MUSIC STRUCTURE DISCOVERY IN POPULAR MUSIC USING NON-NEGATIVE MATRIX FACTORIZATION

Florian Kaiser and Thomas Sikora
Communication Systems Group
Technische Universität Berlin
{kaiser, sikora}@nue.tu-berlin.de

ABSTRACT

We introduce a method for the automatic extraction of musical structures in popular music. The proposed algorithm uses non-negative matrix factorization to segment regions of acoustically similar frames in a self-similarity matrix of the audio data. We show that over the dimensions of the NMF decomposition, structural parts can easily be modeled. Based on that observation, we introduce a clustering algorithm that can explain the structure of the whole music piece. The preliminary evaluation we report in the the paper shows very encouraging results.

1. INTRODUCTION

Music structure discovery (MSD) aims at characterizing the temporal structure of songs. In the case of popular music, this means classifying segments of a music piece into parts such as intro, verse, bridge, chorus or outro. Knowing this musical structure, one can introduce new paradigms in dealing with music collections and develop new applications such as audio thumbnailing and summarization for fast acoustic browsing, active listening (audio based retrieval and organization engines), song remixing or restructuring, learning semantics, etc.

In the past years, MSD has therefore gained an increasing interest in the music information retrieval community. This also led to the constitution of common evaluation data sets and evaluation campaigns (MIREX 09) that strongly stimulate the research in this field.

1.1 Previous work

Structure in music can be defined as the organization of different musical forms or parts through time. How we define musical forms and what builds our perception of these forms is however an open question, and MSD algorithms that have been proposed yet mainly differ in the way they answer those questions. However, Bruderer gives in [2] a general understanding of perception of structural boundaries in popular music, and shows that perception of structure is mainly influenced by a combination of changes in timbre, tonality and rhythm over the music pieces. Therefore, MSD algorithms generally aim at finding similarities and repetitions in timbre, tonality and rhythm based descriptions of the audio signal.

In [4], Foote and Cooper addressed the task of music summarization and proposed to visualize and highlight these repetitions in the audio signal through a self-similarity matrix. The audio signal is therefore parametrized through the extraction of audio features and the similarity between each frame is then measured. Thus using different audio features and similarity measures, most MSD algorithms are a processing of such a self-similarity representation.

In [13], the author distinguishes two categories of structure in the self-similarity matrix: the state representation and the sequence representation. The state representation defines the structure as a succession of states (parts). Each state is a succession of frames that show similar acoustic properties and therefore forms blocks in the self-similarity matrix. This representation is closely related to the notion of structural parts in popular music (intro - verse - chorus - outro), in which the acoustical information does not vary much. Algorithms based on state representation usually start with a segmentation by audio novelty score method [5]. The segments are then merged together with mean of hierarchical clustering, spectral clustering, or HMM.

On the other hand, the sequence representation considers series of times (frames), that are repeated over the music piece. The sequence representation is more related to musical concepts such as melody, progression in chords and harmony. Algorithms based on sequence representation look for repetitions on the off-diagonals of the self-similarity matrix. Matrix filtering of higher-order matrix transformations [14] can also be applied to the self-similarity matrix in order to emphasize off-diagonals. One of the main drawbacks of the sequence representation is that the structure of the music piece can not be fully explained unless all sequences are repeated at least once.

1.2 Approach

Non-negative matrix factorization (NMF) is a low-rank approximation technique that was first introduced in [9]. It is known for extracting parts-based representation of data, that strongly relates to some form of inherent structure in the data. Therefore, it has been successfully used in

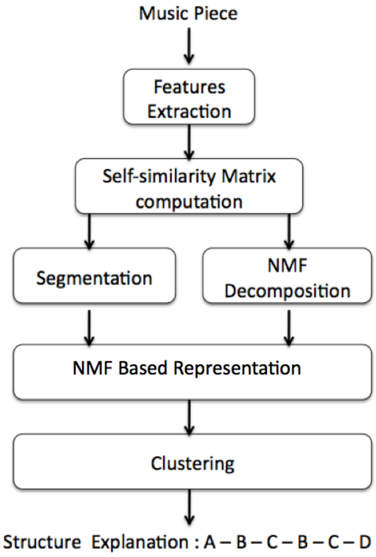

Figure 1. Overview of the proposed music structure discovery system

a wide range of multimedia information retrieval applications such as text summarization [9] or sound classification [1]. Moreover, Foote et al. showed in [3] that decomposing the self-similarity matrix of a video stream via NMF could help separating visually similar segments . We propose to extend the approach of Foote to music data.

Defining structural parts as acoustically similar regions like in the state-representation, we apply NMF to the self-similarity matrix. We show that such structural parts can easily be discriminated over the dimensions of the obtained decomposition. With a clustering approach, we are thus able to merge together similar audio segments in the NMF decomposed matrices, and explain the structure of the whole music piece.

In the next section, we provide a detailed description of our system. Evaluation metrics, data set and results are presented in section 3. Section 4. concludes the paper.

2. PROPOSED METHOD

An overview of our system is shown in Figure 1. In this section each individual block of the system is described.

2.1 Feature Extraction

We first extract a set of audio features that are likely to model variations between different musical parts. As mentioned in the introduction, perception of structural boundaries in music is mostly influenced by variations in timbre, tonality and rhythm [2]. However, few rhythmical changes occur between parts in our evaluation data set (see section 3.) and we thus only focus on the description of timbre and tonality. Nevertheless, the reader might refer to [11] for interesting work also using rhythmical clues for structure discovery.

Timbre properties of the audio signal are described by extraction of the following features: the first 13 MFCC coefficients, spectral centroid, spectral slope and spectral spread.

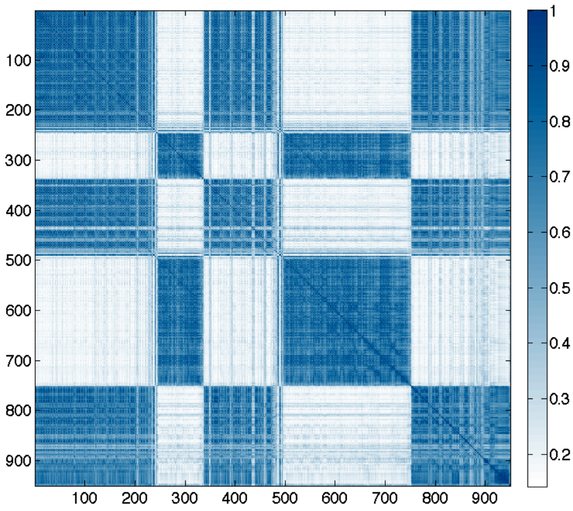

Figure 2. Self-similarity matrix computed on the timbre-related features using the exponential variant of the cosine distance. Audio file : ”Creep” by Radiohead

Tonality can be associated to the concepts of melody and harmony. Songs in a popular music context are however very diverse and a melody extractor would hardly be robust over a whole set of popular songs. We thus only focus on the description of harmonic properties through the extraction of the chroma features. Chroma features are 12 dimensional, each element corresponding to a pitch-class profile of a 12 scaled octave.

The frame analysis is performed with mean of a window size of 400 ms and a hop size of 200 ms. Each feature is normalized to mean zero and variance one.

Timbre-related features and chroma features are stored in two different feature matrices and processed separately.

2.2 Self-Similarity Matrix

After parameterization of the audio, we measure the similarity between each signal frame in a self-similarity matrix $\mathbf{S}$. Each element s_{ij} is defined as the distance between the feature vectors $\mathbf{v}_i$ and $\mathbf{v}_j$, extracted over frames i and j. The cosine angle is used as a similarity measure :

$$d(\mathbf{v}_i, \mathbf{v}_j) = \frac{< \mathbf{v}_i, \mathbf{v}_j >}{||\mathbf{v}_i||||\mathbf{v}_j||} \tag{1}$$

As proposed in [3], an exponential variant of this distance is used to limit its range to [0,1] :

$$de(\mathbf{v}_i, \mathbf{v}_j) = exp(d(\mathbf{v}_i, \mathbf{v}_j) - 1) \tag{2}$$

As an example, we extracted the timbre-related features over the song ”Creep” by Radiohead. The resulting self-similarity matrix is shown in Figure 2. One clearly sees that structural information is conveyed by the self-similarity matrix. Regions of acoustically similar frames form blocks in the matrix and one can also distinguish repetitions of these blocks. This illustrates the state representation of structure, as explained in the introduction. In this specific example, there are few sequence repetitions to see on the off-diagonals. In fact, the clearness of such sequences in the self-similarity matrix pretty much depends on the nature of the song and the features that describe it (chroma features tend to highlight sequences). In our example, blocks are formed because of the strong presence of saturated guitar, which does not yield much timbre evolution within the structural parts.

2.3 Segmentation

Once the audio has been embedded in the self-similarity matrix $\mathbf{S}$, a segmentation step is needed to estimate potential borders of the structural parts. Therefore the self-similarity matrix is segmented using the audio novelty score introduced in [5]. The main idea is to detect boundaries by correlating a Gaussian checkerboard along with the diagonal of the self-similarity matrix $\mathbf{S}$. The checkerboard basically models the ideal shape of a boundary in $\mathbf{S}$. The correlation values yield a novelty score in which local maxima indicate boundaries. We apply an adaptive threshold as described in [6] to detect these maxima and generate the segmentation.

2.4 Non-negative Matrix Factorization

Matrix factorization techniques such as principal components analysis (PCA), independent component analysis (ICA) or vector quantization (VQ) are common tools for the analysis of multivariable data and are mainly used for dimensionality reduction purposes. In [7], Lee and Seung introduced non-negative matrix factorization (NMF), and proposed to build the decomposition additively by applying a non-negativity constraint on the matrix factors. Unlike PCA and other factorization techniques, cancelation of the decomposed data is thus not allowed, leading to a parts-based representation of the data. An intuitive justification is that not allowing negative coefficients in the decomposition will prevent the loss of the physical meaning of the data.

Given an $n \times m$ non-negative matrix $\mathbf{V}$, NMF aims at estimating the non-negative factors $\mathbf{W}$ ($n \times r$) and $\mathbf{H}$ ($r \times m$), that best approximate the original matrix :

$$\mathbf{V} \approx \mathbf{WH} \tag{3}$$

$\mathbf{W}$ contains the basis vectors and $\mathbf{H}$ the encoding coefficients for the best approximation of $\mathbf{V}$. The rank of the decomposition r is usually chosen so that $(n+m)r < nm$, thus providing a compressed version of the original data.

In our approach, we compute NMF on the self-similarity matrix of the audio in order to separate basic structural parts. The algorithm we use for the estimation of the matrix factors $\mathbf{W}$ and $\mathbf{H}$ is detailed in [8]. In the next section, we describe how the factorization via NMF relates to structure and show how we can use that result for music structure discovery.

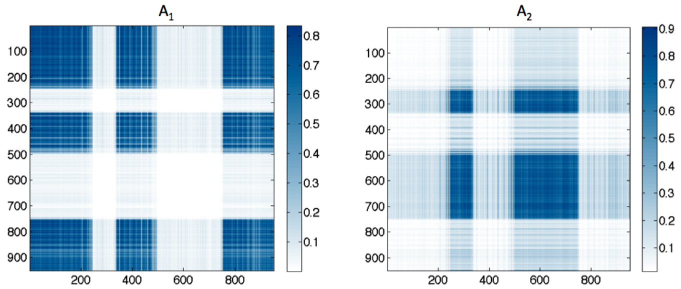

Figure 3. Matrices $\mathbf{A_1}$ and $\mathbf{A_2}$ obtained by NMF decomposition of the timbre self-similarity matrix of the song "Creep" (see Figure 2).

2.5 NMF based feature space

After decomposition via NMF, each element s_{ij} of $\mathbf{S}$ can be written as:

$$s_{ij} \approx \sum_{k=1}^{r} \mathbf{A}_k(i,j) \tag{4}$$

with

$$\mathbf{A}_k = \mathbf{W}(:,k)\mathbf{H}(k,:) \tag{5}$$

To illustrate how NMF can decompose data into basic structural parts, we compute NMF on the self-similarity matrix calculated over the song "Creep" by Radiohead. The rank of decomposition is set to 2 and the decomposed matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ are shown in Figure 3.
According to the timbre description in Figure 2, we can say that the music piece is composed of two main structural parts. Figure 3 shows that these two parts are strongly separated over the two dimensions of the NMF decomposition.

This suggests that each dimension of the NMF decomposition somehow relates the contribution of a structural part in the original data. In other words, that means that there is a specific energy distribution over the dimensions of the decomposition for each structural part.

Therefore it seems relevant to study for each segment how the energy is distributed over the matrices $\mathbf{A_k}$. In order to consider temporal dependencies, we choose to consider segments as successions of frames in matrices $\mathbf{A_k}$, and not as blocks. That means that each frame from the music piece is represented by its corresponding values over the diagonals of matrices $\mathbf{A_k}$. We thus define the feature vector $\mathbf{d}_k$, representing the contribution of the k^{th} decomposition over all frames:

$$\mathbf{d}_k = diag(\mathbf{A}_k) \tag{6}$$

Each frame can then be represented in the $(n \times r)$ feature space $\mathbf{D}$:

$$\mathbf{D} = [\mathbf{d}_1 \mathbf{d}_2 \ldots \mathbf{d}_r] \tag{7}$$

To illustrate this approach, we show an example with the song "Help" by The Beatles. The self-similarity matrix $\mathbf{S}$ computed on the timbre features of the song and the annotated structure are plotted in Figure 4. We compute the

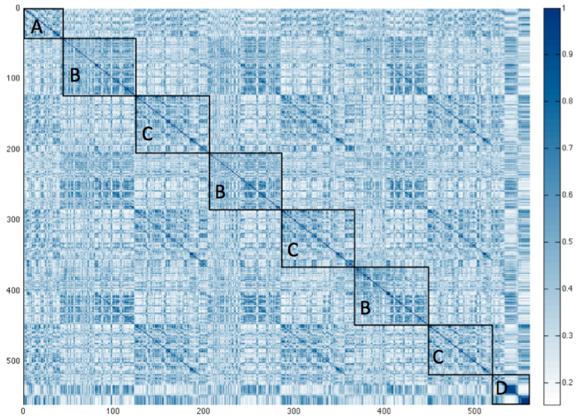

Figure 4. Self-similarity matrix computed on the timbre-related features for the song ”Help” by The Beatles. The black boxes indicate the annotated segments, with A being the intro, B the verse, C the chorus and D the outro.

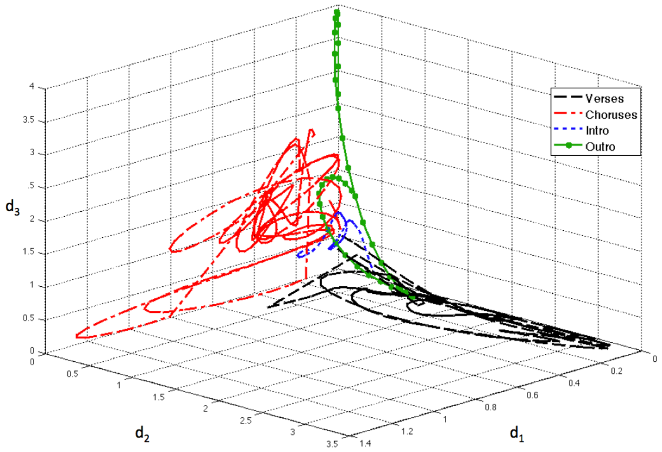

Figure 5. Representation of the structural parts of the song ”help” in the feature space **D**

NMF decomposition of **S**. For visualization purposes, the rank of decomposition is set to 3. In Figure 5, each of the annotated segments is represented in the feature space **D**. It is clear that structural parts chorus, verse and outro tend to be well represented over feature vectors $\mathbf{d}_1$, $\mathbf{d}_2$ and $\mathbf{d}_3$ respectively. In this case, we can say that each dimension of the NMF decomposition relates the contribution of a structural part. It is also interesting to note that segments of the same structural part seem to follow similar trajectories, suggesting that temporal dependencies should also be considered.

In classification problems, a feature space should provide good separability between classes. This means that the set of observations for a single class should have a small variance, whereas the set of all observations (for all the classes) should have a large variance. In that sense and according to Figure 5, representing segments in the feature space **D** should provide a good basis for structural classification.

2.6 Clustering

Each found segment is now represented in the NMF based feature space **D**. In order to merge together segments belonging to the same structural part, we propose to use a classical clustering approach. Therefore, the similarity between segments in **D** is measured with:

- The Bayesian information criterion (BIC)
- The Mahalanobis distance

The clustering is performed using the two measures separately. A comparison of the performance obtained with both measures is done in section 3. The clustering is done with a classical hierarchical approach.

3. EVALUATION

3.1 Data set

The evaluation data set consists of 174 songs from The Beatles, that were first manually annotated at Universistat Pompeu Fabra (UPF) [1] . Some corrections to the annotation were made at Tampere University of Technology (TUT) [2] . We call the data set *TUT Beatles*.

The structure in each music piece is annotated as a state representation and not as sequences (see section 1.). Each frame is thus affected to a label.

3.2 Metrics for the clustering evaluation

Evaluating the performance of a music structure detection algorithm is not simple. In fact musical structures are mostly hierarchical [10], meaning that the structure can be explained at different levels. For example, a structure A-B-A, could be also be described as abc-def-abc. We choose to evaluate our system using the pairwise precision, recall and F-measure. Therefore, we define F_a the set of identically labelled frames in the reference annotation, and F_e the set of identically labelled frames in the estimated structure. Pairwise precision, recall and F-measure, respectively noted P, R and F are then defined as :

$$P = \frac{|F_e \bigcap F_a|}{|F_e|} \tag{8}$$

$$R = \frac{|F_e \bigcap F_a|}{|F_a|} \tag{9}$$

$$F = \frac{2PR}{P + R} \tag{10}$$

These measures are not perfect for evaluating MSD algorithms because they do not reflect hierarchical aspects in the description of structure. Nevertheless, they give an idea of the global performance of the system.

[1] http://www.iua.upf.edu/%7Eperfe/annotaions/sections/license.html

[2] http://www.cs.tut.fi/sgn/arg/paulus/structure.html

	F-measure	Precision	Recall
Timbre	58.6%	58.1%	61.9%
Chroma	50%	46.5%	52.2%
both	53.6%	49%	55%

Table 1. Segmentation evaluation with the *TUT Beatles* database

3.3 Segmentation Evaluation

We evaluate the segmentation step with classical F-measure, precision and recall. Table 1 reports the performance of the segmentation computed on the timbre-related self-similarity matrix, the chroma-related self-similarity matrix and the sum of the two matrices.

The low precision rate in the segmentation suggests that the algorithm tends to over-segment the audio. In fact, structure is hierarchical and the annotation labels high level parts of the structure. The clustering might cope with that by reassembling segments from the same structural part.

3.4 Rank of decomposition

We ran a small experiment in order to choose a suitable rank for the NMF. Over a subset of ten songs from the database, we compute the similarity matrices. Varying the rank of NMF r from 3 to 12, we measure the separability between structural parts along each dimension d_i of $\mathbf{D}$. To do so, we compute the inertia ratio of the variance of d_i within segments belonging to the same structural part and the variance of d_i over the whole music piece [12]:

$$s(i) = \frac{\sum_{k=1}^{K} \frac{N}{N_k}(m_k - m_i)(m_k - m_i)'}{\frac{1}{N}\sum_{n=1}^{N}(d_i(n) - m_i)(d_i(n) - m_i)'} \qquad (11)$$

With K being the number of structural parts, N_k the number of frames in structural part k and N the total number of frames. m_i is the mean of d_i over the all piece and m_k the mean value of d_i over the k^{th} structural part. For a given rank of decomposition r, the separability is then measured as the mean of s:

$$sep(r) = \frac{1}{r}\sum_{i=1}^{r} s(i) \qquad (12)$$

We find a maximum of separability with a rank of 9 for NMF (see Figure 6). It is larger than the median number of annotated parts. In fact, as structure can be explained at different hierarchical levels, we don't expect the NMF decomposition to match the parts described in the annotation one-by-one.

3.5 Experimental set up for the clustering

Self-similarity matrices are computed over the timbre and chroma features separately. As shown in Table 1, segmentation using the timbre features provides better performances. Therefore, in the evaluation of the clustering step, we only use the segments positions extracted over the timbre-related self-similarity matrix. We propose four strategies to evaluate our clustering approach. For the three first strategies, the NMF based feature space is obtained by decomposition of the timbre-related self-similarity matrix (labeled as "Timbre"), the chroma-related self-similarity matrix (labeled as "Chroma") and the sum of the two matrices (labeled as "Fusion 1"). We also study a second fusion strategy where similarity between segments is computed separately in the timbre and chroma related feature spaces and then summed for the clustering algorithm (labeled as "Fusion 2").

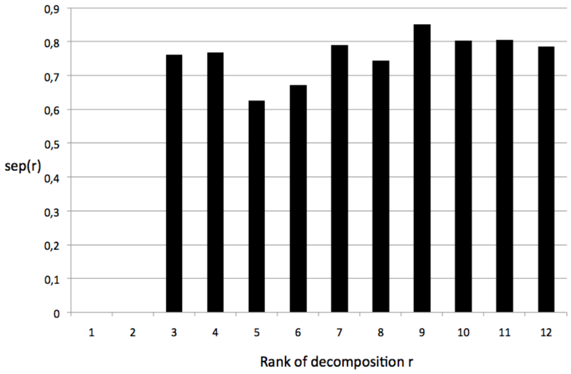

Figure 6. Separability of structural parts given different ranks of decomposition

We also compare the clustering obtained using the automatic segmentation described in section 2. (labeled "auto") and using the annotated segments (labeled "manual"). Finally, each configuration is run using the BIC (Table 2) and the Mahalanobis distance (Table 3) as similarity measure for the clustering algorithm.

The number of clusters is set to 4, which is the median number of annotated parts within a song in our evaluation data set.

3.6 Clustering Evaluation and Discussion

As a reference we use the system described in [11], that was also evaluated on the *TUT Beatles* database. The system is based on a description of the audio signal through MFCC, chroma and rhythmogram features. Each of these features is then used to estimate the probability for two segments to belong to the same structural part and a fitness measure of the description is introduced. A greedy approach is used to generate the candidate descriptions.

Evaluation of the whole system is reported in Tables 2 and 3, using BIC and Mahalanobis distance respectively. Compared to the reference system, our system shows slightly better F-measure rates. The interesting result is that we show significantly better recall rates. This suggests that our algorithm splits the parts in the annotation as sequences of sub-parts. This also explains why we don't match the precision rates in [11]. There again, the annotation relates a high stage of the structure hierarchy, and over-segmentation causes a lack of precision. Modeling sequences of basic parts in our algorithm might cope with that. This also explains the huge gain of performance when using the annotated segments for the evaluation.

Method	*Segmentation*	F	P	R
[11]		59.9%	72.9%	54.6 %
Timbre	auto	60.2%	64.7%	60%
	manual	76.1%	83.6%	72.6%
Chroma	auto	60.5%	**66**%	59.6%
	manual	80%	87%	76.6%
Fusion 1	auto	60.6%	65%	60%
	manual	78.7%	85%	76.4%
Fusion 2	auto	60.2%	64.7%	60%
	manual	80%	86.5%	77%

Table 2. Evaluation on *TUT Beatles*, BIC

Method	*Segmentation*	F	P	R
[11]		59.9%	72.9%	54.6 %
Timbre	auto	61%	62.4%	63.3%
	manual	78.4%	82.1%	78.3%
Chroma	auto	60.8%	61.5%	**64.6**%
	manual	76.6%	81.2%	75.7%
Fusion 1	auto	**62.1**%	63.6%	64.5%
	manual	77.8%	82.3%	77%
Fusion 2	auto	61%	62.4%	63.3%
	manual	78%	81.7%	78.2%

Table 3. Evaluation on *TUT Beatles*, Mahalanobis

Obviously, fusing both timbral and chroma description as in the ”Fusion 1” strategy makes sense and improves the overall performance of the system. Finally, using the Mahalanobis distance yields better performances than the BIC.

4. CONCLUSIONS

We introduced a music structure discovery method that uses the ability of NMF to generate parts-based representation of data. The evaluation conducted on the *TUT Beatles* data set shows that we are able to obtain slightly better performances than the reference system introduced in [11]. The improvements we obtain in the recall rates however suggest that there is still room for improvements. Moreover, the method used for the clustering of segments in the NMF based feature space only considers statistical similarity between the segments over time. We will consider modeling time dependencies between frames and thus model trajectories in the feature space instead of clouds of points. The NMF processing itself could also be enhanced by using sparse constraints on the matrix factors. Further evaluation on more diverse audio material will be done. The first results we obtained are however very encouraging.

5. ACKNOWLEDGMENT

This work was supported by the European Commission under contract FP7-21644 PetaMedia.

6. REFERENCES

[1] Emmanouil Benetos, Margarita Kotti, and Constantine Kotropoulos. Musical instrument classification using non-negative matrix factorization algorithms. In *ISCAS*. IEEE, 2006.

[2] Michael J. Bruderer, Martin F. McKinney, and Armin Kohlrausch. Structural boundary perception in popular music. In *ISMIR*, pages 198–201, 2006.

[3] Matthew L. Cooper and Jonathan Foote. Summarizing video using non-negative similarity matrix factorization. In *IEEE Workshop on Multimedia Signal Processing*, pages 25–28. IEEE Signal Processing Society, 2002.

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

[5] Jonathan Foote. Automatic audio segmentation using a measure of audio novelty. In *IEEE International Conference on Multimedia and Expo (I)*, page 452, 2000.

[6] A.L. Jacobson. Auto-threshold peak detection in physiological signals. In *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE*, volume 3, pages 2194–2195 vol.3, 2001.

[7] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, oct 1999.

[8] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization, July 21 2000.

[9] Ju-Hong Lee, Sun Park, Chan-Min Ahn, and Daeho Kim. Automatic generic document summarization based on non-negative matrix factorization. *Inf. Process. Manage*, 45(1):20–34, 2009.

[10] Namunu C. Maddage. Automatic structure detection for popular music. *IEEE MultiMedia*, 13:65–77, 2006.

[11] Jouni Paulus and Anssi Klapuri. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *IEEE Transactions on Audio, Speech & Language Processing*, 17(6):1159–1170, 2009.

[12] Geoffroy Peeters. Automatically selecting signal descriptors for sound classification. In *ICMC*, 2002.

[13] Geoffroy Peeters. Deriving musical structures from signal analysis for music audio summary generation: ”sequence” and ”state” approach. In Uffe Kock Wiil, editor, *CMMR*, volume 2771 of *Lecture Notes in Computer Science*, pages 143–166. Springer, 2003.

[14] Geoffroy Peeters. Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. In *ISMIR*, 2007.