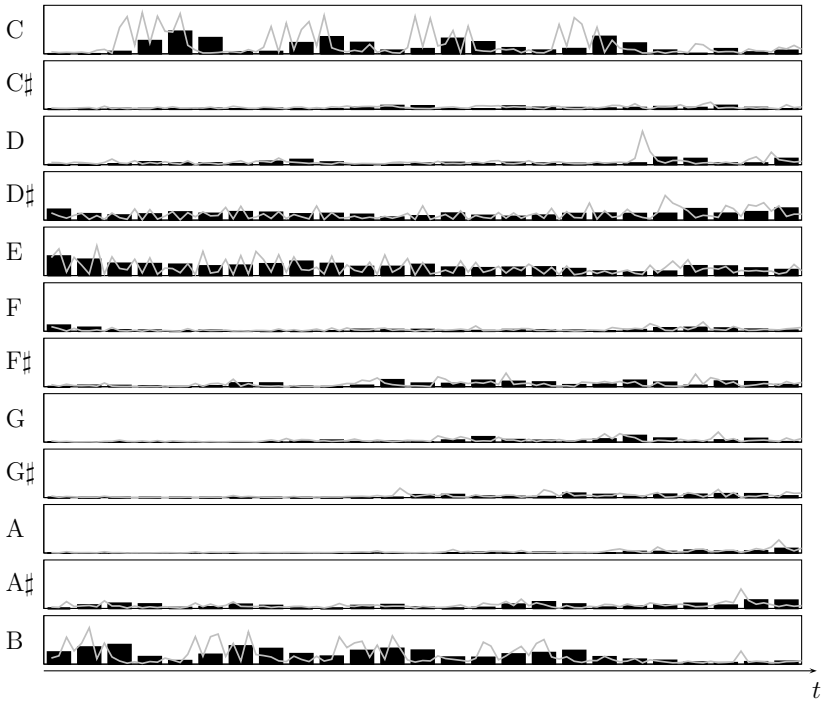

Fig. 3. Harmonic representation of the first 20 seconds of Abba - Mamma Mia. The light curves illustrate the local chroma energy distribution, the dark bars the CENS features.

Next, the envelope of each band is extracted using a half wave raised cosine filter and processed by incorporating the moving average over the previous 10 and the following 20 samples due to the fact that humans perceive note onsets louder if they occur after a longer time of lower sound level. Hence, we determine the lowest metrical level referred to as tatum grid using a bank of 57 phase comb filters with gain $\alpha = 0.8$ and delays ranging from $\tau = 18$ to $\tau = 74$ envelope samples. A comb filter is able to extract a frequency and its multiples by adding to the signal a delayed version of itself specified by the gain α and the delay τ. An example for such a comb filter is depicted in figure 1, its magnitude response for $\alpha = 0.8$ and $\tau = 50$ bpm is illustrated in figure 2. Based on the tatum grid our beat tracker is able to determine meter and tempo features by setting up narrow comb filters centered on multiple tempos of the tatum grid.

2.2 Harmonic Information

In order to incorporate the temporal harmonic structure of a song we use the chroma energy distribution normalized statistics introduced by Müller et al.

[Müller et al., 2005]. These features are based on chroma features which are computed using a fast Fourier transform with a window length of 372 ms and an overlap of 0.5 by taking into account a psychoacoustic model using A-weighting filtering as within the beat tracking according to DIN EN 61672-1:2003-10 and by decomposing the audio signal into frequency bands representing the semitones which are defined for equal temperament as

$$f_i = f_0 \cdot 2^{i/12} \qquad f_0 = f(\text{A0}) = 27.5\,\text{Hz} \tag{1}$$

with $15 \leq i \leq 110$ (corrisponding to the notes C2–B9) and therefore covering 96 semitones (8 octaves). In order to overcome deficient recordings due to mis-arranged recording settings or intentional manipulations of the sound impression, pitch correction is applied. A long term frequency analysis computes the prominent frequency f_p and determines a factor c

$$c = \frac{f_p}{f_r} \tag{2}$$

with

$$f_r = \underset{f_i}{\operatorname{argmin}} \left\| \frac{f_p}{f_i} - 1 \right\| \tag{3}$$

Next, all semitones f_i are multiplied with the factor c to correct their pitch. In order to allocate the frequencies to the semitones a nearest neighbor approach is applied which implies the use of Gaussian bells $g_i(x)$ centered at f_i given by

$$g_i(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{\left(\frac{x-f_i}{f_i-f_{i-1}}\right)^2}{2\sigma^2}} \qquad \sigma = 0.125 \tag{4}$$

Now we normalize the resulting sub-bands s_i by dividing each one belonging to the same octave O by the sum of these sub-bands according to

$$\hat{s}_i = \frac{s_{i,O}}{\sum s_{i,O}} \qquad s_{i,O} = s_i \in O \tag{5}$$

In a final step we add up all sub-bands corresponding to the same relative pitch class, for example for the chroma C we compute $\overline{s}_1 = \hat{s}_{15} + \hat{s}_{27} + \ldots + \hat{s}_{99}$, and normalize the resulting values

$$v_i = \frac{\overline{s}_i}{\sum \overline{s}_i} \qquad 1 \leq i \leq 12 \tag{6}$$

Due to the fact that the local chroma features are too sensitive concerning articulation effects and local tempo deviations we extend the chroma features following Müller et al. [Müller et al., 2005] and apply to each component of $\mathbf{v} = (v_1, \ldots, v_{12})$ a quantization function Q defined as

$$Q(a) := \begin{cases} 4 & \text{for} \quad 0.4 \quad \leq a \leq 1 \\ 3 & \text{for} \quad 0.2 \quad \leq a < 0.4 \\ 2 & \text{for} \quad 0.1 \quad \leq a < 0.2 \\ 1 & \text{for} \quad 0.05 \quad \leq a < 0.1 \\ 0 & \text{for} \quad 0 \quad \leq a < 0.05 \end{cases} \tag{7}$$

In the next step, we convolve 11 consecutive quantized chroma vectors $Q(\mathbf{v}(i))$ component-wisely using a Hann window resulting in a weighted 12-dimensional features vector including temporal harmonic information. As the information changes due to the windowing being quite slow, down-sampling with a factor of 4 is applied. The resulting feature vectors are referred to as chroma energy distribution normalized statistics (CENS) which we will denote from now on as $\mathbf{v} = (v_1, \ldots, v_{12})$. A comparison between the two types of features is visualized in figure 2.2.

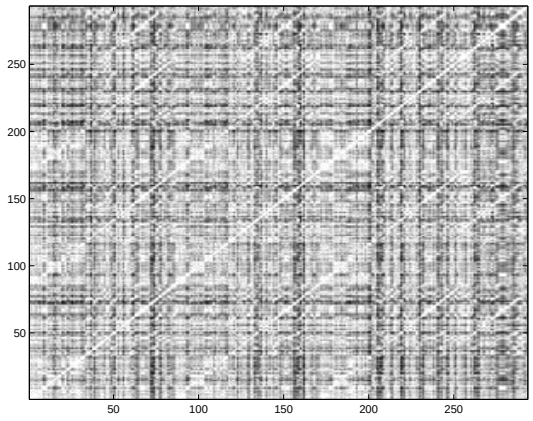

Fig. 4. Similarity matrix for Adriano Celentano - Azzurro. Bright diagonals illustrate a high similarity between two segments.

3 Chorus Extraction

3.1 Similarity matrix

In line with Foote et al. [Cooper and Foote, 2002] we compute a $N \times N$ similarity matrix $\mathbf{S}$ based on the normalized scalar product

$$\mathbf{S}(i,j) = \frac{\langle \mathbf{v}(i), \mathbf{v}(j) \rangle}{\|\mathbf{v}(i)\| \cdot \|\mathbf{v}(j)\|} \tag{8}$$

In terms of music visualization we are looking for bright diagonals (we mean diagonal segments parallel to the main diagonal, c.f. figure 4) in this matrix which correspond to similar segments in a song. Thus, we use an edge filter given by

$$F_{Diag}(i,j) = \begin{cases} 1 & \text{for} \quad i = j \\ c & \text{for} \quad 0 < |i-j| \leq b \\ 0 & \text{for} \quad |i-j| > b \end{cases} \tag{9}$$

with $1 \leq i, j \leq 20$, $b = 5$ and $c = -\frac{2}{17}$, to extract these bright diagonals from the similarity matrix. After a normalization step a threshold δ is subtracted from the filtered image resulting in the matrix $\hat{\mathbf{S}}$ in order to reduce noise that is generated by the edge filtering. δ corresponds to the highest value which is exceeded by at least $10 \cdot N$ values of the filtered image. In a subsequent step, we create a binary matrix $\mathbf{S}_b$ according to

$$\mathbf{S}_b(i,j) = \begin{cases} 1 & \text{for} \quad \hat{\mathbf{S}}(i,j) > 0 \\ 0 & \text{for} \quad \hat{\mathbf{S}}(i,j) \leq 0 \end{cases} \tag{10}$$

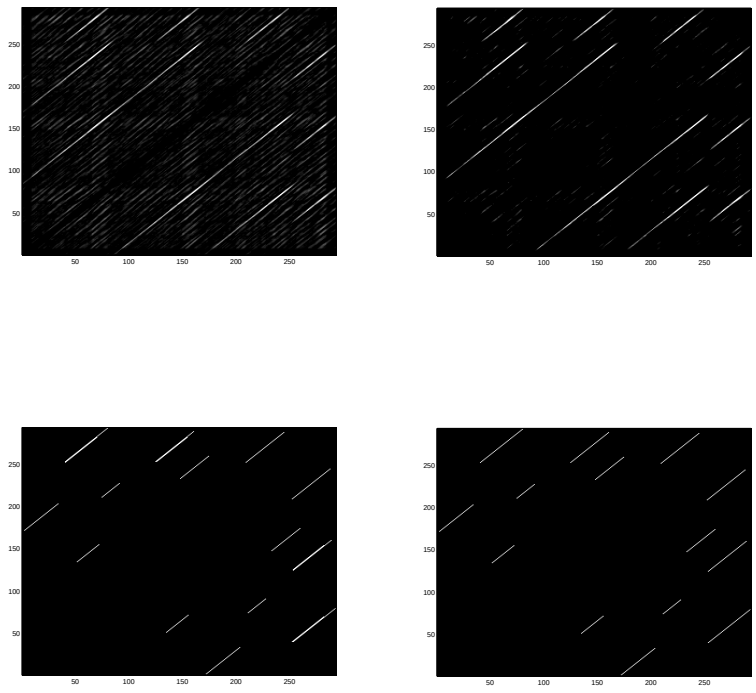

Fig. 5. Similarity matrices after the processing steps: first, edge filtered (top left), then dynamic thresholding (top right). From the resulting matrix $\hat{\mathbf{S}}$, respectively its binary representation $\mathbf{S}_b$, ROI are determined by length and characteristics of each segment (bottom left). Last, adjacent segments are combined (bottom right).

3.2 Regions of interest

Now we determine regions of interest (ROI). Starting and ending points of potential chorus sections are extracted by the following approach: Let $d(i, j)$ denote

the temporal derivate along a diagonal segment $\mathbf{S}_b(i+1, j+1) - \mathbf{S}_b(i, j)$. In a first step segment bounds are estimated by setting starting points at i and at j if $d(i, j) > 0$ and corresponding ending points at i and j if $d(i, j) < 0$. In order to correct these preliminary bounds we introduce a counter c_k for each segment k and define a threshold δ_{Sim} which corresponds to the highest value exceeded by at least $0.1 \cdot N^2$ entries of the matrix $\mathbf{S}$. Starting at the middle (x, y) of each segment, we increment c_k, if $\mathbf{S}(x, y)$ falls below δ_{Sim}, and decrement c_k if it exceeds δ_{Sim} up to a minimal counter value of $c_k = 0$. If c_k is smaller than C we process the next value $(x, y) := (x - 1, y - 1)$. Otherwise we stop and save the corrected starting point $(x + C, y + C)$. Next, we apply these steps for the other directions with $(x, y) := (x + 1, y + 1)$ starting again at the middle of segment k gaining the corrected ending point $(x - C, y - C)$. The algorithm with $C = 4$ has delivered the best performance in practice. The described process is depicted by an example in figure 5.

In order to reduce the amount of regions of interest we define lower and upper limits for the segment length l. A dynamic lower bound given by

$$l \cdot m_{Sim} > 8.7\,\mathrm{s} \tag{11}$$

where m_{Sim} is denoting the mean similarity of the segment in the similarity matrix $\mathbf{S}$. This has proven as an optimal choice to eliminate short repeating sections containing non-relevant segments. Further, a static upper bound given by 29.1 s has shown good results to distinguish between chorus sections and longer sections such as verse or verse plus chorus. In a last step we combine adjacent segments as they do not contain any additional information.

We now define an audio thumbnail by taking the best remaining segment regarding its mean similarity m_{Sim}, as we assume chorus sections to be the most similar sequences among the regions of interest. In order to evaluate the regions of interest and to provide multiple thumbnails for each song, we extract the 3 best segments.

4 Results and discussion

Our approach was tested on a database containing 360 songs of different genres with an overall duration of 23 h 47 min. The database consists on the one hand of 110 songs belonging mainly to rock music and oldies and on the other hand of 250 songs indexed by genre each with 50 songs covering electronic dance, pop, rock, german folk and pop music and oldies.

The evaluation was performed by comparing the initial positions of the extracted thumbnails to those manually annotated. If the deviation between these positions was at most T_{max} the extracted audio thumbnail was assumed correct. The notation Top$X$ represents the percentage of the songs where one of the X best thumbnails was correctly extracted. Table 1 shows the results for a maximal allowed deviation of $T_{max} = 1, 2$ and 3 s. Most of the not correctly extracted thumbnails represent a characteristic part of the song as well, such as

T_{max}[s]	Top1 [%]	Top2 [%]	Top3 [%]
1	22.6	37.8	45.8
2	48.6	67.2	73.3
3	60.6	76.1	81.4

Table 1. Correctly extracted audio thumbnails for different maximal deviations

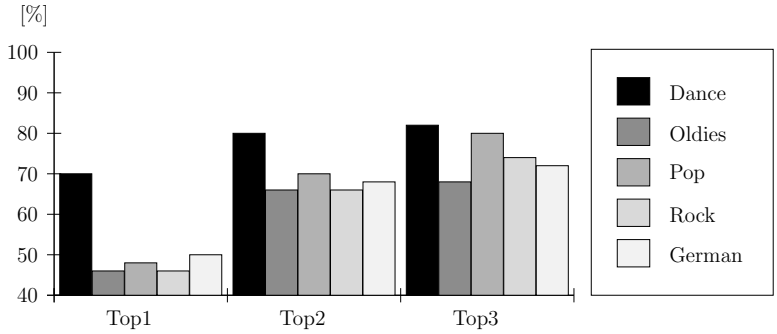

Fig. 6. Correctly extracted audio thumbnails depending on the evaluated genres for $T_{max} = 2\,\mathrm{s}$

the chorus section with a deviation slightly above T_{max}, a non-equivalent repetition of the chorus or the beginning of verse or bridge. Unfortunately, no other work provides quantitative results of the generated audio thumbnails in terms of deviation from the actual chorus sections. Therefore, no objective comparison is possible at this time.

Figure 6 illustrates the results specific to the genre. The notably higher performance for electronic dance results from the fact that electronic music provides higher similarities due to perfect accordance of the electronically produced tones. Further, the musical structure is mostly simpler and fewer variations are found.

5 Conclusion and outlook

Within this paper we presented an effective approach for automatic extraction of audio thumbnails based on rhythmic structure and harmonic similarity analysis. Experimental test runs provided promising results, especially for electronic dance music where we were able to determine the chorus section for 70 % of the songs with a maximal deviation of $\pm 2\,\mathrm{s}$. Likewise, we could "'compress"' one day of music to roughly half an hour of thumbnails.

In future works the algorithm can easily be extended by additional modules to increase the performance by incorporating progression structure or by classifying vocal and non-vocal sequences.

References

[Abdallah et al., 2005] Abdallah, S. A., Noland, K., Sandler, M., Casey, M., and Rhodes, C. (2005). Theory and Evaluation of a Bayesian Music Structure Extractor. In *Proc. 6th ISMIR*, pages 420–425.

[Aucouturier et al., 2005] Aucouturier, J.-J., Pachet, F., and Sandler, M. (2005). "The way it Sounds": Timbre Models for Analysis and Retrieval of Music Signals. *IEEE Transactions on Multimedia*, 7(6):1028–1035.

[Aucouturier and Sandler, 2001] Aucouturier, J.-J. and Sandler, M. (2001). Segmentation of Musical Signals Using Hidden Markov Models. In *Proc. of the Audio Engineering Society 110th Convention*.

[Bartsch and Wakefield, 2001] Bartsch, M. A. and Wakefield, G. H. (2001). To Catch a Chorus: using Chroma-Based Representations for Audiothumbnailing. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 15–18.

[Burges et al., 2004] Burges, C. J. C., Plastina, D., Platt, J. C., Renshaw, E., and Malvar, H. S. (2004). Duplicate Detection and Audio Thumbnails with Audio Fingerprinting. Technical Report MSR-TR-2004-19, Microsoft Research (MSR).

[Cooper and Foote, 2002] Cooper, M. and Foote, J. (2002). Automatic Music Summarization via Similarity Analysis. In *Proc. 3rd ISMIR*, pages 81–5.

[Foote, 1999] Foote, J. (1999). Visualizing Music and Audio using Self-Similarity. In *Proc. 7th ACM Int. Conf. on Multimedia (Part 1)*, pages 77–80.

[Goto, 2006] Goto, M. (2006). A Chorus Section Detection Method for Musical Audio Signals and its Application to a Music Listening Station. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1783–1794.

[Jehan, 2005] Jehan, T. (2005). Hierarchical Multi-Class Self Similarities. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 311–314.

[Logan and Chu, 2000] Logan, B. and Chu, S. (2000). Music Summarization Using Key Phrases. In *Proc. ICASSP*, volume 2, pages 749–752.

[Müller and Kurth, 2006] Müller, M. and Kurth, F. (2006). Enhancing Similarity Matrices for Music Audio Analysis. In *Proc. ICASSP*, volume 5, pages 9–12.

[Müller et al., 2005] Müller, M., Kurth, F., and Clausen, M. (2005). Chroma-Based Statistical Audio Features for Audio Matching. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 275–278.

[Peeters et al., 2002] Peeters, G., Burthe, A. L., and Rodet, X. (2002). Toward automatic music audio summary generation from signal analysis. In *Proc. 3rd ISMIR*, pages 94–100.

[Schuller et al., 2007] Schuller, B., Eyben, F., and Rigoll, G. (2007). Fast and Robust Meter and Tempo Recognition for the Automatic Discrimination of Ballroom Dance Styles. In *Proc. ICASSP*, volume 1, pages 217–220.