Music Structure Segmentation

vorgelegt von
M.Sc. Florian Kaiser

Von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften
- Dr.-Ing. -

genehmigte Dissertation

Promotionausschuss :

Vorsitzender:	Prof. Dr. Thomas Wiegand	-	Technische Universität Berlin
Berichter:	Prof. Dr. Thomas Sikora	-	Technische Universität Berlin
Berichter:	Prof. Dr. Gaël Richard	-	Telecom ParisTech

Tag der wissenschaftlichen Aussprache: 26.4.2012

Berlin 2012
D83

Abstract

Music is above all about expression and communication. For musicians to develop their own musical expression and be able to communicate in a playing situation, music is therefore necessarily a structured language that produces structured musical discourses. One of the problems that Music Information Retrieval tackles and that we propose to study in this document is thus the estimation of structure in music.

In this study, the problem of musical structure estimation is formalized as the task of Music Structure Segmentation and aims at estimating the largest structural entities that compose a music piece. A verse in Popular music, a bridge in Jazz music or a movement in Classical music constitute such structural entities. As a front-end processing for audio indexing applications such as audio browsing, summarization or annotation, the task knows a growing interest in the Music Information Retrieval research community.

Research in this field has been particularly active since the introduction of audio self-similarity matrices and the visualization of the temporal evolution of the harmonic and timbral content of musical signals. We propose to estimate musical structures from such visualizations by means of their sparse decomposition with the Non-negative Matrix Factorization (NMF) algorithm. Indeed, we show that structural sections that are defined by sufficient acoustical homogeneity in terms of harmony or timbre can be easily separated with such a decomposition. We are then able to derive from the NMF of similarity matrices a mid-level representation of structure that allows for its robust classification. This approach is then further developed using the analogy of the visualization of structural sections in the similarity matrices with the segmentation of foreground and background objects in intensity images. Using image segmentation filtering techniques, we are indeed able to strengthen the structure representation and improve its segmentation. Finally, we propose a mid-level descriptor of tonal structures to allow for a better characterization of structural sections in similarity matrices. Analyzing such matrices with the NMF-based segmentation approach we significantly improve the structural segmentation. These results are illustrated with the comparative evaluation on popular and classical music.

Zusammenfassung

Die Musik ist in erster Linie ein Mittel des Ausdrucks und der Kommunikation. Um die Entwicklung eines spezifischen musikalischen Ausdrucks und das musikalische Spiel im Ensemble zu ermöglichen, baut Musik notwendigerweise auf einer strukturierten Sprache auf, und der musikalische Diskurs hat daher eine ihm eigene Form. Eine der Herausforderungen der Extraktion von Informationen aus Musiksignalen die dieser Schrift behandelt wird ist die Erkennung de zugrundeliegenden Musikstruktur.

Das Problem der Abschätzung der musikalischen Struktur wird hier als Identifizierung von Abschnitten auf höherer Ebene formalisiert. Man kann die Analogie mit der Strophe in der populären Musik, mit der Brücke zwischen zwei Teilen eines Jazzstückes oder mit dem Satz einer klassischen Komposition machen. Die Ergebnisse dieser Segmentierung kann sehr effektiv im Rahmen von Audioindizierung für Anwendungen wie Navigation durch große Datenbanken verwendet werden.

Die Forschung in diesem Bereich kennt seit der Einführung durch Foote von Audio-Ähnlichkeitsmatrizen eine starke Entwicklung, die die Visualisierung von Musik-Inhalten auf der Basis ihres Timbres oder Oberwelleninhaltes ermöglicht. Wir schlagen vor, solche Visualisierungen für die Schätzung der musikalischen Struktur mittels ihrer nicht-negativen Matritzenfaktorisierung zu analysieren. Wenn die musikalischen Struktur aus Teilen akustisch homogener Form besteht, zeigen wir dass eine Beschreibung auf mittlerer Ebene der musikalischen Struktur von einer solchen Faktorisierung abgeleitet werden kann und eine robuste Klassifikation ermöglicht. Daraufhin beschreiben wir, wie die Darstellung von Strukturinformation mit Ähnlichkeitsmatritzen durch Anwendung von Forschungergebnissen im Bereich der Bildsegmentierung verbessert werden kann. Im letzten Teil dieser Arbeit schlagen wir eine Beschreibung auf mittlerer Ebene des tonalen Kontextes vor, der das Ziel hat die Beschreibung der Homogenität die Teile einer musikalischen Struktur charakterisiert zu verbessern. Durch die Kombination dieses Ansatzes mit nicht-negativer Faktorisierung der Ähnlichkeitsmatritzen erhalten wir ein System für robuste Schätzung der musikalischen Struktur und evaluieren dieses System auf der Basis einer Datenbank die aus populärer und klassischer Musik besteht.

Résumé

La musique est avant tout un moyen d'expression et de communication. Afin de permettre l'élaboration d'une expression musicale propre et de rendre possible le jeu musical entre musiciens, la musique se construit nécessairement autour d'un langage structuré et un discours musical a donc une structure qui lui est propre. L'un des enjeux de l'extraction d'informations dans les signaux musicaux dont nous proposons l'étude dans de document concerne donc l'estimation de cette structure.

Le problème de l'estimation de la structure musicale sera formalisé ici comme la détection des sections haut niveau qui caractérisent la structure d'une pièce musicale. On pourra faire l'analogie avec le couplet en musique populaire, le pont entre deux parties d'un morceau de jazz ou encore les mouvements d'une composition classique. Les résultats d'une telle segmentation peuvent être efficacement utilisé dans le cadre de l'indexation audio pour des applications telles que la navigation à travers de larges bases de données.

Les recherches dans ce champ ont connu un fort développement suite à l'introduction par Foote des matrices de similarités audio qui permettent la visualisation d'un contenu musical sur la base de ses propriétés de timbre ou harmoniques. Nous proposerons d'analyser de telles visualisations pour l'estimation de la structure musicale au moyen de leur factorisation en matrices non-négatives. Nous montrerons que dans l'hypothèse où les parties qui composent la structure musicale sont caractérisées par une forme d'homogénéité acoustique, une description moyen niveau de la structure musicale peut être dérivée d'une telle factorisation et en permettre une classification robuste. Nous verrons ensuite comment améliorer la représentation de l'information de structure dans les matrices de similarités au moyen des résultats de la recherche en segmentation d'images. Enfin, le dernier volet de ce travail s'attachera à améliorer la description de l'homogénéité qui peut caractériser les parties d'une structure musicale en proposant un descripteur moyen-niveau du contexte tonal. En combinant cette approche à l'analyse par factorisation en matrices non-négatives des matrices de similarités, nous proposerons un système pour l'estimation de la structure musicale robuste et présenterons son évaluation sur une base de données composée de musique populaire et classique.

Acknowledgments

I would first like to express my honnest gratitude to my supervisor, Thomas Sikora, for his support and for giving me the opportunity to work on this thesis at the Communication Systems Group of the Technical University in the best conditions. I would also like to thank Gaël Richard for accepting to review this thesis so promptly.

My deepest gratitude goes to my dear friend and colleague Marina Arvanitidou. Her constant support through the steps of this PhD work and life has meant a lot to me. Thanks to my friend and colleague Engin Kurutepe, especially for opening the doors of Istanbul and Turkey for me. Birgit Boldin for making things so easy in the everyday life of the group. Many thanks to Jan Weil and Amjad Samour for the interesting discussions at the beginning of this work and to all of my colleagues at the Technical University of Berlin.

When I started this work I did not know much about Berlin. Since then, I have met here amazing people that have made me love this city much more than I had expected it. Persa, Despina, Bianca, Maria, Erkal, Cagri, Ahmet... Thanks guys for the great time here. Also, this long escape to Berlin wouldn't have been the same without the refreshing visits and welcomes of my friends in Lille.

I want to dedicate this work to my parents, Brigitte and Andreas. As the years go by, I realize how lucky I am to receive so much love from them. To my brothers, Louis, Theo, Gauderic and Johannes. I might be the oldest one, it seems that I still have a lot to learn from these guys. And to Julie, for having changed everything a couple of years ago.

Contents

Chapter 1

Introduction

This PhD Thesis studies the problem of Music Structure Segmentation by means of audio signal processing. The task that is addressed is thus the analysis of the musical properties of an audio signal that convey information about its structural entities such as a chorus in Popular music, a movement in Classical music, or a bridge in Jazz music. Computational methods for the practical segmentation of such properties are also studied.

This introduction chapter will first aim at explaining the motivation of this work in the actual music consumption context. Some applications of music structure segmentation will then be presented. Finally, the main contributions that were proposed within this thesis are summarized and the organization of the document is introduced.

1.1 Introduction

Together with the expansion of advanced electronic devices and social networks have emerged new ways in listening, producing, sharing and interacting with multimedia content. In the vanguard of this digital revolution, music has become a networked media that is available and produced everywhere, and is associated with an expanding list of additional content, e.g. artist information, recommendation, annotation, etc. And while hanging out in record stores to discover new music has become a quite marginal habit, our relation to music has deeply changed. The growth of online available music content has thus been massive in recent years, raising the issue of accessing this content without drowning in the quantity. Consequently, Music Information Retrieval (MIR) research has opened a wide range of new research areas that aim at describing and understanding musical audio signals automatically. New services for accessing, editing, annotating music have emerged and are taken up by a growing community of people. SoundCloud, Shaazam or last.fm are just a few examples of such services.

However, while the available computing power is always increasing, computers are still far from understanding and apprehending music as humans do. In particular, there is still a gap between the computer description of a musical event

and its interpretation by a listener in a given context. In this work, we address one particular aspect of the ongoing research in this field that is the estimation of musical structures, better known as the task of music structure segmentation. Music is indeed a media that is built around a highly structured language and thus produces structured musical discourses. For instance, the organization of tones in a chord sequence is supposed to be coherent with the rules of harmony just as grammar dictates the organization of words in a sentence. Harmony, instrumentation and rhythm offer countless possibilities for composers to articulate such musical sentences and develop their own musical discourse. In order for a computer to understand such discourses, it should naturally be able to understand their structure.

With the task of music structure segmentation we aim at estimating the highest level of musical structures. This level is the one of musical sentences that stand on their own as structural entities of the music piece and are not subparts of any other sections. We can use here the analogy of sections such as the verse or chorus in Popular music. Finding the boundaries of these sections within a music piece and explaining its structure as their combinations and repetitions defines the structural segmentation that we will be aiming at automatically derive from music in this work.

As illustrated with the application examples in the next section, estimating such musical structures is of great interest for the content-based discovery of large audio databases. Moreover, the estimation of structural information necessitates a deep understanding of musical contexts. We therefore think that progress in this research will also bring breakthroughs in future computer music interaction. In most music genres, structure is indeed essential in the experience of playing and enjoying music.

1.2 Applications

In itself, structural segmentation of music can have several understandings. It can first be the semantic audio segmentation of a music piece. Structural sections as well as their associated semantic labels then have to be detected. The segmentation can then be reduced to the detection of sections' boundaries and of their repetitions. This is the kind of segmentation this work addresses and that is also known as the task of music structure discovery. Finally, the segmentation can be limited to the detection of a single section, the chorus for example, or to the detection of the most repeated segment. Diverse applications follow from these different types of segmentations. Some of them are described in the following list:

- Active Listening: this is a direct application of the semantic audio structural

segmentation. Indeed, knowing the structure of a song, one can easily imagine browsing through its structural units. This allows the user to have a fast overview of the content of a song without having to randomly browse within it. The quality of such applications is however tied to the quality of the structural segmentation and especially to the performance of the labeling. Such an application with chorus detection was proposed in [Goto 2003].

- Audio Summarization: an other way to give a quick overview of a musical content is to generate its summary or thumbnail. Diverse strategies can be used. One can either use the chorus or most repeated section as thumbnail or concatenate the two main sections of the song for example.

- Audio Indexing: the structural information can be efficiently used to measure the similarity between audio samples of a database. For instance, the task of song versions identification uses the structure information as a feature to retrieve different versions of a same music piece. Also, the problem of the musical alignment of these versions can be tackled with music structure segmentation.

- Music Annotation: annotating a music piece is a time consuming process. Providing users with structural sections would definitely improve the quality and efficiency of this process. Moreover, web services such as SoundCloud have revealed a great interest for collaborative annotation. Providing users of such services with automatically generated tags would be interesting.

1.3 Summary of Contributions

As explained through this document, information about the structure of a music piece is mainly contained in its harmonic and timbral properties. Visualizing the evolution within a musical signal of such properties by means of the audio similarity matrix [Foote 1999] has thus largely contributed to the understanding and development of music structure segmentation systems. This work therefore proposes in a first stage a further analysis of such visualizations of the audio signal's content by means of sparse decomposition techniques, and in particular of the Non-negative Matrix Factorization. Such methods are known for deriving parts-based representation of data. The notion of part being closely related to structure, we show that such decompositions of audio similarity matrices are closely related to their structural content and can be used as a mid-level representation of the structural information for its segmentation.

Obviously, many algorithms for the estimation of musical structures are based on the analysis of visualizations of the structural information. We have thus been

interested in studying wether image processing could help for segmenting structure in the audio similarity matrices taken as intensity images. The next contribution of this thesis therefore proposes a framework for the enhancement of the structure visualization by means of image filtering techniques. This approach is then combined with the NMF-based structure segmentation algorithm.

Finally, the last contribution of this work is the definition of a mid-level descriptor of the tonal structure of an audio signal. Structural sections can be characterized by several musical properties, one of which is their harmonic or tonal context. While tonal structures shouldn't vary much within sections, their variations can be a strong indicator of structural changes. Thus, describing the local properties of the harmonic information, we are able to visualize a rather high-level structural information and improve the structure segmentation.

1.4 Organization of the Thesis

Chapter 2 is an introduction to Music Structure Discovery in which the main solutions that have been proposed in the literature are reviewed. Audio signal representations commonly used are first presented and methods for the structural segmentation are then distinguished according to two definitions of structure: the homogeneity-based and the repetition-based approaches. Metrics for the performance evaluation for music structure segmentation algorithms are then introduced.

Chapter 3 introduces our method for the segmentation of audio self-similarity matrices by means of non-negative matrix segmentation that was presented at the 11^{th} *International Symposium for Music Information Retrieval* (ISMIR) Conference in 2010 [Kaiser 2010]. The output of the Non-negative Matrix Factorization (NMF) of similarity matrices is used as a mid-level feature for the structure classification. Further contributions in the remainder of this thesis aim at improving the performance of this system by either post-processing the similarity matrices, or further describing the audio signal itself.

Chapter 4 details a method that consists in using filtering techniques from image segmentation research to sharpen the structure visualization in similarity matrices. With this technique we improve the accuracy of the mid-level structure representation derived from the similarity matrices and perform better structural segmentations. This work was presented at the 9^{th} *International Workshop on Content-Based Multimedia Indexing* (CBMI) 2011 [Kaiser 2011a].

Chapter 5 addresses the issue of the lack of coincidence between the temporal

scale of extraction of low-level audio features and the definition of a structure as a sequence of high-level musical segments. Exploiting properties of tonal structures in music, a mid-level feature that models the tonal context of mid-scale audio portions is proposed. This mid-level feature is then applied to music structure segmentation. This work was presented at the 14^{th} *International Conference on Digital Audio Effects* (DAFx) 2011 [Kaiser 2011b].

Chapter 6 proposes a comparative evaluation of all methods proposed in this thesis on a corpus composed of popular and classical music. Evaluation of the temporal segmentation and structure classification are separately provided and advantages and inconvenients of each method are discussed.

Chapter 7 summarizes and concludes this thesis. Future directions for music structure segmentation research are proposed.

Chapter 2

Music Structure Segmentation

"J'ai dit quelque part qu'il ne suffisait
pas d'entendre la musique,
mais qu'il fallait encore la voir"
Igor Stravinski

Contents

Since the introduction of music visualization by means of audio self-similarity matrices in [Foote 1999], the task of Music Structure Segmentation or Music Structure Discovery (MSD) has gained an increasing interest in the music information retrieval research community. In this chapter we will aim at drawing the context of this research and propose an overview of the methods that have been proposed yet.

2.1 Introduction

2.1.1 Problem Definition

Literally, structural segmentation of music means dividing an input audio signal according to its structural parts. A finer definition of the task then depends on the understanding we have of what defines sections of a musical structure. This is the first issue music structure segmentation research faces: there is no precise consensual definition of musical structure that would apply to any musical genre. For instance, is structure defined by harmony or instrumentation? Or can a definition of structure in popular music be applied in Jazz? There are no strict answers to these questions and the understanding of structure in music structure segmentation research is therefore rather general and does not really refer to any deep musical knowledge. A couple of features of musical structure can however be defined as follows.

First of all, the task of music structure segmentation deals with the entire structure of music pieces and we thus refer with the term "structure" to the concept of musical form in music theory. Of course musical forms vary with the genre and style of the considered music. Nonetheless, most musical forms can be divided into sectional units whose combination and eventually repetitions define the global structure of the music piece. Such units however appear at various hierarchical levels in music. Indeed, many aspects of music relate to some sort of structure. For example the structure of chords and melody phrases in a given harmony system, or the structure of rhythmical patterns. These are stand alone research topics that might help for the structural segmentation of music but are not sufficient to define its goal.

Here we consider sectional units defined by large-scale distinct and stand-alone segments of the music piece. Stand-alone means that the segments should have clear discrete boundaries in the audio signal and do not overlap with any other sections. This thus excludes the musical form of Fugue in which a theme recurs at different pitches and voices and might be superposed with decay for example.

Finally, most methods for the structural segmentation of music focus on the western popular music genre and aim with this understanding of structure at segmenting the music piece into sections that are commonly labeled as intro, verse or chorus. Algorithms have however weak assumptions on the musicological meaning of these sections and rather seek for homogeneous segments or repetitions. A part of this thesis also deals with the harmonic similarity between sections and mono-instrumental classical music pieces will thus also be studied (see description of databases in Appendix B).

The task of music structure segmentation algorithms thus consists in retrieving a sequence of musically meaningful sections within the audio signal. The output of such algorithms is illustrated in Figure 2.1 with the song *Drive My Car* by the Beatles.

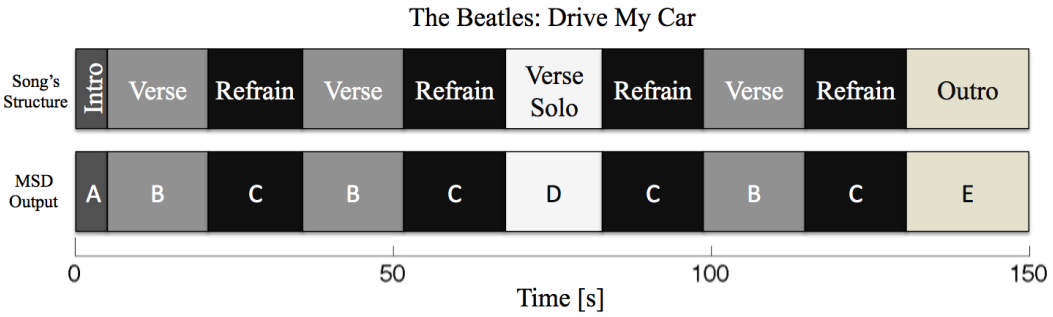


Figure 2.1: Output example of the structural analysis of the song *Drive My Car* by the Beatles

2.1.2 Organization of the Chapter

Algorithms for the structural segmentation of music take as input an audio signal and yield as output information about its temporal structure. From the waveform to this structural information a couple of subtasks can be identified. A rough overview of these subtasks is given in Figure 2.2. The main sections of this chapter will be organized accordingly.

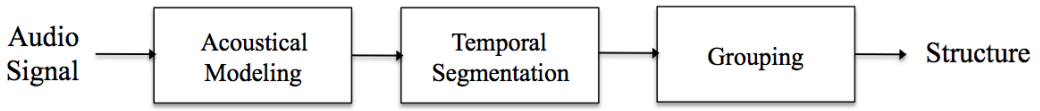


Figure 2.2: Subtasks of music structural segmentation

The first subtask consists in describing the acoustical properties of the audio signal that might characterize and/or discriminate musical sections. These properties are usually isolated referring to studies on the perception of boundaries in music. For example, a change of instrumentation can define such a boundary. The audio signal can then be accordingly described through the extraction of adequate audio features, i.e. descriptors of the audio signal's temporal or spectral properties that are extracted on the whole signal or short-term portions. This first step will be discussed in section 2.2.

The temporal evolution of the signal's acoustical properties can then be analyzed on the audio features vectors. Boundaries or time points of critical changes in the features distribution are usually detected by means of a self-similarity matrix visualization and novelty measure approaches. These methods will be discussed in section 2.3.

After the segmentation step, segments of the same structural part are merged together in the grouping step. Most of the proposed methods in the literature are driven by either one of the two following approaches: the "State" or the "Sequence"

approach. For the state approach, or homogeneity-based, structural-sections are assumed to have some uniformity in their acoustical content. Based on this assumption, segments are merged together or detected as stand-alone structural sections. On the other hand, sequence, or repetition-based, methods seek for repetitive patterns in the features distribution. Methods proposed in the literature for both approaches are discussed in section 2.4.

Finally, we introduce in section 2.5 the metrics that serve for the evaluation of music structure segmentation.

2.2 Audio Features for Structure Segmentation

The audio waveform of a music piece is a rather noisy signal that is hardly analyzable in its raw form. Indeed, the link from the audio waveform of the song *Drive My Car* (Figure 2.3) and its structure (Figure 2.1) is not really flagrant, and the amplitude of the waveform does not seem to be a satisfying descriptor. The first step of most MIR processing thus consists in deriving from the audio signals a couple of descriptors of their content in the temporal and spectral domains. Such audio descriptors, also called audio features, can model musically meaningful properties of the audio signal such as its pitch content, its sound color or its percussive degree. Along with the history of MIR, a large set of audio features have been proposed. In this section we will first discuss the acoustical properties that should be extracted in the context of music structure segmentation and then introduce a couple of audio features that might help for this task.

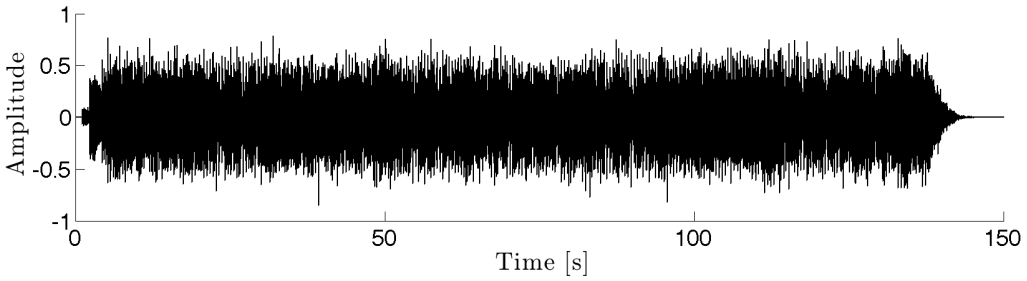


Figure 2.3: Audio waveform of the song *Drive My Car* by the Beatles

2.2.1 Perceptual Attributes of Structure

What discriminates sections or makes two segments sound similar in a music piece? Prior to applying any acoustical model on the audio signal, answering this question is the starting point of music structure segmentation. To this end, most methods refer to the work of Bruderer on perception of boundaries in western

popular music [Bruderer 2006]. Bruderer used a set of six songs that cover different western popular music styles. Each song is annotated at the beginning and ending of instrumental voices and at the end of harmonic cycles. Boundaries based on musicological considerations are also annotated. A group of eighteen subjects of diverse musical experience is then asked to listen to the songs and indicate time points on which they perceived a boundary by pressing a key. A second experiment let subjects judge of the saliency of the boundaries. The experiment showed that there was quite a large agreement between subjects on most boundaries but not all. Moreover, Bruderer noted a high correlation between the number of key presses on a boundary and its perceived salience. Based on this experiment and the comments of subjects on their perception of boundaries, Bruderer concludes that changes in music are mainly determined by a combination of changes in timbre and harmony. An interesting outcome of this work is also that perception of structure is not binary. This is a big issue for music structure segmentation algorithms that consider a time point in the audio signal as either indicating a boundary or not. He observed a wide range of perceived salience on the boundaries suggesting that finding a match between human annotation and automatic segmentation is hardly feasible.

Methods for the automatic segmentation of structure nevertheless extract audio features that describe timbral and harmonic properties of sounds. Some work also segment the music piece by finding rhythmical patterns (see Table 2.1).

2.2.2 Features Extraction

In order to analyze the temporal evolution of the audio content, the audio signal is first segmented in a sequence of overlapping frames. Feature vectors, either in the temporal or spectral domain, are then computed on single frames. Such features are denoted as low-level descriptors. Note that global descriptors can also be computed on the whole signal or by means of a statistical modeling of the feature frames sequence. To ensure a reasonable compromise between the temporal and spectral precision of the Short-Time Fourier Transform (STFT), the usual size of feature frames in audio signal processing is of 20 - 30 ms. A spectra-temporal resolution of a dozen of milliseconds is however unnecessary to capture the dynamics of musical sections, for which the time-scale of a note should rather be considered. Slightly higher-level features can be computed using the mean and variance of low-level descriptors. An alternative that is often used consists in extending the size of signal frames to a hundred of milliseconds. Synchronizing the feature frames on the estimated beat positions is also a musically meaningful way to extend the size of analysis windows [Bartsch 2005] [Eronen 2007] [Levy 2008].

A comprehensive overview of audio features extraction and their musical meaning can be found in [Peeters 2004b] and [Kim 2005].

2.2.3 Timbre Description

Timbre of a sound, or its tone color, denotes all properties of its frequency spectrum apart from its pitch and loudness. While two tones of different pitch can share the same timbre, two tones of the same pitch played by two different instruments differ in timbre. Timbre is rather a subjective concept than a deterministic property of sounds and is thus hardly describable from a signal point of view. However, studies on timbre and instrument classification have shown that perceptual attributes of timbre such as brightness or warmth could be related to the shape of the audio signal's spectral envelope [Wegener 2008]. Most of the features used for analyzing timbre are thus descriptors of the spectral shape. The most popular are the Mel Frequency Cepstral Coefficients (MFCC's) [Imai 1983] that describe the frequency spectrum transposed in a perceptual frequency scale. After introducing MFCC's we will review a set of other spectral features that were found to convey timbral information.

2.2.3.1 Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCC) features were first introduced in the speech processing community but became very popular in MIR processing as a poweful description of the global shape of the spectral envelope with a few coefficients. One of the main advantage is that it translates the linear frequency scale in a scale that is inspired from human perception of sound. Indeed perception and especially discrimination ability between close frequencies varies with the frequency. It means that two adjacent low frequencies can be easily discriminated, whereas two close high frequencies can not be discriminated. This is reflected in the mel frequency scale in which critical perceptual sub-bands are modeled (see Figure 2.4). MFCCs coefficients sum the Spectrum energy in each of these sub-bands and thus isolate the salient properties of the spectral shape..

Secondly, MFCCs use a Cepstrum representation of the Spectrum. This is motivated by the source-filter model that can be applied to human speech production, with a source signal s(t), i.e. the vocal tract, and a filter h(t), i.e. the mouth cavities (see illustration in Figure 2.5).

In the temporal domain, the filtering of the source signal is defined by the convolution of s(t) with h(t) :

$$x(t) = s(t) * h(t) \tag{2.1}$$

with x(t) the speech signal. In the frequency domain, the convolution becomes a product,

$$X(f) = S(f) \times H(f) \tag{2.2}$$

with X(f), S(f) and H(f) the spectrum of x(t), s(t) and h(t) respectively. The

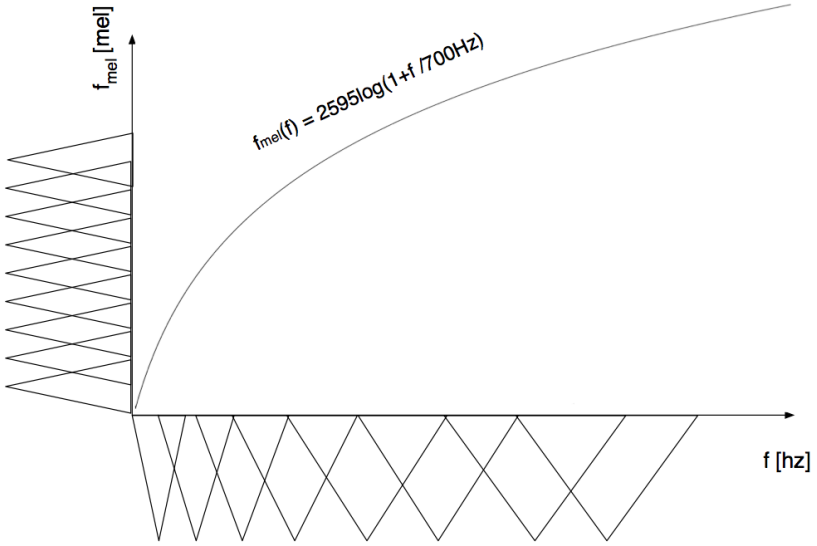


Figure 2.4: Mel Frequency Scale

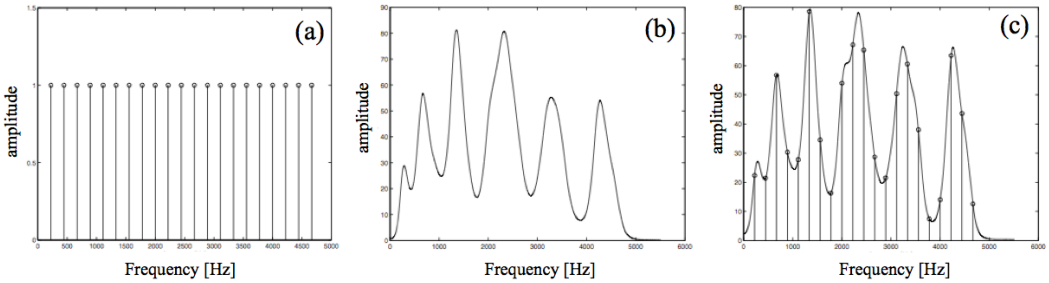


Figure 2.5: Source-filter model for speech production: (a) : Spectrum of the source (Vocal folds) ; (b) : Frequency response of the filter (Vocal Tract) ; (c) : Resulting spectrum (Speech)[Obin 2006]

contribution of the source and filter can be then easily discriminated using a logarithmic representation:

$$\log|X(f)| = \log|S(f)| + \log|H(f)| \tag{2.3}$$

The cepstrum coefficients c(n) are then defined as:

$$c(n) = DCT(\log|X(f)|) = DCT(\log|S(f)|) + DCT(\log|H(f)|) \tag{2.4}$$

$$c(n) = \frac{1}{N}\sum_{k=0}^{N} \log|X(k)| \exp^{\frac{2j\pi}{N}} \tag{2.5}$$

2.2.3.2 Spectral Features

Within studies on musical instrument classification, a couple of audio features have been found to be strongly correlated with timbral properties of sounds. Most of these features describe the shape of the spectral envelope which is known to be related with timbre. We briefly introduce in this subsection the spectral centroid, spread and skewness that are computed as different order moments of the spectrum distribution and therefore relate to its shape. We also introduce the spectral flatness.

Spectral Centroid the spectral centroid can be defined as the barycenter, or first order moment, of the spectrum. It is known in the literature to be related with the brightness of sounds. The higher the spectral centroid, the brighter the sound color. For its computation we consider the distribution of the normalized frequency ν and the normalized spectrum $a(\nu)$ as their observation probability:

$$Centroid = \mu = \int_0^1 \nu.a(\nu)d\nu \tag{2.6}$$

Spectral Spread this feature is a value of the spread of the spectrum around its mean value and is computed as the 2^{nd} order moment of the normalized frequency distribution:

$$Spread = \sigma^2 = \int_0^1 (\nu - \mu)^2 a(\nu)d\nu \tag{2.7}$$

with μ the spectral centroid.

Spectral Skewness the skewness measures the asymmetry of the normalized spectrum distribution around its mean value. We first compute the 3^{rd} order moment of the normalized frequency distribution:

$$m_3 = \int_0^1 (\nu - \mu)^3 a(\nu)d\nu \tag{2.8}$$

The spectral skewness is then computed as:

$$Skewness = \frac{m_3}{\sigma^3} \tag{2.9}$$

Spectral Flatness the spectral flatness measures the spectrum's tonality, in contrast to its noisiness. It can be computed on the whole frequency range or in sub-bands. Considering the amplitude spectrum $A(f)$ and the sub-band of N frequencies $F = \{f_1, ..., f_n\}$ it is computed as:

$$Flatness(F) = \frac{(\prod A(k))^{\frac{1}{K}}}{\frac{1}{K}\sum A(k)}, k \in F \tag{2.10}$$

2.2.3.3 Higher-level Features

Modeling the temporal evolution of the spectral envelope has also shown good results for the the task of instrument classification. Peeters introduced in [Peeters 2002b] dynamic features that capture short-time statistics of the spectral shape. Therefore, the audio signal is passed through a filter bank of N mel filters, and the Short-Time Fourier Transform (STFT) is computed for each output over a window of size L. Playing with the resolution of windows, dynamic features can capture timbral characteristics at different temporal scales.

2.2.4 Low-level Harmonic Description

Harmony relates to the organization of tones in a music piece and thus relates to musical concepts such as tonality, chords construction, melody, etc. Low-level features that might help for its description are based on the analysis of the pitch content of the audio signal. Pitch is an auditive sensation that results from a periodicity in the audio signal but that is not uniquely defined by physical frequencies. It is however often associated to the musical tones of the western music scale in the MIR community and can be estimated both in the temporal and spectral domains. Methods in the temporal domain are mainly based on the autocorrelation of the signal. This however only robustly works for highly periodical signals. Because of the particular properties of musical signals, i.e. presence of harmonics and timbre, pitch is preferentially estimated in the frequency domain.

The relative frequencies of tones of the western music scale are geometrically spaced. For example A4 is at the frequency of 440 Hz, A5 is at 880 Hz and A6 at 1760 Hz. Frequency bins estimated by means of the Fourier Transform are however equally distant and do not map the pertinent musical frequencies. Moreover, the resolution of the transform remains constant for all frequency bins. This motivated the introduction of the Constant-Q Transform (CQT) for the spectral analysis of musical signals [Brown 1991]. The CQT is a filter bank of geometrically spaced filters centered at frequencies f_k:

$$f_k = f_0.2^{\frac{k}{b}} \tag{2.11}$$

with b the number of filters per octave. The bandwidth δf_k of the k-th filter is then defined as:

$$\delta f_k = f_{k+1} - f_k = f_k(2^{\frac{1}{b}} - 1) \tag{2.12}$$

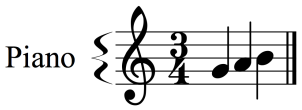
thus providing a constant ratio between frequency bin and frequency resolution:

$$Q = \frac{f_k}{\delta f_k} = \frac{1}{(2^{\frac{1}{b}} - 1)} \tag{2.13}$$

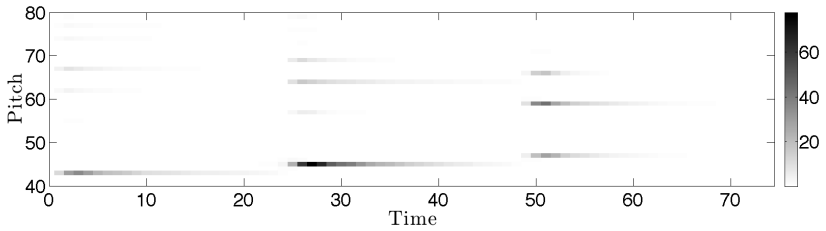
Choosing a number of 12 filters per octave and a reference fundamental frequency f_0 of the chromatic scale, each bin of the CQT is centered on the frequency

of a musical tone. The CQT of musical notes thus form constant patterns and pitch estimation can be performed by means of pattern recognition techniques.

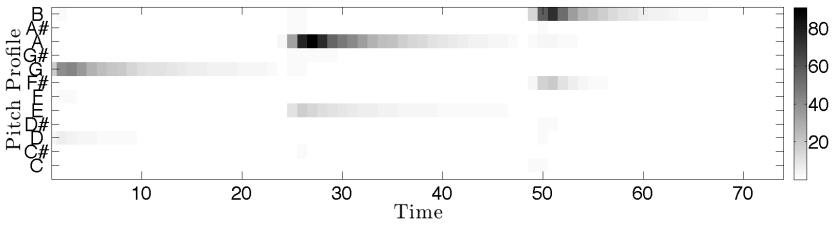
To reduce the dimensionality of the pitch content, it is often preferred to extract the chroma vectors that use the concept of pitch class [Müller 2007a]. Each note of the western music scale defines a particular pitch class profile in the frequency spectrum composed of its fundamental frequency f_0 and all its corresponding harmonics. Each bin of the twelve-dimensional chroma vector thus sums the contribution of a pitch class in the CQT transform.



(a) Score



(b) Pitch Estimation



(c) Chroma Vector

Figure 2.6: Example of the pitch and chroma estimation with the audio recording the note sequence G - A - B played on acoustic piano

For the computation of chroma features, we use in this thesis the Matlab Chroma Toolbox proposed in [Müller 2011]. Pitch content is first estimated by means of a constant Q multirate pitch filter bank that decomposes the signal into 88 frequency bands centered on musical tones from $A0$ to $C8$. The filter bank can be shifted accordingly to a tuning deviation estimation. The chroma bins for each frame are then obtained by adding the pitch energy in sub-bands corresponding to the 12 pitch profiles of the well-tempered western music scale.

We illustrate in Figure 2.6 the pitch and chroma estimation on a recording of a sequence of notes played on an acoustic piano. The dominant chroma bins extracted on each note correspond with the pitch class of the note.

2.3 Music Visualization and Segmentation

We have seen in the previous section how timbral and harmony related characteristics of sounds can be described by means of adequate audio features. A common approach to segment structural elements in these features distributions consists in visualizing their self similarity by means of the audio self-similarity matrix. We now introduce this technique and a method to derive a temporal segmentation of the audio signal from it. Enhancement of the structure visualization in similarity matrices is also discussed.

2.3.1 Self-Similarity Matrix

Music is made of repetitions of musical patterns and the question of measuring musical similarity is thus essential to its analysis. For example, parts of a popular music song are successively repeated in a given sequence that defines its structure. Inspired by the technique of recurrence plots introduced in [Kamphorst 1987] for detecting the recurrence of states in dynamical systems, Foote proposed in [Foote 1999] to visualize the similarity of a musical signal with itself by measuring the pairwise similarity of its audio feature vectors in a Self-Similarity Matrix (SSM).

In the SSM, each feature frame of the audio signal, and thus each time point, is compared to all other feature frames by means of a suitable distance measure. The resulting representation is a square matrix as illustrated in Figure 2.7.a. Each element s_{ij} of the SSM $\mathbf{S}$ is defined as the distance between the feature vectors $\mathbf{v}_i$ and $\mathbf{v}_j$, extracted over frames i and j. Euclidean and cosine distance are usually used as similarity measures between feature vectors. For our illustrations, we use the cosine angle:

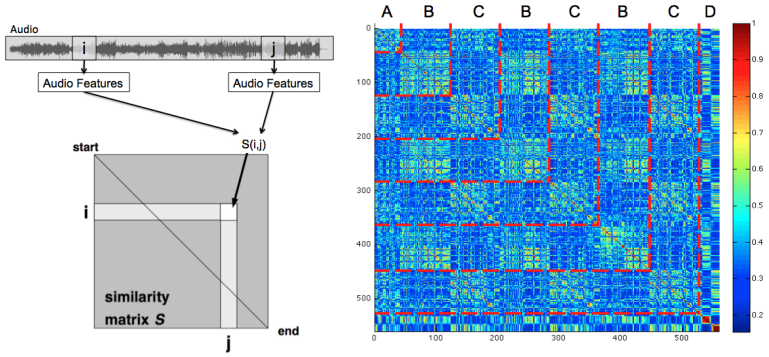
$$s_{ij} = d(\mathbf{v}_i, \mathbf{v}_j) = \frac{< \mathbf{v}_i, \mathbf{v}_j >}{||\mathbf{v}_i|| ||\mathbf{v}_j||} \tag{2.14}$$

with $<,>$ the dot product and $||.||$ the vector nom. As proposed in [Cooper 2002], an exponential variant of the cosine distance is also often used to limit its range to [0,1] :

$$de(\mathbf{v}_i, \mathbf{v}_j) = exp(d(\mathbf{v}_i, \mathbf{v}_j) - 1) \tag{2.15}$$

Points of high luminance in the similarity matrix indicate a high similarity between the compared time points. Hence, if musical sections in the audio file consist of segments of acoustically similar frames, they will form dense regions of high similarity in the matrix. There is thus a strong correlation between structural sections as musical objects and visual objects formed in the similarity matrix.

In Figure 2.7.b, we show an example with the song *Help* by *The Beatles*. The similarity matrix is computed on the MFCC features with the cosine distance. The matrix thus measures the similarity in timbre between each time instants of the song. In order to see how the timbre property can relate to structural information,



(a) Similarity Matrix Computation

(b) Example with the song Help with MFCC audio features and cosine distance

Figure 2.7: Similarity Matrix Computation and example with its annotated structure

the structure of the song (A, B, C and D) is also annotated on the similarity matrix. In the visualization, frames belonging to a same section are highly similar and form regions of high luminance in the matrix. We also note that part A, which is the introduction, mainly consists in similar sonorities as part C. The conclusion of the song (part D) is a short repetition of parts B and C.

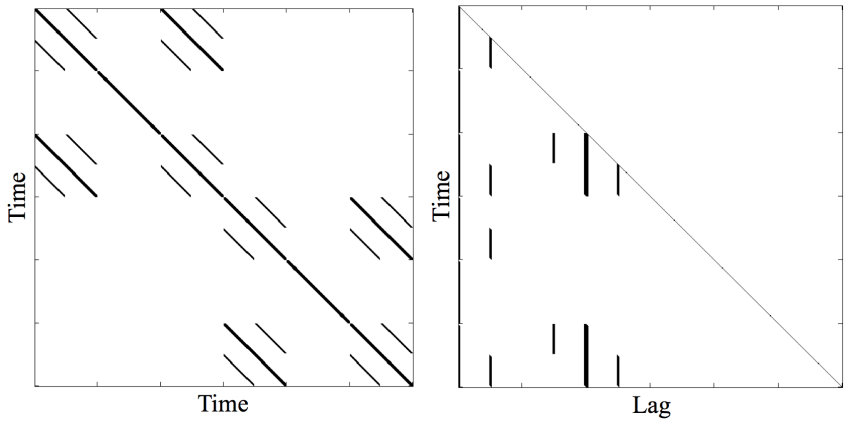
Time-Lag Matrix When looking for repetitive patterns in the similarity matrix, it might be useful to rotate the coordinate of the matrix so that patterns appear on the horizontal or vertical axes. This can be done by computing the time-lag matrix [Goto 2003] which is calculated as follow:

$$L(l,t) = S(t, t-l) \tag{2.16}$$

with L the time-lag matrix and S the original similarity matrix. In case of a repetition in the audio signal, a sequence in the audio features distribution is repeated with a constant lag. Such a repetition appears as a line on the vertical axis of time-lag matrix whose abscissa indicate constant lags (see Figure 2.8).

2.3.2 Audio Novelty Measure

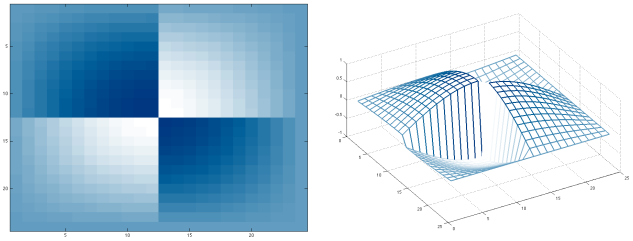
A boundary between two sections in a music piece indicates a significant change in the audio signal. The perceptual studies on music structure perception led by Bruderer in [Bruderer 2006] indicated that salient characteristics of such changes are to be found in the change of instrumentation (timbre) or harmony. As a description of the self-similarity of audio feature vectors that relate to harmony or timbre, such boundaries have a particular shape in the similarity matrix. Indeed, a boundary is characterized by self-similar past samples, self-similar future samples, with past



(a) Similarity Matrix (b) Corresponding time-lag matrix

Figure 2.8: Similarity matrix and its time-lag version

and future being unsimilar.



(a) 2D view (b) 3D view

Figure 2.9: Gaussian checkerboard kernel

Using this property of similarity matrices, a popular method to retrieve potential boundaries between musical sections consists in using the audio novelty method introduced in [Foote 2000]. The idea is to detect boundaries by correlating a gaussian checkerboard kernel along with the main diagonal of the similarity matrix S. The kernel is a $M \times M$ matrix that has a 2×2 checkerboard like structure. Radial smoothing by means of a gaussian function is also often used (see Figure 2.9). This can be seen as an image processing pattern recognition approach, where the checkerboard is a template of the ideal shape of a boundary in S. As illustrated in Figure 2.10, the correlation of S with the kernel yields a novelty score in which local maxima indicate boundaries.

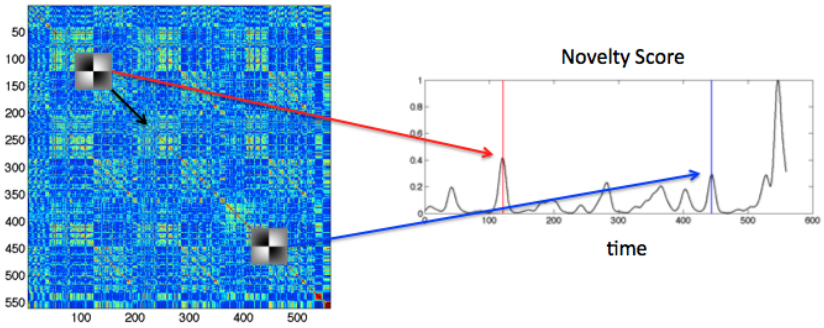


Figure 2.10: Segmentation of the similarity matrix with the Audio Novelty approach

2.4 Structure Segmentation

A comprehensive overview of music structure segmentation methods that was proposed in [Paulus 2010] is reproduced in Table 2.1. There are a variety of tasks associated to music structure segmentation:

- **Segmentation:** estimate boundaries between structural sections
- **Chorus detection:** detect boundaries of the most repeated section
- **Thumbnailing:** extract the most repeated section
- **Repetitions detection:** detect all repeated segments
- **Full structure:** merge together all segments belonging to the same structural section
- **Summarization:** assemble a portion of all structural sections to generate of summary of the music piece

Independently of the final task and as previously mentioned, there are two main approaches to the structural segmentation of music: the state and sequence approaches [Peeters 2004a]. We introduce the assumptions on music structure underlying these two approaches and then review methods that have been proposed in the literature for both cases. Further comparative studies and evaluation of these methods can be found in [Paulus 2010] and [Smith 2010].

State Approach A state denotes the homogeneity within structural segments with respect to a particular acoustic property. Say this acoustic property is instrumentation, e.g. a vocoder is used in the chorus and not the other sections, values of feature frames extracted on the chorus define a timbre state that relates to its particular instrumentation. The structure is then explained as the succession of states in the audio signal. With this assumption, feature frames within sections are rather self-similar and therefore form particular patterns within similarity matrices. With the state approach, structural sections are ideally represented in the SSM as blocks of high-similarity with low similarity with unrelated other sections. A state thus defines itself within its boundaries and not with its repetitions.

Sequence Approach On the other hand, the sequence approach assumes that the structure defines itself by its repetitions. Rather than characterizing the self-similarity within a feature segment, repetitions of feature frames sequences are sought. Say a melody C - D - E is repeated, chroma vectors extracted on both occurrences have low self-similarity but are exact repetitions. Such repetitions in the feature vectors from lines of high similarity on the off-diagonals of the similarity matrices.

2.4.1 State or Homogeneity-based Methods

The ideal shape of a similarity matrix in the state hypothesis is shown in Figure 2.11. This representation of structure is well-suited with the audio novelty score previously introduced. For similarity matrix based structure segmentation methods, novelty functions are thus often applied as a first step to estimate boundaries between sections.

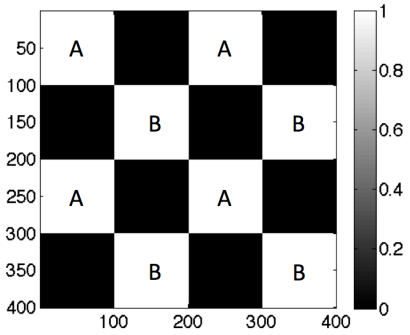


Figure 2.11: Ideal similarity matrix in the state hypothesis

Content of segments is then analyzed to build homogeneous clusters that explain the structure of the music piece. Segments can for example be modeled as normal distributions and the pairwise similarity between segments measured by means of the Kullback-Leibler divergence. Clustering techniques such as hierarchical clustering [Logan 2000], spectral clustering or k-means clustering [Peiszer 2007] are then applied to retrieve the song's structure from the pairwise segments distance matrix. Inspired from image processing, authors in [Cooper 2003] factorize the distance matrix between segments by means of a Single-Value Decomposition (SVD) to detect clusters of similar feature frames. A similar approach in the context of video summarization was introduced in [Cooper 2002] where a representative thumbnail of the video is extracted by means of the Non-negative Matrix Factorization (NMF) of a similarity matrix computed on visual features. The segment of highest energy among the NMF decomposed vectors is chosen as the thumbnail. The method we proposed in [Kaiser 2010] and that we describe in Chapter 3 is inspired from this work.

Musical patterns within the audio signal are detected in [Jehan 2005] by means of multiple similarity matrices that relate to different hierarchy-level of the musical structure, starting from a frame level through note onsets and beats until detected short-term musical patterns.

An other popular method to segment the structure in the feature vectors under the state assumption consists in using Hidden Markov Model (HMM) [Aucouturier 2005] [Peeters 2004a] [Levy 2008]. Musical segments compose the states of a HMM and segment's transitions the transition probabilities between states. One of the drawback of applying HMM on the feature vectors is that it might capture short-term musical events and detect rapid changes rather than long-term musical patterns. Indeed, even if there is a sort of acoustic homogeneity within segments, local changes in the spectral content are still captured in the feature vectors. A method that can be employed to cope with that consists in averaging the feature vectors over a longer time-scale. In [Peeters 2004a] authors introduce after a SSM-based segmentation a two step initialization of the HMM: a prior grouping of nearly identical segments that present up to 99% similarity (*initial states*) and a k-means clustering of the feature vectors (*middle states*). Authors find in these two steps a convenient way to estimate a reasonable number of states and prove better results than with random initialization. The HMM is trained on the middle states by means of the Baum-Welch algorithm. Decoding on the feature vectors with the HMM model is then done with the Viterbi algorithm. In [Abdallah 2005] and [Levy 2008] a large number of states is used for the HMM. The output states of the HMM are used as mid-level features for further temporal clustering with some constraints on the length of structural sections.

2.4.2 Sequence or Repetition-based Methods

Repetitions in the audio signal can be visualized in the SSM as stripes on the off-diagonals (see Figure 2.8.a), and methods thus aim at detecting theses stripes. Variations in timbre, dynamics, tempo, notes execution, etc. however often render the stripes noisy. Stripe enhancement has thus been proposed by means of low-pass filtering for smoothing and noise removal [Wellhausen 2003] [Bartsch 2005]. Goto proposed in [Goto 2003] to enhance the stripes in the time-lag matrix by means of a two-dimensional local filter and remove noise by means of a threshold inspired from the Otsu method [Otsu 1979]. Repetitions are then detected by integrating over time the constant-lag vertical lines in the time-lag matrix. Methods inspired from image processing are applied In [Lu 2004] and [Ong 2007]. Thresholding is first used to binarize time-lag matrices and stripes are then reinforced by means of morphological filtering. Morphological filtering is also applied in the work we proposed in [Kaiser 2011a] and is further described in Chapter 4.

The sequence structure of similarity matrices is enhanced in [Peeters 2007] by means of higher-order similarity matrices. The idea is that a repetition at time

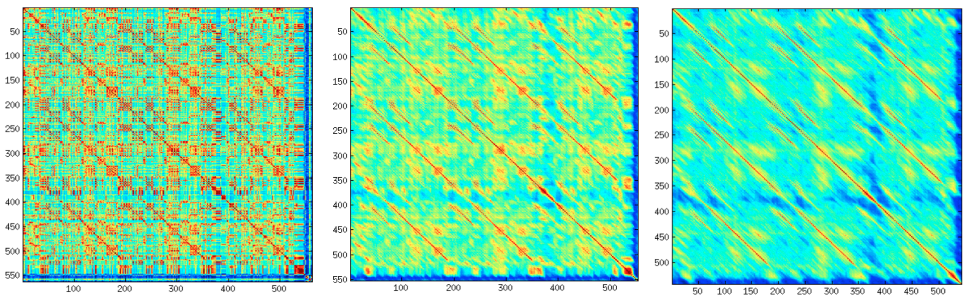
instant t_z of the time instants t_x and t_y should imply a high similarity in the similarity matrix S between t_x and t_y. Peeters uses this property and defines the 2^{nd} order similarity matrix $S_2(t_x, t_y)$ as the similarity between t_x and t_y through all possible t_z:

$$S_2(t_x, t_y) = \int S(t_x, t_z) S(t_z, t_y) dt_z \tag{2.17}$$

Repeated segments are then detected in a manner that is similar to [Goto 2003]. In [Mueller 2006], authors propose to enhance stripes by defining the *contextual similarity measure* d_L:

$$d_L(\mathbf{v}_i, \mathbf{v}_j) = \frac{1}{L} \sum_{l=0}^{L-1} d(\mathbf{v}_{i+l}, \mathbf{v}_{j+l}) \tag{2.18}$$

with L the length of the considered sequences of frames and $\mathbf{v}$ the feature vectors. Thus considering a larger observation horizon for the measure of similarity, one favors the visualization of sequences of similar frames. Hence, off-diagonals consisting of repetitions are amplified in the similarity matrix. We show an example in Figure 2.12. The standard similarity matrix is computed on the chroma features extracted over the song *Help* (see Figure 2.12.a). We then compute the contextual similarity matrices for the lengths L = 10 and L = 20 frames ($\approx$ 2.5 and 5 seconds respectively).



(a) Original Matrix (b) L = 10 frames (c) L = 20 frames

Figure 2.12: Contextual Similarity Matrices computed on the chroma features

One can similarly enhance the structure visualization in the similarity matrices using the dynamic features introduced in [Peeters 2002b]. Varying the resolution of the dynamic features, authors derive similarity matrices that either relate to short-term or long-term structures.

Repetitions can be alternatively detected in the feature sequences by means of pattern recognition techniques [Rhodes 2007].

2.4.3 Dynamic Time Warping

For both the state and sequence hypothesis, proposed methods for the structural segmentation measure the similarity between segments of the audio signal, either in the similarity matrix or with the corresponding feature vector sequences. A problem that thus often occurs is the alignment of two compared segments. To allow for the detection of repeated segments executed at different time or speed, Dynamic Time Warping (DTW) is used [Dannenberg 2002] and aims at finding the shortest path between two sequences. If the sequences are compared in a N-by-K matrix D, the shortest path is found as:

$$\tilde{D}(x,y) = D(x,y) + min \begin{cases} D(x-1,y) \\ D(x,y-1) \\ D(x-1,y-1) \end{cases}$$

Applying DTW when comparing segments allows to have a better precision in the detection of repetitions.

2.4.4 Information Rate

The Audio Oracle algorithm is a recently proposed alternative approach for repetitions detection that is inspired from information theory [Dubnov 2011] [Cont 2011]. Especially, authors use the notion of mutual information between the current sample x_n and past events $x_{past} = x(1, ..., n-1)$. In a classic information theory context, the mutual information is then defined as follows:

$$I(x_n, x_{past}) = H(x_n) - H(x_n|x_{past}) \tag{2.19}$$

with $H(x)$ the entropy of the variable x with probability distribution $P(x)$:

$$H(x) = -\sum P(x) log_2 P(x) \tag{2.20}$$

Entropy is a measure of the uncertainty of a given random variable. In an audio analysis context, authors put aside statistical consideration on the audio signal's statistics and keep the idea of uncertainty in the signal and relate it to the notion of complexity. Doing so and in order to measure the amount of change between past and present, authors introduce the Information Rate:

$$I(x_n, x_{past}) = C(x_n) - C(x_n|x_{past}) \tag{2.21}$$

with $C(.)$ a measure of the audio signal's complexity. In Audio Oracle, complexity of the audio signal relates to a compression algorithm and the number of bits that is needed to encode the signal. Musical events can then be automatically segmented with the Information Rate I.

Publication	Task	Acoustic Features	Approach	Method
[Aucouturier 2005]	full structure	spectral envelope	state	HMM
[Barrington 2010]	full structure	MFCC / chroma	state	dynamic texture
[Bartsch 2005]	thumbnailing	chroma	sequence	stripe detection
[Chai 2005]	full structure	chroma	sequence	stripe detection
[Cooper 2003]	summarization	magnitude spectrum	state	segment clustering
[Dannenberg 2002]	repetitions	chroma	sequence	dynamic programming
[Eronen 2007]	chorus detection	MFCC + chroma	sequence	stripe detection
[Foote 1999]	visualization	MFCC		self-similarity matrix
[Foote 2000]	segmentation	MFCC		audio novelty
[Goto 2003]	repetitions	chroma	sequence	stripe detection (RefraiD)
[Jehan 2005]	pattern learning	MFCC+chroma+loudness	state	hierarchical SSM
[Jensen 2007]	segmentation	MFCC+chroma+rhythmogram	state	diagonal blocks
[Levy 2008]	full structure	mpeg7 timbre descriptor	state	temporal clustering
[Logan 2000]	key phrase	MFCC	state	HMM / clustering
[Lu 2004]	thumbnailing	constant-Q spectrum	sequence	stripe detection
[Maddage 2006]	full structure	chroma	state	rule-based reasoning
[Marolt 2006]	thumbnailing	chroma	sequence	RefraiD
[Mauch 2009]	full structure	chroma	sequence	greedy selection
[Müller 2007b]	multiple repetitions	chroma statistics	sequence	stripe search & clustering
[Ong 2007]	full structure	multiple	sequence	RefraiD, morphological filtering
[Paulus 2006]	repeated parts	MFCC+chroma	cost func.	
[Paulus 2009]	full description	MFCC+chroma+rhythmogram	cost func.	
[Peeters 2004a]	full structure	dynamic features	state	HMM, image filtering
[Peeters 2007]	repeated parts	MFCC+chroma+spec. contrast	sequence	stripe detection
[Rhodes 2007]	hierarchical structure	timbral features	sequence	string matching
[Shiu 2005]	full structure	chroma	state	state model stripe detection
[Turnbull 2007]	segmentation	various		various
[Wellhausen 2003]	thimbnailing	mpeg7 timbre descriptor	sequence	stripe detection

Table 2.1: Overview of methods for music structural segmentation (credits to [Paulus 2010])

2.5 Music Structure Segmentation Evaluation

Independently of the chosen approach, music structure segmentation algorithms provide with a set of boundaries and labels within the studied music piece. This output segmentation is thus closely related to a state description of the musical structure. For this segmentation to be evaluated and compared with a groundtruth, music pieces must be previously manually annotated. Example annotations can be found in the description of our databases in Chapter 6. While the produced annotation mostly fit a state representation to match the output format, there are however no standardized annotation procedure. Indeed and as already mentioned in the introduction of this chapter , there are no commonly accepted definition of structure and the characterization of structural parts thus remain rather vague. Musical structures being highly hierarchical [Maddage 2006], one particular issue with the evaluation is the fact that the hierarchy-level of the annotated structure is quite high while algorithms might describe a lower level of this hierarchy.

While this remain an ongoing issue in the music structure segmentation research community, a couple of consensual metrics have been developed and employed for the MIREX structural segmentation evaluation task [1]. These metrics will used for the evaluation of our algorithms and are now introduced.

2.5.1 Boundary Retrieval Evaluation

The temporal segmentation step produces a set of boundaries between sections. These are evaluated by their comparison with the annotated boundaries by defining the following terms:

- **True Positives**: number of correctly retrieved boundaries
- **False Positives**: number of unexpected retrieved boundaries
- **False Negatives**: number of missing boundaries

The precision and recall are then defined as in equations 2.22 and 2.23.

$$Precision = \frac{true\_positive}{true\_positive + false\_positive} \tag{2.22}$$

$$Recall = \frac{true\_positive}{true\_positive + false\_negative} \tag{2.23}$$

For classification tasks, the precision is interpreted as the probability that a retrieved document is relevant, whereas the recall is the probability that a relevant document is retrieved in a search. Both precision and recall can then be combined in the F-Measure defined in equation 2.24.

$$F_\beta = (1 + \beta^2)\frac{precision.recall}{\beta^2.precision + recall} \tag{2.24}$$

With β a scalar. In our comparative evaluation in Chapter 6 we use $\beta = 2$.

[1] http://www.music-ir.org/mirex/wiki/2011:Structural_Segmentation

2.5.2 Frame Labeling Evaluation

2.5.2.1 Pairwise F-Measure

A largely consensual method to evaluate the frame labeling consists in using the pairwise precision, recall and F-measure introduced in [Levy 2008]. Considering 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, the pairwise precision, recall and F-measure, respectively noted P, R and F are then defined as:

$$P = \frac{|F_e \bigcap F_a|}{|F_e|} \tag{2.25}$$

$$R = \frac{|F_e \bigcap F_a|}{|F_a|} \tag{2.26}$$

$$F = \frac{2PR}{P+R} \tag{2.27}$$

An inconvenient of these metrics is that they are quite dependent on the temporal segmentation performance. Plus, they do not exactly reflect hierarchical aspects in the description of structure. For example the over-segmentation of a state 'A' explained as the sequence 'a-b-c' will be penalized in the evaluation. It is however the most robust evaluation measure of the performance of a music structure segmentation system up to now. Pairwise precision and recall rates can be interpreted in terms of over- and under- segmentation. An over-segmentation causes an increase of the precision rates and low recall rates. Under-segmentation has in contrast the reverse effect.

2.5.2.2 Normalized Conditional Entropies

An other type of evaluation metric that is being commonly used for the evaluation of music structure segmentation relies on information theoretic background and uses the notion of conditional entropy illustrated in Figure 2.13. These measures were proposed to provide a segmentation evaluation independent of the number of labels given by the estimated segmentation and allow for a better comparison of systems.

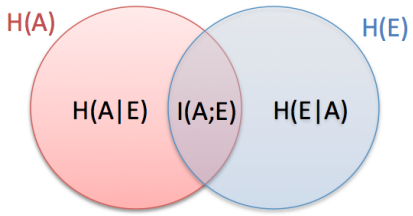


Figure 2.13: Conditional entropies $H(A|E)$ and $H(E|A)$ of two variables A and E, with $I(A;E)$ their mutual information and $H(A)$ and $H(E)$ their entropies

A normalized version of the conditional entropies was recently proposed in [Lukashevich 2008]. In a few words, the conditional entropy $H(A|E)$ measures the amount of information missing in the estimation, and $H(E|A)$ the amount of erroneous information. While turning to zero in case of a perfect segmentation, the two entropies are not higher bounded. Therefore the introduction of normalized conditional entropies.

Structural segmentations are seen as discrete time sequences of numeric labels, namely A and E for the annotated and estimated segmentations. The joint distribution p_{ij} of the two sequences and the marginal distributions for the annotated and estimated segmentations p_i^a and p_j^e are defined in equations 2.28 2.29 and 2.30.

$$p_{ij} = \frac{n_{ij}}{\sum_{i=1}^{N_a}\sum_{j=1}^{N_e} n_{ij}} \tag{2.28}$$

$$p_i^a = \frac{n_i^a}{\sum_{i=1}^{N_a}\sum_{j=1}^{N_e} n_{ij}} \tag{2.29}$$

$$p_j^e = \frac{n_j^e}{\sum_{i=1}^{N_a}\sum_{j=1}^{N_e} n_{ij}} \tag{2.30}$$

With N_a and N_e the number of different labels in the annotated and estimated segmentations respectively, n_{ij} the number of frames that simultaneously belong to the label i in the annotation and to the label j in the estimation, n_i^a the total number of frames belonging to the label i in the annotation and n_j^e the total number of frames belonging to the label j in the estimation. Conditional probability distributions of i given j and j given i is noted in equation 2.31.

$$p_{ij}^{a|e} = \frac{n_{ij}}{n_j^e} \qquad and \qquad p_{ji}^{e|a} = \frac{n_{ij}}{n_i^a} \tag{2.31}$$

The conditional entropies $H(E|A)$ and $H(A|E)$ are then defined as in equations 2.32 and 2.33.

$$H(E|A) = -\sum_{i=1}^{N_a} p_i^a \sum_{j=1}^{N_e} p_{ji}^{e|a} \log_2 p_{ji}^{e|a} \tag{2.32}$$

$$H(A|E) = -\sum_{j=1}^{N_e} p_j^e \sum_{i=1}^{N_a} p_{ij}^{a|e} \log_2 p_{ij}^{a|e} \tag{2.33}$$

It is proposed in [Lukashevich 2008] to normalize these conditional entropies. Given the annotated segmentation and the estimated number of labels N_e, the maximal conditional entropy $H(E|A)_{max}$ is obtained when the estimated labels are uniformly distributed over the annotation. This implies conditional distributions such that $p_{ij}^{a|e} = \frac{1}{N_e}$, resulting in a maximum conditional entropy as in equation 2.34.

$$H(E|A)_{max} = -\sum_{i=1}^{N_a} p_i^a \sum_{j=1}^{N_e} \frac{1}{N_e} \log_2 \frac{1}{N_e} = \log_2 N_e \tag{2.34}$$

Over-segmentation and under-segmentation scores are then defined as in equation 2.35 and 2.36 respectively.

$$S_o = 1 - \frac{H(E|A)}{\log_2 N_e} \tag{2.35}$$

$$S_u = 1 - \frac{H(A|E)}{\log_2 N_a} \tag{2.36}$$

A score So = 1 indicates that there was no over-segmentation in the song's structure clustering while a score Su = 1 indicates that there was no under-segmentation.

2.6 Chapter Summary

We have introduced the task of music structure segmentation and presented audio signal processing solutions. To summarize in a few words, most methods describe either timbral or harmonic aspects of the audio signal and then estimate the musical structure assuming either the state or sequence hypothesis. Methods for the evaluation of such segmentations were also introduced.

This introduction to MSD will serve as context for the methods proposed in the next chapters of this thesis. We now introduce in Chapter 3 a method that uses non-negative matrix factorization to segment states in similarity matrices.

Chapter 3

Non-negative Self-similarity Matrix Factorization

"The whole is greater than the parts"
Gestalt Psychology

Contents

With the state assumption, structural sections are represented in similarity matrices as blocks of high similarity, forming clear visual patterns. We aim in this chapter at benefitting from this regularity in the structure visualization and separate the contribution of each state in a similarity matrix by means of its Non-negative Matrix Factorization (NMF). Showing that such matrix factors can be used as mid-level features for the classification of structure, we introduce a system for music structure segmentation that shows performances comparable to the state of the art.

3.1 Introduction

Non-negative Matrix Factorization (NMF) is a low-rank approximation method that constraints the matrix factors to non-negativity and yields parts-based decomposition of data, allowing for a semantic interpretation of the decomposition. In its introduction in [Lee 1999], this property of the factorization was illustrated with the decomposition of face images. While standard methods such as the Single Value Decomposition (SVD) tended to yield rather holistic decompositions of the original data, authors showed that structural components of the face images, i.e. nose, mouth, eyes, etc. were separated in the dimensions of the NMF decomposition. Hence, assuming that perception of a whole is based on the perception of its parts, this kind of decomposition potentially mimics human perception. NMF has therefore raised an increasing interest in multimedia content analysis research in recent years. In Music Information Retrieval, NMF has been generally applied to the decomposition of audio spectrums. Reducing the spectrum to a set of basis spectral patterns and their time activations, NMF was successfully applied to audio source separation for example.

In this chapter we intend to further develop an application of NMF that was proposed in [Cooper 2002] in the context of video summarization. To generate a summary, authors defined the most representative video segment as the longest visually self-similar segment. Computing a similarity matrix on visual features, authors visualized such a segment as a block of high similarity and showed that the NMF decomposition of the similarity matrix could be used for its segmentation.

This definition of the most representative segment is rather similar to the definition of states introduced in music structure segmentation and we claim here that the decomposition of similarity matrices by means of NMF can help for their segmentation. Quality of the mapping between such a matrix decomposition and structural states in the similarity matrix naturally depends on the ability of the chosen audio features to characterize the inner homogeneity of structural sections. Moreover, the approach can be generalized using the fact that symmetric solutions to the NMF problem are equivalent to k-means classification. The system that we introduce thus makes use of the output of the NMF as a mid-level representation, or pre-classification, of the similarity matrices to segment its structure.

We first introduce NMF and the estimation of the matrix factors in section 3.2. The particular cases of symmetric matrices and symmetric solutions are also presented. We then discuss in section 3.3 the interpretation of the NMF decomposition of similarity matrices for structure segmentation, especially in the cases of state and sequence representation. Based on these results we introduce a system for music structure segmentation in section 3.4 and propose a preliminary evaluation in section 3.5.

3.2 Non-negative Matrix Factorization

Non-negative Matrix Factorization (NMF) was introduced in [Lee 1999] by Lee and Seung and aim at the factorization of matrices in the form of equation 3.1.

$$V \approx WH \tag{3.1}$$

Given the $n \times m$ non-negative matrix V, NMF thus aims at estimating the non-negative factors W $(n \times r)$ and H $(r \times m)$, that best approximate the original matrix. Standard factorization such as Singular Value Decomposition (SVD) allow negative entries in the matrix factors even if the original matrix is nonnegative. Interpretation of the original data with the matrix factors is thus highly holistic and does not reflect the constitutive parts or objects in the original data. NMF estimates matrix factors that are constrained to non-negativity. When reconstructing the original data, additive combinations of matrix factors are thus not allowed, leading to a parts-based representation. For this particular reason, NMF has met a large success in Multimedia Information Retrieval tasks such as face recognition, text mining, or sound source separation.

3.2.1 Algorithm

NMF matrix factors are first initialized. Most of the time this initialization is random. Estimation of the matrix factors is then an iterative process in which a cost function between the original data and the matrix factors is minimized by means of multiplicative update rules.

Lee and Seung formulate in [Lee 2000] two cost functions for measuring the quality of the estimated factors and propose two algorithms that aim at minimizing these two cost functions. The first measure is the square of the Euclidean distance between two nonnegative matrices A and B as in equation 3.2. In the case of NMF we refer to equation 3.1 and have $A = V$ and $B = WH$.

$$||A - B||^2 = \sum_{ij} (A_{ij} - B_{ij})^2 \tag{3.2}$$

$||A - B||^2$ equals zero if and only $A = B$.

Secondly, they propose to minimize the measure in equation 3.3.

$$D(A||B) = \sum_{ij} (A_{ij} \log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij}) \tag{3.3}$$

$D(A||B)$ is also lower bounded by zero if and only if $A = B$. The measure is however non symmetric and is thus referred to as a a divergence instead of a distance.

To minimize these two measures, taken as cost functions between V and WH, Lee and Seung propose two update rules for the matrix factors as in Theorem 1 and Theorem 2.

Theorem 1: The euclidean distance $||V - WH||$ is non increasing under the update rules:

$$H_{a\mu} \leftarrow H_{a\mu} \frac{(W^T V)_{a\mu}}{(W^T W H)_{a\mu}} \qquad W_{ia} \leftarrow W_{ia} \frac{(V H^T)_{i\alpha}}{(W H H^T)_{ia}} \tag{3.4}$$

The Euclidean distance is invariant under these updates if and only if W and H are at a stationary point of the distance.

Theorem 2: The divergence $D(V||WH)$ is non increasing under the update rules:

$$H_{a\mu} \leftarrow H_{a\mu} \frac{\sum_i W_{ia} V_{i\mu}/(WH)_{i\mu}}{\sum_k W_{ka}} \qquad W_{ia} \leftarrow W_{ia} \frac{\sum_i H_{a\mu} V_{i\mu}/(WH)_{i\mu}}{\sum_\nu H_{k\nu}} \tag{3.5}$$

The divergence is invariant under these updates if and only if W and H are at a stationary point of the divergence.

Proofs of theorems 1 and 2 can be found in [Lee 2000]. Algorithm for the estimation of NMF matrix factors thus consists in iteratively applying one of these update rules until a minimum error ε is reached.

3.2.2 NMF and Symmetric Matrices

Similarity matrices are usually computed by means of a symmetric positive distance and are therefore symmetric positive semidefinite. One can thus wonder if the symmetry of the original matrix has an impact on its NMF solution. Especially, we may wonder whether the solution is also symmetric and verifies $W = AH^T$ with A a $k \times k$ diagonal matrix. A rough observation of the reconstructed matrices obtained by the NMF decomposition of similarity matrices indeed suggests such a result. This question was studied in [Catral 2004] and authors however showed that the NMF solution as proposed by Lee and Seung for the decomposition of symmetric matrices is not symmetric.

Nevertheless, Ding et al. further studied the case of symmetric NMF in [Ding 2005] and highlighted the equivalence of symmetric NMF solutions with kernel k-means clustering. They thus propose to constraint the NMF solution for the decomposition of $n \times n$ matrix S to be in the form of:

$$S \approx HH^T \tag{3.6}$$

with H a $n \times k$ matrix. Such a solution is obtained by applying the following update rule:

$$H \leftarrow max\left(SH(H^TH)^{-1}, 0\right) \tag{3.7}$$

Applying this version of the NMF algorithm to similarity matrices, we thus perform a previous clustering of the structural information that we hope will benefit from the ability of NMF to yield parts-based representations of data.

3.3 NMF and Similarity Matrices

We are now interested in understanding wether the NMF decomposition of similarity matrices can help for music structure segmentation. Prior to applying the decomposition to real case similarity matrices, we study how NMF handles ideal state and sequence representations. Secondly, we compare the decompositions of a real case similarity matrix obtained by means of NMF and of another technique that is the Single Value Decomposition (SVD).

3.3.1 State vs. Sequence Representation

As seen in Chapter 2, there are two distinct visualizations of structure in similarity matrices: the state and the sequence representations. The work in [Cooper 2002] suggests that the NMF decomposition of similarity matrices can help for the interpretation of states. In this section, we will aim at understanding how the NMF handles both representations.

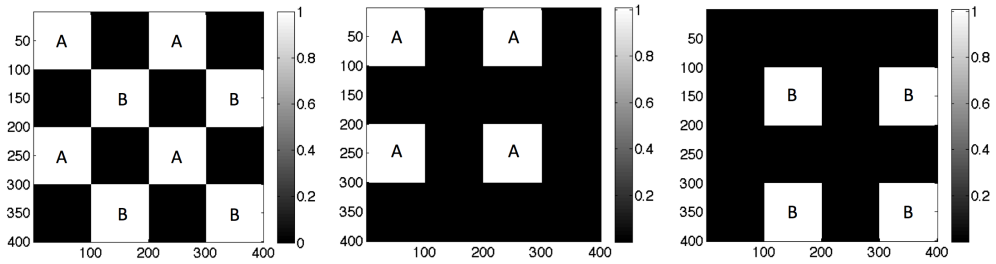
State Representation: an ideal state-shaped similarity matrix displays structural sections as uniform self-similar blocks that may be repeated over time as illustrated in Figure 3.1.a. Considering such a n×n similarity matrix S that contains two ideal-shaped states 'A' and 'B', we perform its 2^{nd} order NMF decomposition $S = WH$ and define the n×n matrix reconstructions of each dimension of the decomposition D_k as in equation 3.8.

$$D_k(i,j) = W(i,k)H(k,j) \tag{3.8}$$

Reconstruction D_1 and D_2 of both dimensions of the decomposition of S are shown in Figure 3.1.

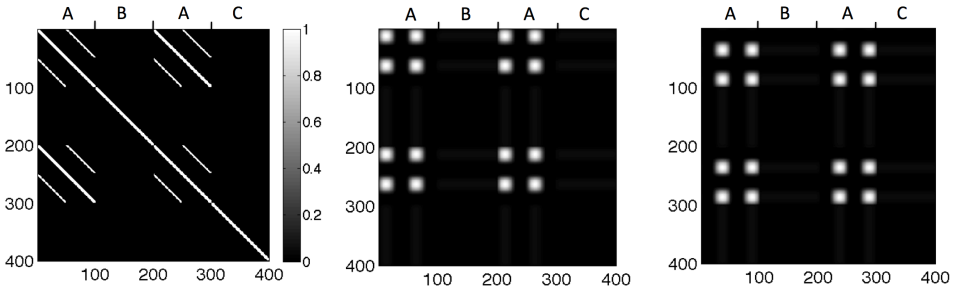
There is absolutely no ambiguity in the decomposition provided by NMF: each dimension separates the contribution of a state in the original similarity matrix and the matrix can be exactly reconstructed with a rank of decomposition r = 2.

Sequence Representation: we perform the same experiment for an ideal sequence representation. S now contains three sections 'A', 'B' and 'C' that fit the state hypothesis and are thus represented as stripes (Figure 3.2.a). Structure consists of only one repetition with the sequence 'A-B-A-C'. Matrix reconstructions D_1 and D_2 of both dimensions of the decomposition of S are shown in Figure 3.2.



(a) Ideal sate similarity matrix S (b) 1^{st}st Decomposition Reconstruction D_1 (c) 2^{nd} decomposition reconstruction D_2

Figure 3.1: Ideal state representation similarity matrix and its NMF decomposition of rank 2



(a) Ideal sequence similarity matrix S (b) 1^{st} decomposition reconstruction D_1 (c) 2^{nd} decomposition reconstruction D_2

Figure 3.2: Ideal sequence representation similarity matrix and its NMF decomposition of rank 2

Obviously, low-rank NMF decomposition does not allow for the reconstruction of the diagonal stripes that compose the structure of the similarity matrix. Motifs displayed in both reconstructions correspond to repetitions on the horizontal and vertical lines of the matrix but not on the off-diagonals. Moreover, each decomposition is a rough approximation of the original data and only show regions in which a repetition has occurred. We now set the rank of decomposition to r = 40. The reconstructed matrix WH is shown in Figure 3.3.

Increasing r, one obtains a satisfying but not ideal reconstruction of the original matrix. However, such a high rank of decomposition prevents from finding a particular mapping between the structural sections of a music piece and the dimensions of the NMF decomposition in the case of a sequence representation.

To conclude, we can say that the similarity matrix decomposition by means of NMF generally tends to separate patterns that are simultaneously repeated on the vertical and horizontal lines of the matrix and is thus adapted for the segmentation of states. Moreover, a single sectional block, as a subset of consecutive repeated frames, does not need to be repeated to be detected by the NMF. In contrast repeti-

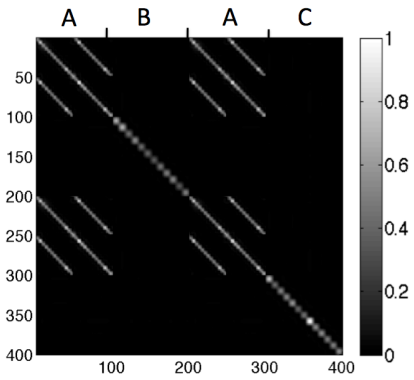


Figure 3.3: Reconstructed similarity matrix by means of a NMF decomposition with r = 40

tions on the off-diagonals as in the sequence representation seems hardly analyzable by means of NMF.

3.3.2 NMF vs. SVD

We now compare the decompositions of similarity matrices obtained by means of NMF and Single Value Decomposition (SVD). This time we use a real case similarity matrix computed on the spectral moments, i.e. spectral centroid, spread and skewness, extracted on the song *Creep* by Radiohead (Figure 3.4.a). The song contains sections played with either acoustic or saturated guitar that have very different and homogeneous timbral properties and thus form two distinct states that we labeled as 'A' and 'B'.

In the case of a real symmetric $n \times n$ matrix S, SVD aims at its factorization in the form of equation 3.9.

$$S \approx U\Sigma U^T \tag{3.9}$$

U is an orthogonal $n \times n$ real matrix containing the left and right eigenvectors, and Σ is a $n \times n$ non-negative diagonal matrix containing the eigenvalues of S on its diagonal. We recall that an eigenvector u and associated eigenvalue σ are such that equation 3.10 is verified.

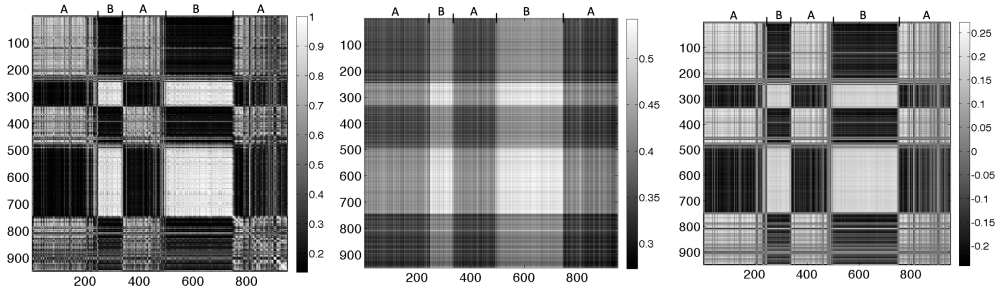
$$S.u = \sigma.u \tag{3.10}$$

Vectors u of U can constitute an orthogonal basis in which S can be generated. Each dimension i of the basis generates a matrix A_i as defined in equation 3.11.

$$A_i = u_i \sigma_i u_i^T \tag{3.11}$$

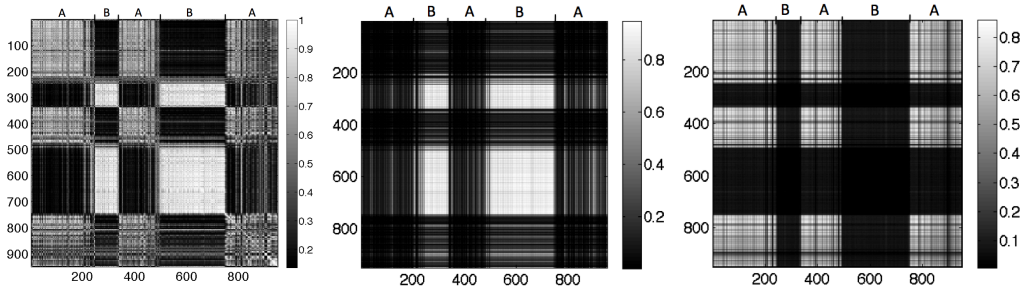
In the case of low-rank approximation, we retain only the r largest singular values in Σ. The remaining singular values are set to zero. For the decomposition of the similarity matrix of the song Creep that contains two states, we set r to 2.

The 2^{nd} order SVD decomposition and reconstructions are shown in Figure 3.4 while the 2^{nd} order NMF decomposition and reconstructions are shown in Figure 3.5.



(a) Similarity Matrix (b) 1^{st} SVD reconstruction (c) 2^{nd} SVD reconstruction

Figure 3.4: Timbral similarity matrix for the song *Creep* and its SVD low-rank approximation of rank 2



(a) Similarity Matrix (b) 1^{st} NMF reconstruction (c) 2^{st} NMF reconstruction

Figure 3.5: Timbral similarity matrix for the song *Creep* and its NMF decomposition of rank 2

We can easily say that the two dimensions of the NMF decomposition almost perfectly separate the the two states 'A' and 'B' of the song. In contrast, while the first dimension of the SVD shows a pattern that is quite coherent with the 'B' state, the second decomposition is completely holistic and does not help for any interpretation of the structure of the song. This comparison further illustrates the fact that not allowing for negative entries in the decomposition does help for a semantic interpretation of data in general, and for the segmentation of structural states in particular.

3.4 NSMF Structure Segmentation

As seen in the previous section, NMF seems to have a strong potential for segmenting states in similarity matrices. To exploit this in the context of music structure segmentation, we introduce a system that uses the NMF decomposition of similarity matrices as a mid-level representation to classify the structure. We call this system NSMF for Non-negative Similarity Matrix Factorization. In this section, we first introduce the NMF-derived mid-level representation and then present a detailed description of the structure segmentation algorithm.

3.4.1 NSMF Data Representation

Considering the NMF factorization of a similarity matrix S, each element s_{ij} of S can be written as in equation 3.12.

$$s_{ij} \approx \sum_{k=1}^{r} W_{ik} H_{kj} \tag{3.12}$$

To allow for a better separation of states, we however constrain the NMF solution to be symmetric and have $H = W^T$. Unlike most NMF applications in audio signal processing where W contains a dictionary of basis patterns and H their activations in time, the decomposition information in our case is thus contained in the matrix W solely. Information in H is indeed redundant with W and we do not consider it in our data representation.

Each music piece can thus now be represented in a r-dimensional space defined by W. To illustrate how the structural information is visualized in such a space, we extract timbral features on the song *Help* by the Beatles and compute the similarity matrix. The song consists of an introduction, a verse that is repeated three times, a verse that is also repeated three times, and an outro. The similarity matrix is plotted with its annotated structure in Figure 3.6. It shows that states are reasonably well represented for this song. We then estimate the low-rank approximation of rank 3 by means of the symmetry-constrained NMF. Each section is represented in the space defined by the three vectors of the matrix W in Figure 3.7.

Though the two main sections, i.e. verse and chorus, are not each exclusively represented over a single basis vector of the factorization, as it is the case with ideal state representations, the obtained representation space seems to allow for the discrimination of structure. Indeed, feature points of each of the two main sections spreads over a particular area of the space in which repetitions follow similar trajectories. In this sense, observation of sections by means of this mid-level representation defines clusters of points whose properties should allow for an efficient classification of structure.

We can conclude that though the NMF decomposition does not provide a one-to-one mapping between its dimensions and the musical structure, it can help for

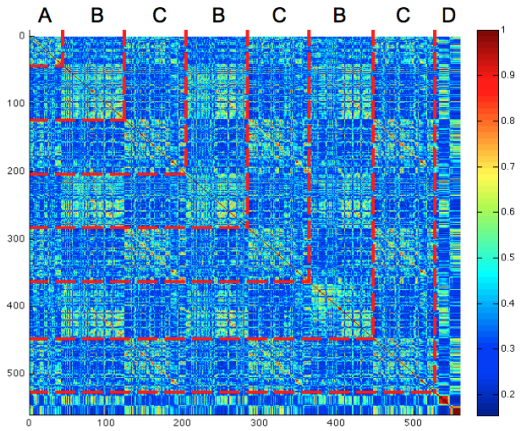


Figure 3.6: Similarity matrix computed on the timbral features for the song *Help* by the Beatles

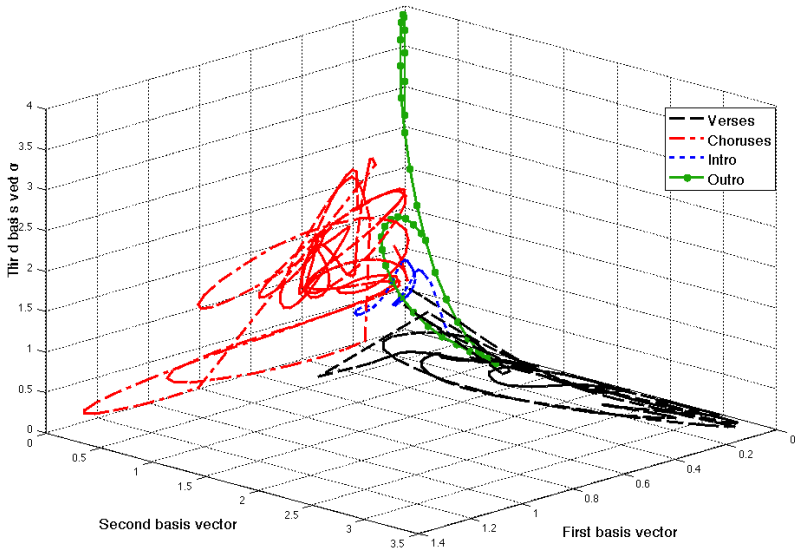


Figure 3.7: Projection of structural parts on the NSMF basis vectors

the discrimination of structural sections by defining a rather small dimensional mid-level feature representation of structure. We thus now introduce an algorithm that benefits from this representation to estimate musical structures. The constitutive blocks of the system are shown in Figure 3.8. Computation of the similarity matrices and their temporal segmentation are first detailed in sections 3.4.2 and 3.4.3. We then discuss the rank of decomposition that should be chosen for the NMF in section 3.4.4. The clustering approach is finally presented in section 3.4.5.

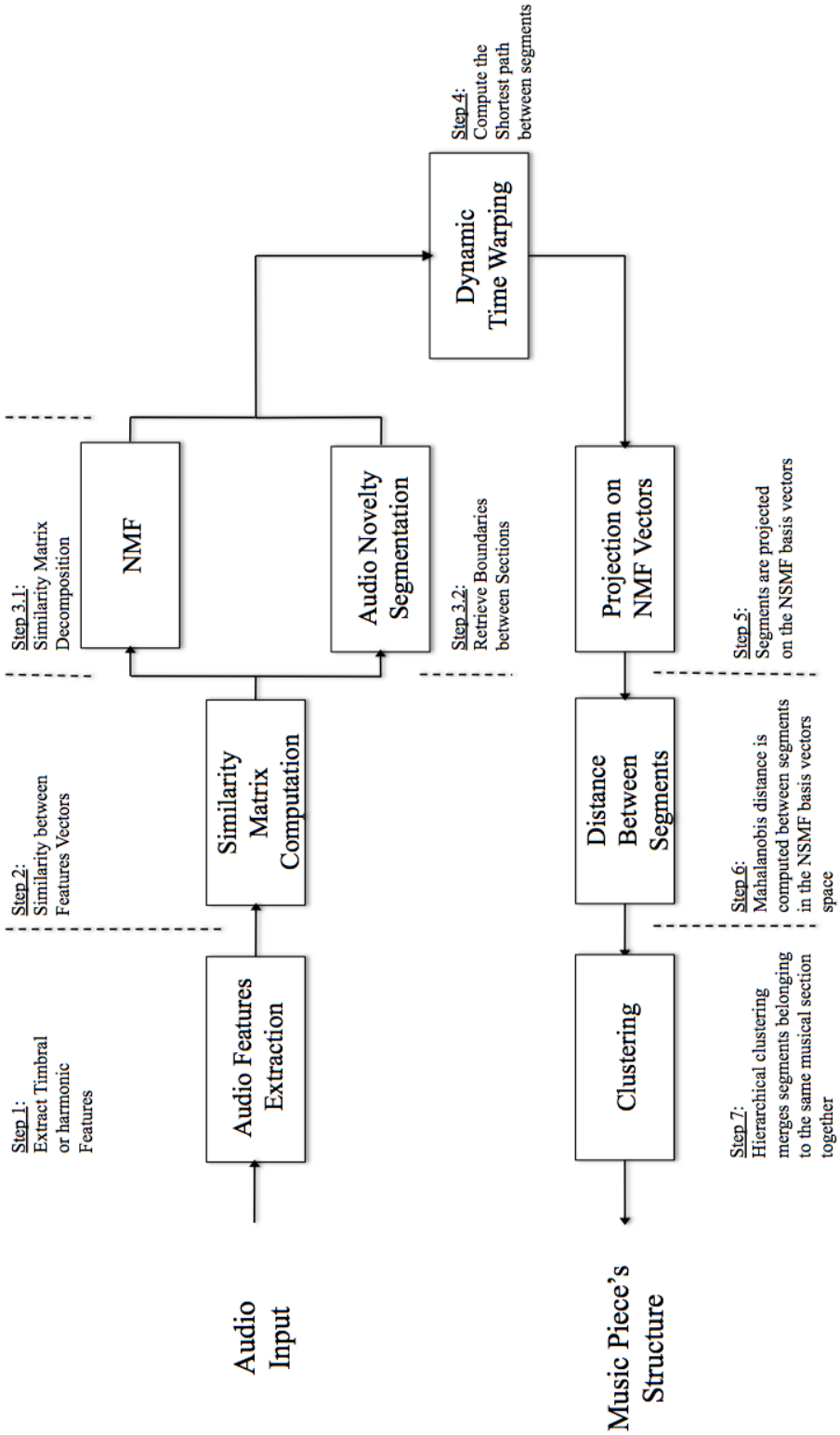


Figure 3.8: NSMF Structure Detection Overview

3.4.2 Features Extraction and Similarity Matrices

Generally, structure is more likely to appear in the similarity matrix as a state representation if one uses low-level description of timbre rather than harmony. However, chroma representations quite often yield small blocks of similarity that are not of the scale of structural sections but still constitute mid-states. Since we don't expect a one-to-one mapping between the NMF decomposition of similarity matrices and the structure of the songs, it would be interesting to segment such intermediate states for a mid-level representation.

We thus extract for each song two similarity matrices computed by means of chroma features and timbral features. For the audio features extraction to be representative of the scale of musical events such as notes, the signal is framed by overlapping windows of 250 ms, yielding feature vectors sampled at 4 Hz. The timbral content of music pieces is described by means of the following spectral low-level features:

1. MFCC
2. Spectral Centroid
3. Spectral Spread
4. Spectral Skewness
5. Spectrum Flatness

Feature vectors are normalized to mean zero and variance one and stored in a same feature matrix that serves for the similarity matrix computation. Similarity matrices are then computed by means of the exponential variant of the cosine distance (see eq. 2.15).

3.4.3 Applying the Audio Novelty Function

Prior to clustering the structure in the NSMF representation, boundaries are retrieved in the similarity matrix by means of the audio novelty score as described in section 2.3.2. Our goal is to detect boundaries between large scale audio segments and the length of the gaussian checkerboard should thus be carefully chosen. The description of databases in section 6.1 relates that the average length of sections in our popular music database is of about 20 seconds. Therefore, to allow for the detection of such segments as well as of shorter segments that might occur such as introductions, we use a checkerboard of length 15 seconds. This means that each point of the novelty curve measures the dissimilarity between the past- and forth-coming 7,5 seconds. The audio features' sample rate being 4 Hz, the novelty function is thus a 60×60 gaussian checkerboard, with variance sigma equal to 0,5.

Once the novelty function is computed, a threshold is applied to detect peaks that indicate boundaries in the audio signal. Dissimilarity range in the novelty curves however varies between music pieces and we apply the adaptive threshold estimation described in [Jacobson 2001]. Therefore, a peak P in the audio novelty

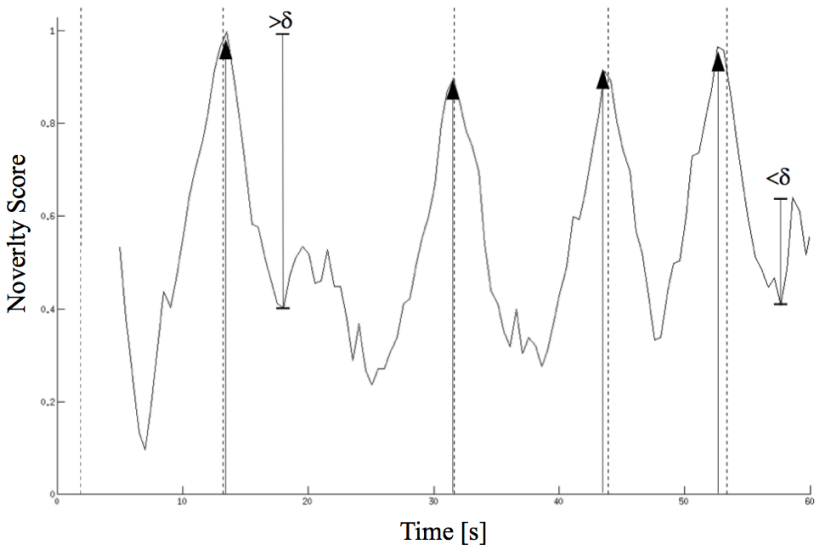


Figure 3.9: Boundaries detection in the audio novelty score. Upwards arrows indicate detected boundaries.

function is defined as the local maxima between two consecutive local minima, and a trough T as a local minima between two local maxima. A boundary is then defined as a peak whose value exceeds its two neighboring minima of at least a threshold δ as described in equation 3.13 and illustrated in Figure 3.9.

$$P_i \quad \equiv \quad T_i + \delta \leq P_i \cap T_{i+1} + \delta \leq P_i \tag{3.13}$$

To estimate the threshold δ, the novelty curve is clustered by means of the nearest neighbor technique into two clusters : the highest and lowest values. Instead of choosing the mean value of the novelty curve, the threshold δ is then chosen as the mean value of the largest cluster.

An example segmentation obtained for the similarity matrix computed on the timbre features of the song *Help* is shown in Figure 3.10. An evaluation of the performance of this technique on our datasets is proposed in Chapter 6.

3.4.4 Rank of Decomposition

A one-to-one mapping between the dimensions of the NMF and the structural sections is not to be expected. Indeed, there are different hierarchy levels in musical structures, and a structural section as perceived by humans is often not represented by a single homogeneous block in the similarity matrix. The dimension of the feature space obtained with the NSMF representation that is determined by the rank of decomposition is thus a crucial parameter to allow for the description of sections. Estimating the optimal rank for each song is rather time-consuming [Chen 2011] and we prefer to run an experiment to estimate a good compromise in the choice of

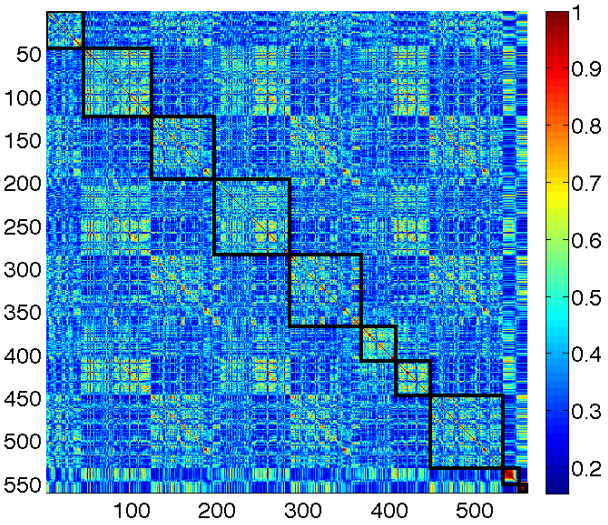


Figure 3.10: Segmentation of the song *Help* by means of the audio novelty score

the rank for popular music. We therefore select a subset of ten songs of the *TUT Beatles* database (see Chapter 3.10) and extract their timbral features similarity matrices. Varying the rank of NMF r from 3 to 12, we measure the separability between structural parts along each dimension w_i of W. To do so, we compute the inertia ratio of the variance of w_i within segments belonging to the same structural part with the variance of w_i over the whole music piece [Peeters 2002a]:

$$s(i) = \frac{\sum_{k=1}^{K} \frac{N}{N_k}(m_k - m_i)(m_k - m_i)'}{\frac{1}{N}\sum_{n=1}^{N}(w_i(n) - m_i)(w_i(n) - m_i)'} \tag{3.14}$$

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 w_i over the all piece and m_k the mean value of w_i over the k^{th} structural part. For a given rank of decomposition r, the separability is then measured as the mean of s as in equation 3.15.

$$sep(r) = \frac{1}{r}\sum_{i=1}^{r} s(i) \tag{3.15}$$

We find a maximum of separability with a rank of 9 for NMF (see Figure 3.11) that is as expected significantly larger than the median number of annotated structural sections, i.e. 5,12 (Chapter 3.10).

3.4.5 Clustering Approach

The task of music structure segmentation now consists in merging together the objects formed by the structural components of songs in the NSMF feature

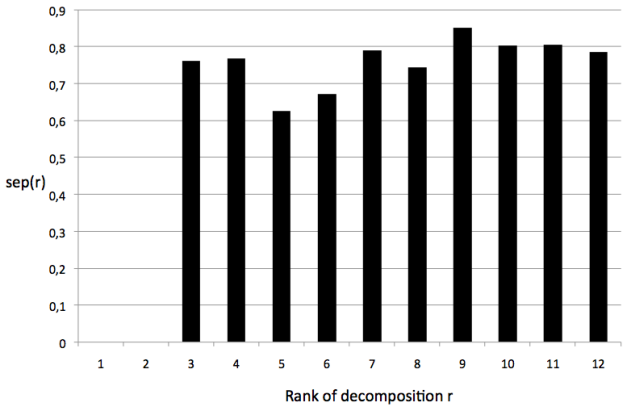


Figure 3.11: Separability of structural parts given different ranks of decomposition

space. We have few assumptions on these objects apart from the fact that they should be distinct from each other in the feature space and that the occurrences of a same section should produce similar forms. Moreover, structure representation is completely dependent on the song and a supervised approach for the structure classification seems inefficient. We have however a very useful information with the segmentation step that indicates the borders of these structural objects in the NSMF space, allowing for their comparison. This information can be efficiently exploited for estimating the song's structure using hierarchical clustering approaches that allow to refine a given segmentation either by starting with single clusters and joining similar segments, as in agglomerative algorithms, or by first considering a unique cluster and splitting intrinsically dissimilar clusters, as in divisive algorithms.

Using an appropriate distance measure within the NSMF feature space, a cluster can then be defined by a maximum distance that merges segments together. This induces a hierarchy in the clusters that can be represented in a dendogram in which the y axis indicates the maximum distance needed to link segments in a cluster. Segments that form the clusters are indicated on the x axis. An example dendogram is shown in Figure 3.12 for the song *Help*.

The linkage of segments in a cluster is thus centered on the notion of cluster dissimilarity. Assuming a distance measure d between segments within the feature space, the dissimilarity $D(C_i, C_j)$ between two clusters C_i and C_j can be defined in the following manner:

- The ***single*** linkage D_S defined as the shortest distance between all pairs of segments x_a and x_b forming the clusters:

$$D_S(C_i, C_j) = \min_{a \in C_i} \min_{b \in C_i} (d(x_a, x_b))$$

- The ***complete*** linkage D_C defined as the largest distance between all pairs of

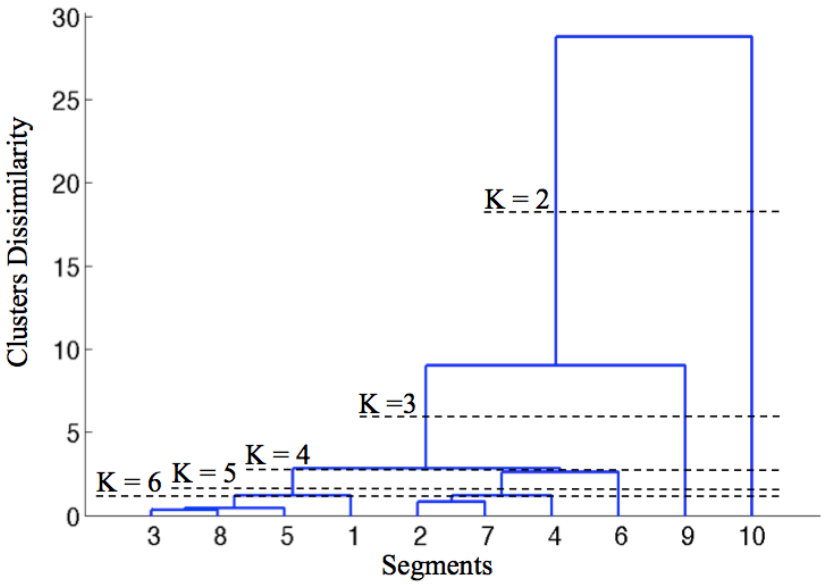


Figure 3.12: Dendogram with the *complete* linkage obtained for the song *Help*

segments x_a and x_b forming the clusters:

$$D_C(C_i, C_j) = \max_{a \in C_i} \max_{b \in C_i} (d(x_a, x_b))$$

- The ***average*** linkage D_{Avg} defined as the mean distance between all pairs of segments x_a and x_b forming the clusters:

$$D_{Avg} C_i, C_j) = \frac{1}{|C_i|.|C_j|} \sum_{a \in C_i} \sum_{b \in C_i} d(x_a, x_b)$$

The refined segmentation is derived from the hierarchical cluster tree by cutting it at a given number of clusters K. Examples of structural segmentations of the song *Help* obtained for different numbers of K is shown in Figure 3.13. The segmentation obtained for $K = 4$ and its comparison with the annotation is shown in Figure 3.14.

3.4.6 Distance Between Segments

In order to efficiently compare segments in the NSMF feature space, we need an appropriate distance that measures the statistical correlation between features' sequences of structural segments.

We first use for that purpose the Mahalanobis distance that is defined in equation 3.16.

$$d_{Mahal}(x_1, x_2) = \sqrt{(x_1 - x_2)^T \Sigma^{-1} (x_1 - x_2)} \tag{3.16}$$

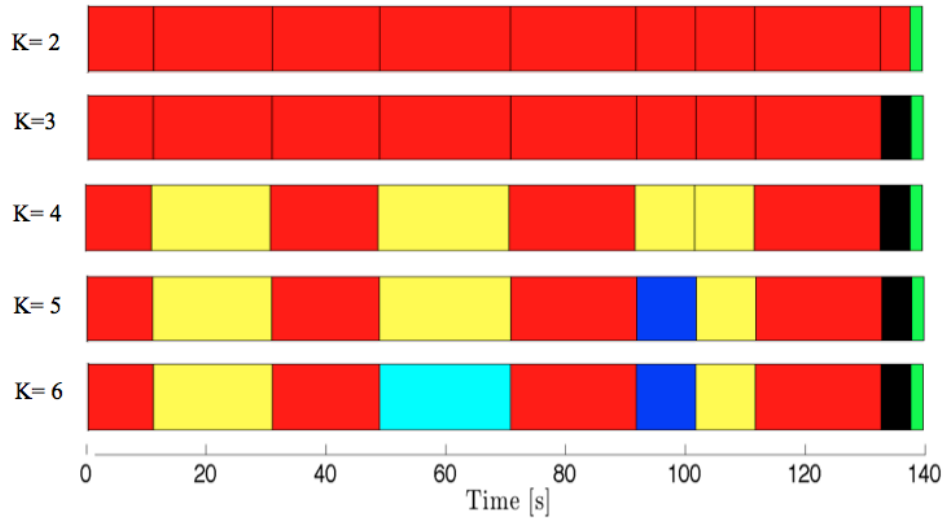


Figure 3.13: Hierarchical clustering of the song *Help* for different number of clusters K

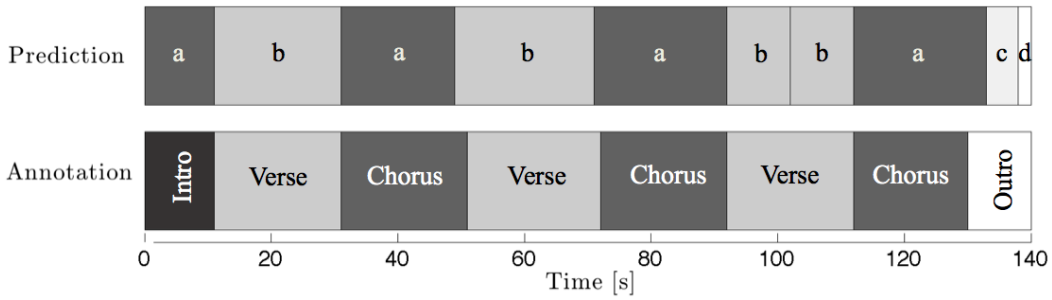


Figure 3.14: Structure clustering of the song *Help* by means of the NSMF Approach

With x_1 and x_2 two feature vectors and Σ their covariance matrix. A low value of the mahalanobis distance indicates that the two feature vectors can be seen as two observations of the same distribution.

We also measure the dissimilarity between segments by means of the Bayesian Information Criterion (BIC) that is defined in equation 3.17

$$BIC = (n_1 + n_2)\log(\Sigma) - n_1\log(\Sigma_1) - n_2\log(\Sigma_2) - \lambda P \tag{3.17}$$

With n_1 and n_2 the lengths of the compared segments, Σ_1 and Σ_2 the covariances of their corresponding feature vectors and Σ the covariance of the whole feature sequence. P is a penalty factor and λ a scalar comprised between 0 and 1. Note that the BIC is not symmetric and is thus not a distance. It is however a relevant indicator of the segments' dissimilarity in terms of their statistical properties and shows whether or not two feature vectors can be modeled by a same single-gauss distribution. In the literature, the BIC is often used as a criterion to

merge segments together. In speaker diarization for example, if the BIC between two speaker segments is under a given threshold, the segments belong to a same speaker and should be merged together.

In order to account for the tempo variations and eventual over-segmentation of sections, Dynamic Time Warping is also applied to find the shortest path between the two compared segments in the similarity matrix. Only the corresponding feature points in the NSMF space are then considered for the distance computation.

3.5 Preliminary Evaluation

3.5.1 Introduction

The system is preliminarily evaluated on the *TUT Beatles* database. The database consists of 145 songs of the Beatles. Performance is measured with the pairwise F-Measure, Precision and Recall described in Section 2.5 of Chapter 2. A further evaluation is provided in Chapter 6.

The system is evaluated extracting both timbral and chroma features. Results obtained for these similarity matrices are labelled as *Timbre* and *Chroma* respectively. We also consider two fusion strategies of the timbral and chroma information. The system is therefore run on a similarity matrix for each song that is the sum of the timbre and chroma similarity matrices. This means that the NMF is applied on this particular matrix and that the dimension of the NSMF feature vectors is of 9. This method is labelled as *Fusion 1.* For the second fusion strategy, the NSMF features are obtained with the combination of the NMF of the timbral and chroma matrices separately and are thus 18 dimensional. This method is labelled as *Fusion 2.* We also compare the performances obtained using the boundaries estimated by means of the audio novelty score (labelled *auto*), and using the annotated boundaries (labelled *manual*). The parameters used for the evaluation are summarized in Table 3.1.

Audio features	Timbre and/or Chroma
Distance between features	Cosine distance
Checkerboard's length	60 samples
NMF rank	9
Dissimilarity measure within NSMF	Mahalanobis and BIC
Linkage method	complete
Number of clusters K	4

Table 3.1: Parameters for the structure segmentation algorithm used for the evaluation

3.5.2 Results and Discussion

The evaluation for all different inputs is reported in Tables 3.2 and 3.3. The two tables differ in the distance used for the segments comparison. The performances are compared with the state-of-the-art system described in [Mauch 2009] evaluated on the same dataset and that won the MIREX 09 evaluation campaign.

Method	*Segmentation*	F	P	R
[Mauch 2009]		60.0%	56.1%	71.0%
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 3.2: Evaluation on *TUT Beatles*, Mahalanobis

Method	*Segmentation*	F	P	R
[Mauch 2009]		60.0%	56.1%	71%
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 3.3: Evaluation on *TUT Beatles*, BIC

The first result is that we show an overall performance that is comparable with the state-of-the-art, achieving F-Measures until 62,1% with the Mahalanobis distance and the first fusion strategy. The NMF decomposition of similarity matrices thus proves to be a satisfying model for the structural segmentation of music. The precision and recall rates however suggest that the structural segmentation estimated with our algorithm is of a different nature. Indeed, we achieve precision rates up to 10% higher than the reference system, but compensated with a de-

crease in the recall rates. Looking at the definitions of the pairwise precision and recall rates in equations 2.22 and 2.23, we note that precision can be virtually increased by over-segmentation, while recall increases with under-segmentation. As reported in the boundary retrieval evaluation in Chapter 6.1, the segmentation step in our algorithm has a tendency to over-segmentation. The high precision rates should therefore be counterbalanced with the fact that over-segmentation might have contributed to this increase. Indeed, states annotated are of a rather high temporal scale. However, the low extraction time-scale of low-level audio features probably prevents from characterizing states of such lengths and this results in over-segmentation.

There is naturally a big gap between the performances obtained with the automated segmentation and with the annotated segments' boundaries. This raises the issue of the matching between the annotation and the structural hierarchy level that is described by the algorithm. Indeed, a structural section might also have an inner structure that is reflected in the estimated musical structure but not in the annotation. The actual performance evaluation do not qualify eventual under- or over- segmentations.

Finally, fusing both timbral and chroma description as in the "Fusion 1" strategy seems to make sense for this database and improves the overall performance of the system. For further experiments only the Mahalanobis distance will be used as there seem to be no particular interest in using the BIC criterion.

3.6 Chapter Summary and Outlook

In this chapter, we illustrated how the NMF decomposition of similarity matrices can efficiently model structural sections in the case of a state representation. Using this observation, we introduced a method that uses the basis vectors of the NSMF as mid-level features for the estimation of the musical structure. The reported evaluation shows promising results.

However, quality of the modeling of structural sections by means of NSMF relies on the display of structure in the similarity matrices as states. The method thus depends on the ability of the chosen low-level audio descriptors to fulfill the following assumptions:

- Low statistical variance within musical sections and their repetitions
- High statistical variance between different musical sections

Considering the musical landscape, these conditions are rarely fulfilled. Indeed, acoustic properties easily vary within the melodic progression of a musical section, whereas different sections often share harmonic and instrumentation properties.

Thus, considering the short extraction time-scale of the audio features, i.e. 250 ms, the nature of the extracted information does not allow to describe the musical forms or patterns that eventually reflect a kind of invariance within a section. And this results in not well-defined state representations in the similarity matrices, which are then hardly understandable for human as well as for the algorithm.

The remainder of this thesis thus aims at enhancing the state representation in similarity matrices in order to strengthen their separation in the NSMF feature space. Two approaches are followed. First we apply in Chapter 4 image processing techniques to generate from similarity matrices a binary segmentation mask that provides a simpler structure visualization that tends to strengthen the state representation. The original similarity matrices are then enhanced by means of the mask.

In the second approach, developed in Chapter 5, we intend to model local harmonic properties by means of a mid-level audio feature. The idea is to probe local pitch class intervals within chroma sequences to characterize tonal contexts.

Chapter 4

Visualization Enhancement in an Image Processing Framework

Contents

Similarity matrices provide visualizations of the audio signal's content. While state or sequence representation are not always well defined in these visualizations, structural sections however often draw clouds of self-similar frames that are perceptually visible. In this chapter, we thus consider structural sections as visual objects in an image representation that is the similarity matrix, and consider the problem of their segmentation in an image processing framework. Indeed, image segmentation techniques allow to ignore noise in images in order to draw their contours. Applying such techniques to similarity matrices, we are able to strengthen the state representation of structure and facilitate its segmentation by means of the NSMF algorithm introduced in Chapter 3.

4.1 Image Segmentation

4.1.1 Background

Image segmentation tackles the problem of partitioning an image into *meaningful* segments. The term *meaningful* can be interpreted in the context of human perception, where the segments should be distinguishable for humans or in the context of an application where the segments will be accordingly processed and thus improve the applications performance. The main steps involved in an image segmentation problem according to [Salembier 1999] are:

a) Simplification
b) Feature Selection
c) Decision

Noise filtering serves usually as simplification step by removing components that distort the image such as high frequencies noise. Additionally, for maintaining the image's major features, such as edges or corners, Perona and Malik made a significant contribution in the area of noise filtering by proposing a nonlinear diffusion algorithm [Perona 1990].

Regarding feature selection, in the general case of image segmentation the feature space range is wide and may include color, motion, depth, texture etc. In the case of intensity images, it is limited to features such as pixel intensity, texture, histogram and shape. Multiple features' fusion towards object segmentation is overviewed in an extensive study of segmentation algorithms [Alatan 1998].

Finally the decision based on the feature space heavily influences the segmentation result. In the context of decision strategies we distinguish mainly:

a) homogeneity-based approaches [Shi 2000], [Salembier 1999], [Deng 2001]
b) discontinuity-based ones [Pollak 2000], [Ma 1997]
c) hybrid ones [Fan 2001]

Homogeneity-based (region-based) methods exploit the spatial homogeneity of features with respect to the feature space and attempt to produce coherent regions by region growing, or merging together and splitting subgroups of pixels. On the other hand, discontinuity-based algorithms rely on estimating the discontinuities in an image, usually using gradient and proceed by linking together edge pixels to form closed regions. Hybrid algorithms use both the homogeneity and discontinuity to perform segmentation.

4.1.2 Approach

In many image processing applications, the gray levels of pixels belonging to the foreground object are substantially different from the gray levels of the pixels belonging to the background. Thresholding then becomes an effective tool to separate objects from the background [Sezgin 2004]. Considering similarity matrices

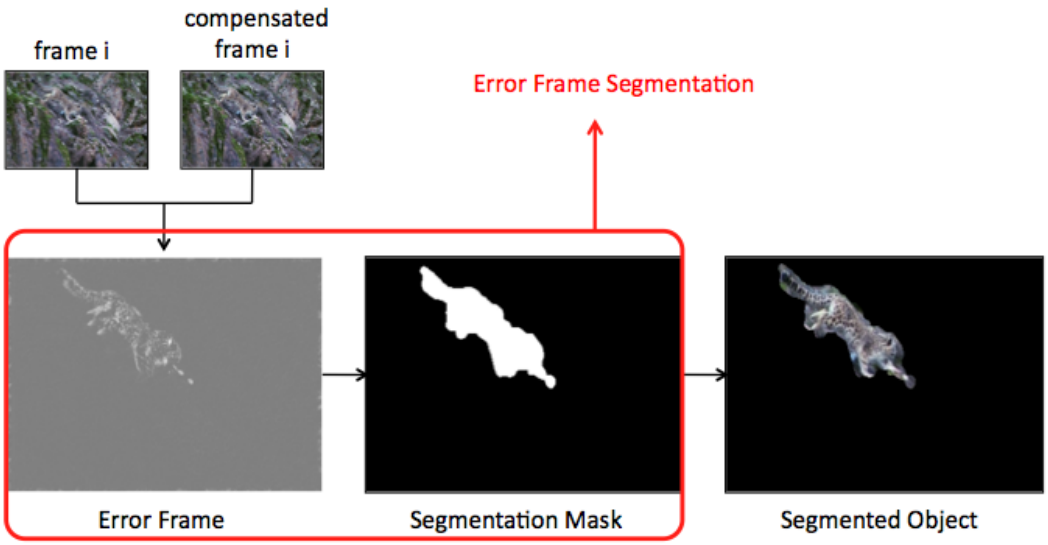


Figure 4.1: Moving objects segmentation in the video sequence "Mountain" (MPEG testset)

as intensity images, structural sections are represented by pixel groups of similar intensity levels that are coherently localized. We thus decide to apply thresholding oriented techniques instead of a classification approach by means of visual features extraction.

Among the image segmentation problems that more or less neighbor the problem of similarity matrices segmentation, we find analogies with the problem of objects segmentation in video sequences. Indeed, moving objects are segmented in error frames that are intensity images generated by means of motion compensation (see Figure 4.1). The definition of objects in error frames as groups of higher gray level intensity pixels that are not completely homogeneous fits the definition of structural objects in similarity matrices images.

We study in particular the segmentation algorithm proposed in [Krutz 2007] that consists of three steps:

a) Anisotropic diffusion filtering
b) Binarization
c) Morphological operations on the binary frames

As illustrated in Figure 4.2, low-pass filtering by means of anisotropic diffusion aims at smoothening images and enhance objects homogeneity in pixel intensity levels. Binarization then segments foreground and background objects by means of thresholding. Finally, pixels connectivity is used to refine objects applying morphological operations.

This has proved to be a quite efficient strategy for various segmentation tasks

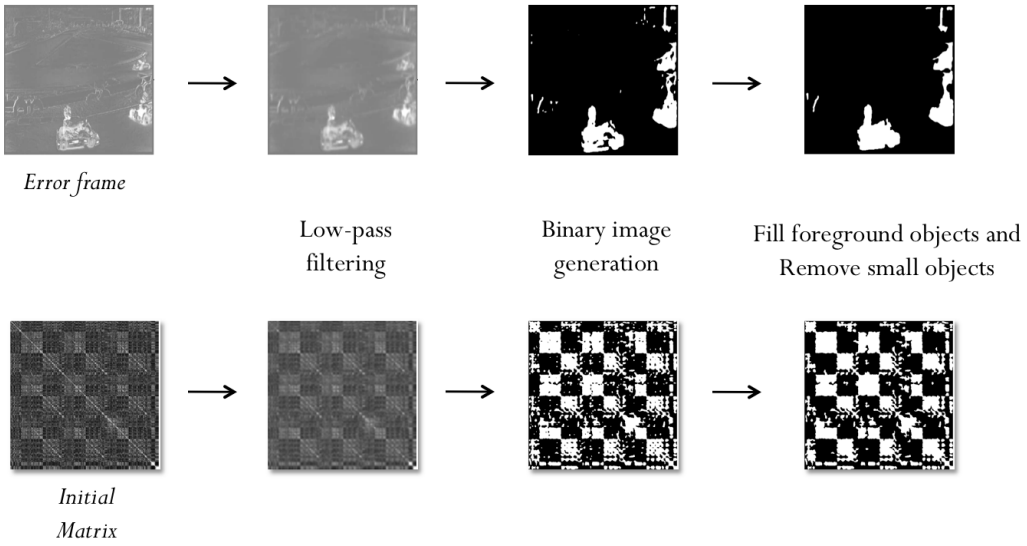


Figure 4.2: Parallel between the segmentation of error frames [Krutz 2007] and the segmentation of similarity matrices. Video sequence: "Race1 (View 0)" (544 × 336, 100 frames), part of an MPEG testset for multiview sequences

[Krutz 2009], [Kunter 2009], [Arvanitidou 2009], [Kunter 2009], and we therefore integrate parts of the algorithm in the processing of similarity matrices in order to simplify the analysis of structural objects. The algorithm is further described in section 4.2. An image enhancement procedure by means of the segmentation mask is then introduced in section 4.3. Application of the enhanced matrices for the task of music structure segmentation is then discussed in section 4.4.

4.2 Algorithm Description

Processing blocks of the segmentation algorithm in [Krutz 2007] are illustrated in Figure 4.2 on an error frame example and on an audio similarity matrix. In an image context, the final binary segmentation mask should draw the contours of foreground objects. Applying the algorithm to similarity matrices, we hope to generate segmentation masks that draw the contour of the state representation of structure.

4.2.1 Anisotropic Diffusion

In image processing, error frames are generated by means of motion compensation and image background modeling. However, error accumulation in the global motion estimation causes distortions in the error frame and high values in the

background regions remain [Krutz 2010]. As these distortions are characterized by high-frequency noise, low-pass filtering is used to minimize the errors. Problem is, object's edges are also defined by high frequency in the image, and employing a classical low-pass filter, e.g. Gaussian filter, blurs the edges and affects the segmentation performance.

To cope with this edge blurring, low-pass filtering is done by means of anisotropic diffusion [Perona 1990]. Indeed, anisotropic diffusion filtering provides smoothing of intra-region areas preferentially over inter-region areas, thereby providing an efficient way for removing unwanted noise while preserving edges.

The filtered image S_t is obtained by the diffusion equation that is formulated as follows:

$$S_t = div(c(x,y)\nabla S(x,y)) \tag{4.1}$$

with $c(x,y)$ the diffusion coefficient and $\nabla S(x,y)$ the gradient of the original intensity image $S(x,y)$. The ideal diffusion coefficient should be equal to 0 at coordinates corresponding to edges and equal to 1 for homogeneous regions. In that manner, intra-objects regions are smoothed and edges are kept sharp. However, objects edges are not previously known. The diffusion coefficient is thus chosen as a function of the image's gradient:

$$c(x,y) = g(||\nabla S(x,y)||) \tag{4.2}$$

with g(.) a non-negative monotonically decreasing function such that:

$$||\nabla S(x,y)|| \to 1 \quad \Rightarrow \quad g(||\nabla S(x,y)||) \to 0$$

Gradient of images is employed in image processing for edges detection purposes. Indeed, for homogenous pixels regions that characterize the inner part of objects, the gradient is low, whereas it increases at edges. Employing such a function of the gradient thus provides a smoothing of homogeneous regions, whereas edges are kept.

We select the following diffusion coefficient function g(.) [Perona 1990]:

$$g(|\nabla S|) = \frac{1}{1 + (|\nabla S| / K)^2} \tag{4.3}$$

where K is a constant that controls the sensitivity of the algorithm to objects edges whose value should be carefully chosen. The value of the parameter K is discussed in section 4.2.4.

Example of a low-pass filtered similarity matrix is shown in Figure 4.7c.

4.2.2 Binarization

After filtering, the similarity matrix image is binarized. In [Krutz 2007], this is done by means of a weighted mean thresholding, but we prefer to employ the *Otsu*

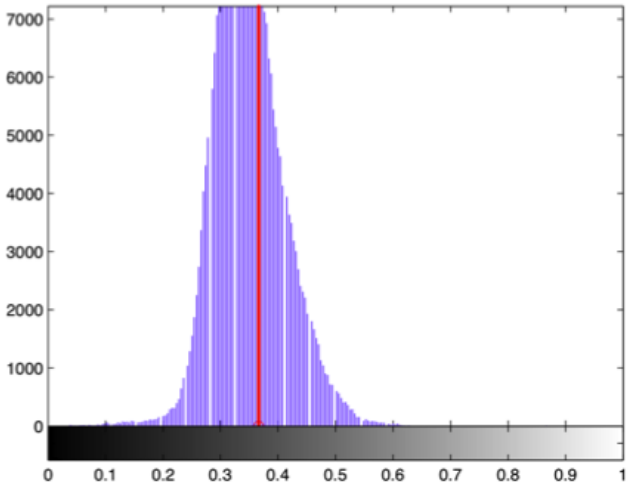


Figure 4.3: Illustration of Otsu Threshold estimation with the image histogram probability

thresholding [Otsu 1979] that performs better with images that have unimodal histograms as we do. The image is assumed to be composed of two classes of pixels: the background pixels and the foreground pixels. A threshold that minimizes the intra-class variance of both classes is then estimated (see Figure 4.3 for illustration). Given a threshold T, the global intra-class variance $\sigma^2_{Intra}(T)$ is defined as a weighted sum of intra-class variance of both classes, $\sigma^2_B(T)$ and $\sigma^2_F(T)$:

$$\sigma^2_{Intra}(T) = \omega_B(T)\sigma^2_B(T) + \omega_F(T)\sigma^2_F(T) \tag{4.4}$$

where

$$\omega_B(T) = \sum_{i=1}^{T-1} p(i) \tag{4.5}$$

and

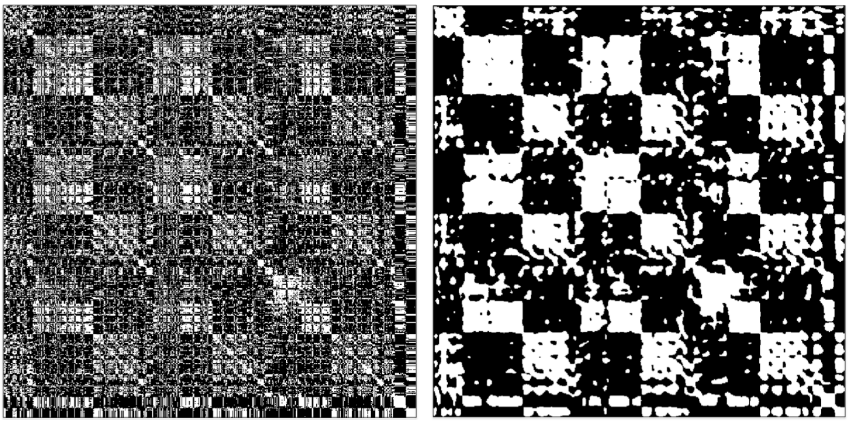
$$\omega_F(T) = \sum_{i=T}^{N-1} p(i) \tag{4.6}$$

with p the histogram probability distribution and [0 , N-1] the range of intensity levels.

Instead of minimizing the intra-class variance, Otsu proposed to maximize the between-class variance $\sigma^2_{Between}(T)$ that is calculated as:

$$\sigma^2_{Between}(T) = \sigma^2(T) - \sigma^2_{Intra}(T) \tag{4.7}$$

$$\sigma^2_{Between}(T) = \omega_B(T)[\mu_B(T) - \mu]^2 + \omega_F(T)[\mu_F(T) - \mu]^2 \tag{4.8}$$



(a) Binarization without low-pass filtering

(b) Binarization after anisotropic low-pass filtering

Figure 4.4: Comparative binary masks with and without low-pass filtering of the similarity matrix computed on the song *Help*! by the Beatles (mfcc features are used)

with μ the mean intensity level of the image, $\mu_B(T)$ and $\mu_F(T)$ the mean intensity levels of the foreground and background respectively. Replacing $\mu = \omega_B(T)\mu_B(T) + \omega_F(T)\mu_F(T)$ we obtain:

$$\sigma^2_{Between}(T) = \omega_B(T)\omega_F(T)[\mu_B(T) - \mu_F(T)]^2 \tag{4.9}$$

which is more easily computable as the intra-class variance. The optimal threshold is then computed in an iterative manner.

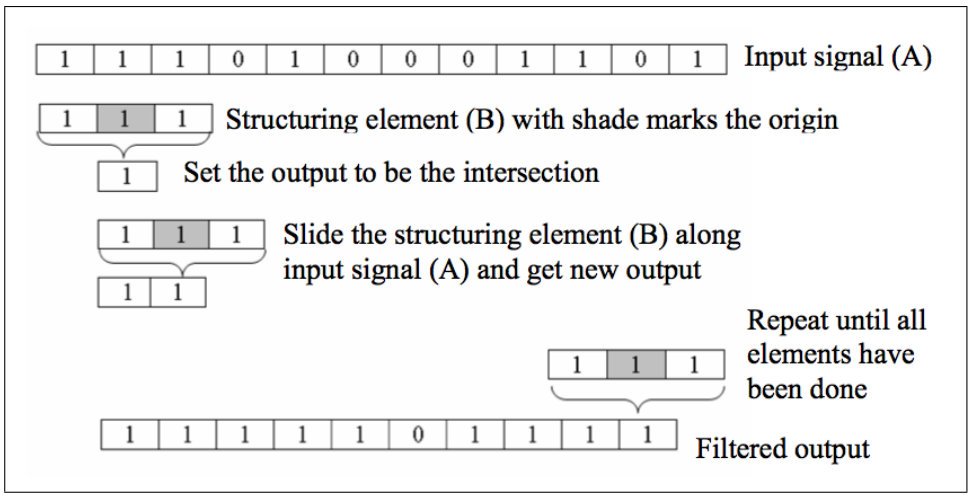
Two examples of binarized similarity matrices are shown in Figure 4.4. The matrices were computed with the song *Help* by the Beatles. The first matrix is the binary mask computed on the original similarity matrix and the second the mask computed on the low-pass filtered version. It clearly shows the benefit of previously filtering the images and smoothing the structural objects. While the mask in Figure 4.4.a did segment both foreground and background objects, blocks of high similarity were significantly preferred as foreground objects in the mask using low-pass filtering of the matrix.

4.2.3 Morphological Operators

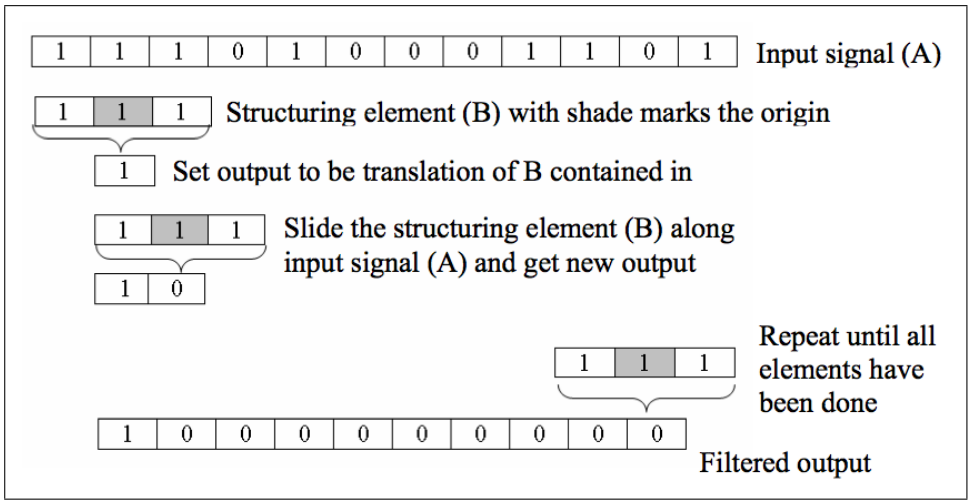
As shown in Figure 4.4.b, low-pass filtering and binarization of similarity matrices does not yield complete blocks of similarity and dissimilarity as defined in state representation of structure. Indeed, small regions of similarity remain in larger regions of dissimilarity and vice versa. To cope with this problem we apply morphological filter operations [Soille 2003]. Morphological filters analyze the signal

in terms of shape and may help to refine objects in binary images. We utilize in particular the 'Opening' and 'Closing' operations.

All morphological filters are based on the dilation and erosion operations. Both operations consist in moving a structuring element over a binary input signal A. The output of the dilation operation is set to one unless the input signal is the inverse of the structuring element. The erosion of the input signal with the structuring element is set to zero unless the input equals the structuring element. Generally image objects tend to grow in size when dilation operation is applied, whereas they crumble under the erosion operation. Illustration of both operators is shown in Figure 4.5.



(a) Example of dilation operation



(b) Example of erosion operation

Figure 4.5: Illustration of morphological dilation and erosion operations with a one dimensional binary signal [Ong 2007]

Opening and Closing then consists in the combination of the dilation and erosion operations. Closing first dilates the input signal and then erodes the output. Opening first erodes and then dilates the output. Considering an input signal A and a structuring element B, we thus have:

$$Close(A, B) = erosion(dilation(A, B), B) \tag{4.10}$$

$$Open(A, B) = dilation(erosion(A, B), B) \tag{4.11}$$

To refine structural objects in the similarity matrices, we use the Open-Close operators:

$$Open - Close(A, B) = close(open(A, B), B) \tag{4.12}$$

As illustrated in Figures 4.7.e and 4.7.f, applying the opening operation on the binary similarity matrix removes small objects in large regions of high dissimilarity. In contrast, the closing operation fills the objects of high similarity. Hence, applying morphological operators enhances the musical interpretation of similarity matrices in terms of structure.

Morphological filtering has been applied in [Ong 2007] to enhance the intelligibility of similarity matrices for segmentation purposes. Our approach however differs in the sense that we first apply low-pass filtering to enhance the precision of the binary masks. As illustrated in Figure 4.4, this operation largely contributes to simplify the structure visualization in the similarity matrix and especially strengthens the states representation.

4.2.4 Parameter K estimation

In order to estimate an optimal value for the parameter K of the diffusion equation for the segmentation of similarity matrices, we run a small experiment on the TUT Beatles dataset (see Chapter 6) for evaluating the quality of the segmentation masks obtained for different values of K. The groundtruth masks are generated by means of the annotated audio segmentation. Each structural part is thus displayed as a block in the mask. The estimated and groundtruth masks are then compared using the true positive and false positive rates (see Figure 4.6).

In order to have a good compromise between the coverage of the structural parts (true positive rate) and over-segmentation (false positive rate) we set the value of K to 40. Indeed, groundtruth masks are generated from a rough annotation of the audio data and do not account for dissimilarity regions within structural parts. They thus shouldn't be taken as the ideal masks. However, our masks should contain as less energy as possible in non-structural regions. Setting K to 40 we are still able to cover 50% of the groundtruth masks while keeping the false positive rate under 30% and limit the risk of over-segmentation.

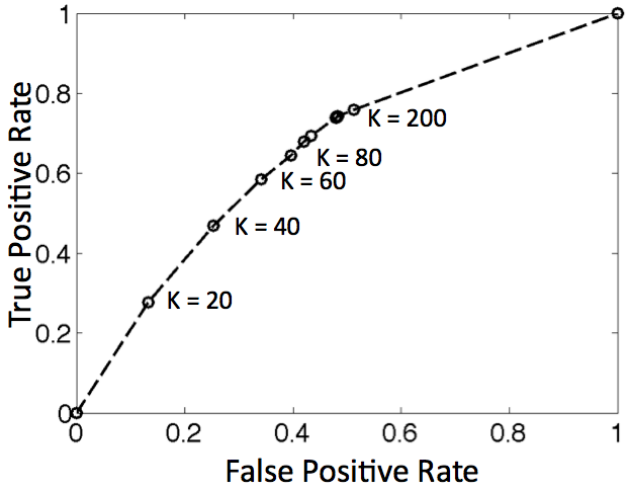


Figure 4.6: Segmentation masks true positive vs false positive rates for different values of K

4.2.5 Summary

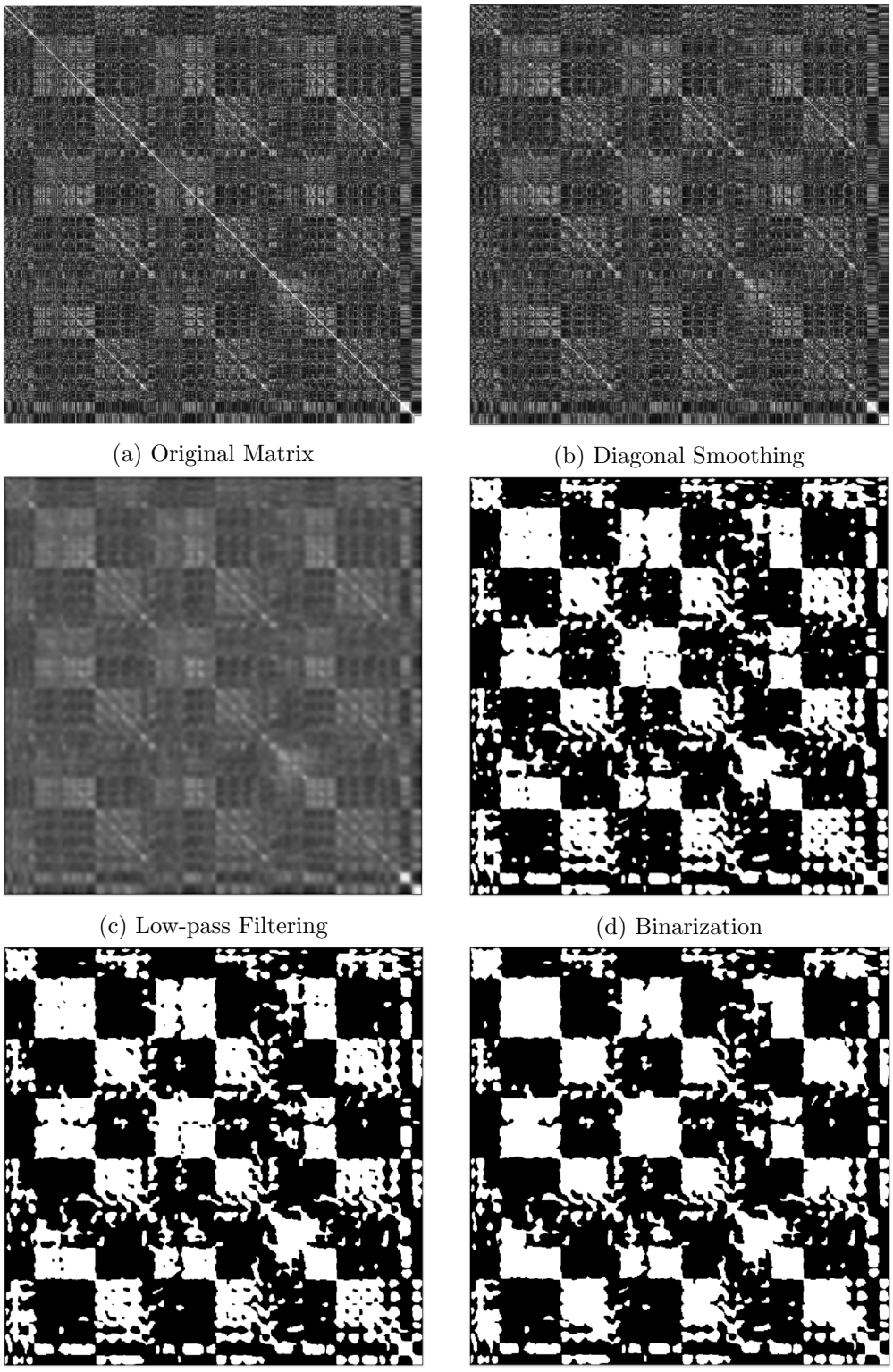
Each step of the segmentation mask generation is illustrated in Figure 4.7 with the song example *Help* by the Beatles. It shows that the approach clearly enhances a visualization of regions of similarity in the audio signal that fits a state representation of structure. The comparison of the estimated masks with the annotated structure suggests that objects segmented as foreground rather correspond to sections of the musical structure in the similarity matrices.

4.3 Enhancing Similarity Matrices

4.3.1 Enhancement Procedure

Generating a segmentation mask from an audio similarity matrix obviously helps discriminating blocks of higher similarity when they are already suggested in the original matrix. Moreover, considering the song *Help* by the Beatles as an example, it seems that the audio sections that were retained in the mask are coherent with the annotated structure (see Figure 4.8).

Segmentation masks thus contain very relevant information regarding musical objects in the original image and it is therefore very tempting to directly segment the objects defined by the mask for further audio processing. However, the segmentation mask is binary and even though indicating meaningful regions, it loses information about the similarity of the structural part with itself. In another word, temporal evolution of similarity within the parts is lost. Also, and as shown in



(a) Original Matrix

(b) Diagonal Smoothing

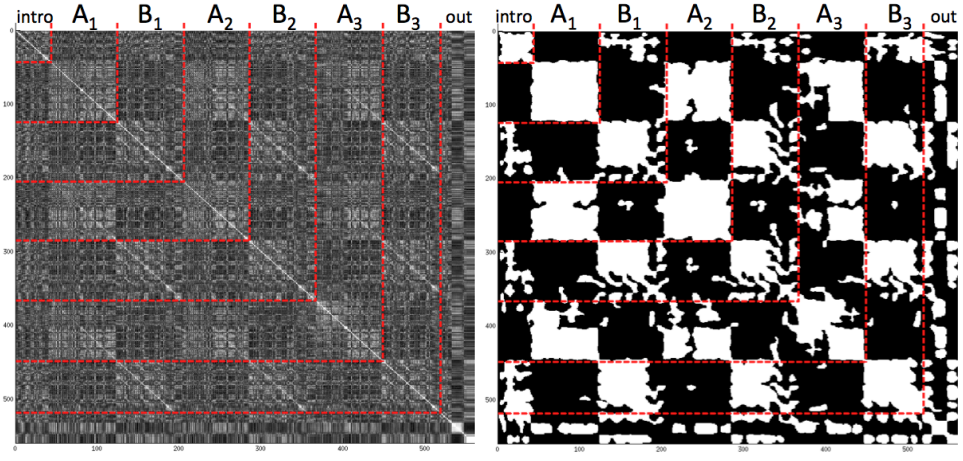
(c) Low-pass Filtering

(d) Binarization

(e) Open Operation

(f) Close Operation

Figure 4.7: Segmentation mask generation with the song *Help* by the Beatles



(a) Original Matrix (b) Segmentation Mask

Figure 4.8: Original Similarity Matrix S (a) and the obtained segmentation mask M (b) with annotated structure

Figure 4.11, the mask might under-segment the similarity matrix, in which case structural information is completely lost. Using the binary mask only can thus be prejudicial to the description of music pieces and to the final application. In order to maintain this temporal information, we consider the segmentation mask as a weighting matrix for the enhancement process. Elements in the original matrix $\mathbf{S}$ that were retained in the mask $\mathbf{M}$ are multiplied by a certain weight w, whereas unretained elements remain unchanged. In the resulting matrix $\mathbf{S}_e$, regions of musical interest are strengthened and the variation in similarity levels is kept.

$$\mathbf{S}_e = \mathbf{S} \cdot (w - 1) \cdot (\mathbf{M} + \mathbf{O}_1) \tag{4.13}$$

where $\mathbf{S}_e$ and $\mathbf{M}$ are the enhanced similarity matrix and the mask respectively. $\mathbf{O}_1$ is a matrix of the same size as $\mathbf{S}$ whose elements all equal to one. It ensures that the original matrix information $\mathbf{S}$ will be retained. w is a scalar that weights the elements of the segmentation masks in the original similarity matrix.

Figure 4.9 shows the final enhanced matrix for our example using a weight of 3.

4.3.2 Related Issues

Our matrix enhancement approach seems to work fine with state representations of structure that assume that musical sections are rather self-similar and therefore form blocks in the similarity matrix. Considering the music piece example shown in figures 4.8 and 4.9 the matrix enhancement clearly strengthened the state representation. We show an other example of a well defined state representation in Figure 4.10. The song example is *Everybody is trying to be my baby* by the Beatles.

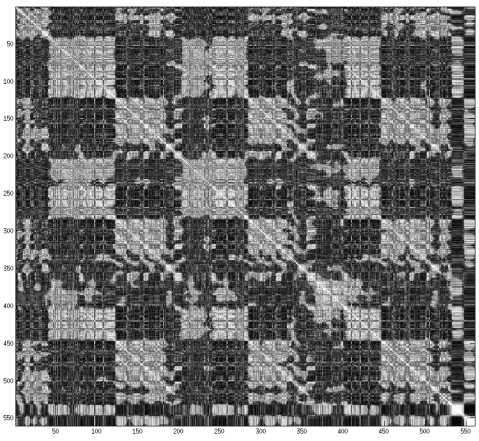


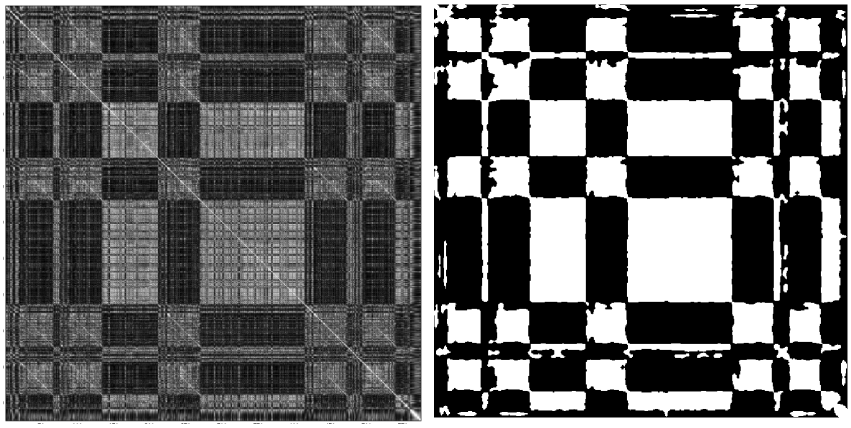
Figure 4.9: Enhanced similarity matrix S_e with mean of the segmentation mask for the song *Help* by the Beatles. Original similarity matrix is computed by means of the mfcc features

Musical sections are very homogenous in timbre in the piece, thus yielding well-defined structural blocks in the similarity matrix. In that case the segmentation mask perfectly retains musical sections and discards regions of dissimilarity.

In contrast to the state representation, the sequence representation considers series of frames that are repeated over the music piece, frames within a section not necessarily being similar. Structure is then displayed in the similarity matrix by dominant repetitive motives on the off-diagonals. For audio material that fits the sequence representation and as shown in Figure 4.11, our enhancement processing is not adequate yet. Indeed, while segments represented as states are retained in the segmentation mask, most of the sequence motives on the off-diagonals are discarded. Image segmentation techniques that fit the sequence representation should be considered in further developments.

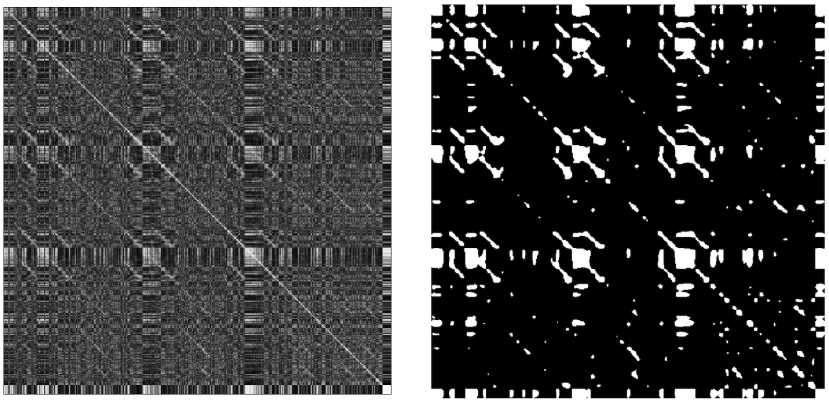
4.4 Application to NSMF

With the enhancement of state representations we initially intended to improve the performance of the NSMF structure segmentation algorithm introduced in Chapter 3. In this section, we will first illustrate how the separation of structural sections can be improved by means of an enhanced similarity matrix example. We will then present a preliminary evaluation.



(a) Original Matrix (b) Segmentation Mask

Figure 4.10: Similarity matrix for the song *Everybody is trying to be my baby* performed by the Beatles (a) and the corresponding segmentation mask (b)



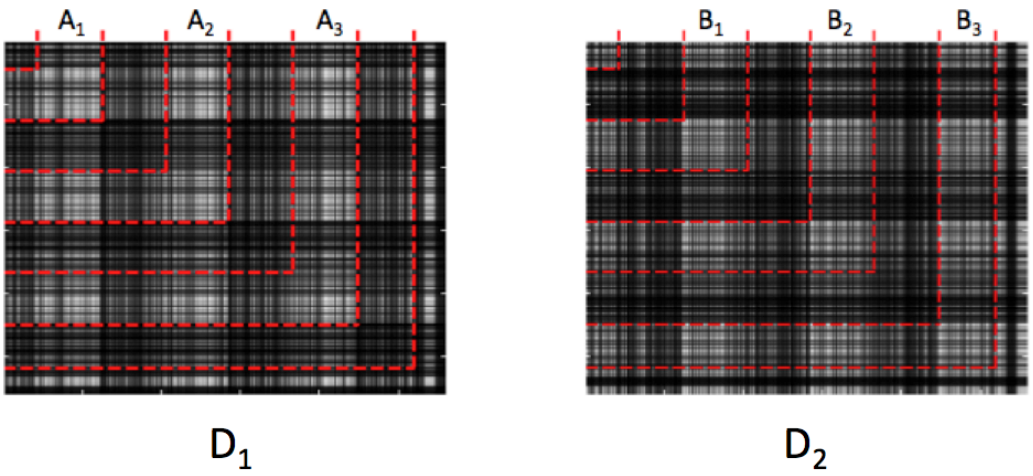
(a) Original Matrix (b) Segmentation Mask

Figure 4.11: Similarity matrix for the song *Think for your self* by the Beatles (a) and the corresponding segmentation mask (b)

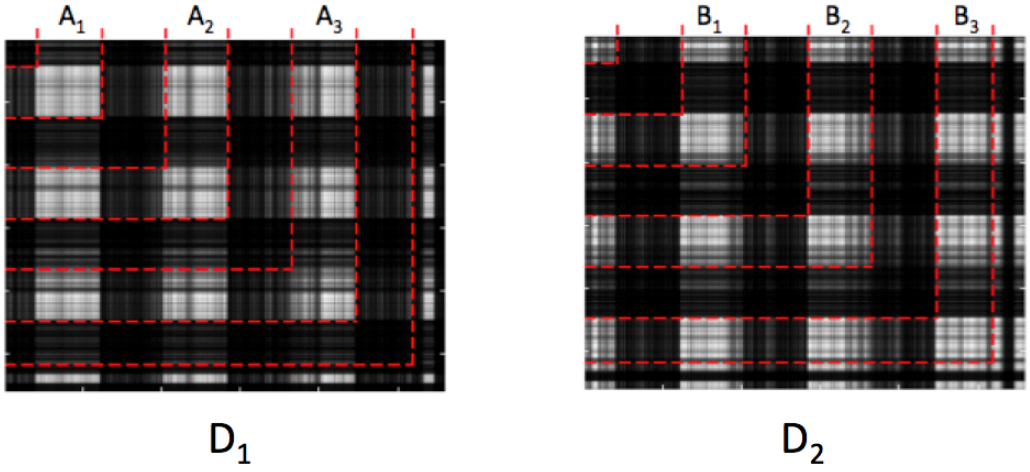
4.4.1 Decomposition of Enhanced Matrices

We use once again the song example *Help* by the Beatles. After extracting the timbral similarity matrix, we apply our matrix enhancement procedure with a weight of 2. The NMF of rank 2 of the original and enhanced matrices are shown in Figure 4.12 together with the structure annotation.

As already shown in the previous Chapter of this thesis, the NMF decomposition of the original matrix can be interpreted for the structure estimation. The state enhancement however significantly strengthens the separation between parts, yielding a much sparser decomposition. Indeed, each of the two main parts of the song are



(a) 2^{nd} order NMF Decompositon of the original similarity matrix



(b) 2^{nd} order NMF Decompositon of the enhanced similarity matrix

Figure 4.12: Separation of structural parts of the song *Help* by *The Beatles* with mean of the NMF decomposition of the original matrix (a), and of the enhanced matrix (b)

almost perfectly reconstructed over a single dimension of the decomposition. This is further illustrated in Figure 4.13 where the enhanced matrix is now decomposed by means of a NMF of rank 3 and the structural sections are represented over the dimensions of the matrix W.

While in Chapter 3 several dimensions of the NMF where needed to define clusters of points specific to a structural section, each dimension of the NMF now almost exclusively models a single structural section.

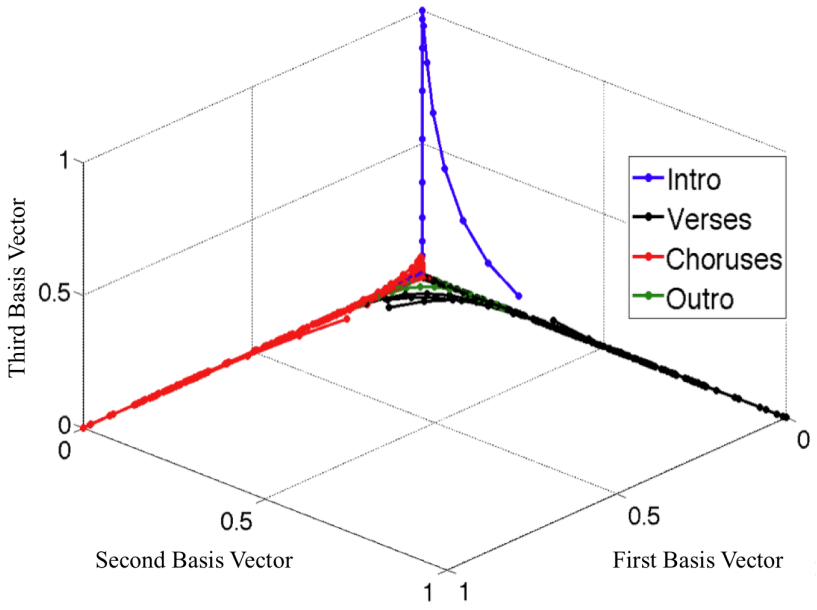


Figure 4.13: Projection of structural parts on the NSMF basis vectors

4.4.2 Preliminary Evaluation

We now present a small evaluation of the combination of the NSMF structure segmentation and the introduced matrix enhancement procedure. A deeper evaluation with larger datasets is provided in Chapter 6.

We use for that purpose the album *Help!* by the Beatles that consists of 14 songs. Annotation was taken from the *TUT Beatles*[1] dataset. Similarity matrices are computed by means of timbre-related audio features, and their enhanced versions are used as inputs for the NSMF structure segmentation. Evaluation of the structure segmentation is done using the pairwise F Measure, Precision and Recall (*cf* Section 2.5.2). We compare the performance of the structure segmentation using the segmentation masks solo (referred to as *NSMF + Mask*) and using the enhanced matrices for different weightings (referred to as *NSMF + Enhancement*). Results are reported in Table 4.1.

Using the enhanced similarity matrices clearly seems to improve the overall performance of the structure detection, gaining up to 3.4% in terms of F-measure. While the precision is not increased but remain in the same range as NSMF solo, we achieve much better recall rates (up to 8.3 %) in any of the proposed cases. We raised in the last chapter the issue of over-segmentation and this increase in the recall rates suggests that the enhancement procedure can better cope with it. Indeed, while low-level audio features are easily changed by small acoustic events, the smoothing of the enhancement procedure provides more homogeneity in the

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

Algorithm	F	P	R
NSMF solo	63.5%	**64.7%**	65.7%
NSMF + Mask	64.8%	62.5%	69.5%
NSMF + Enhancement ($w = 2$)	65.0%	61.7%	72.3%
NSMF + Enhancement ($w = 3$)	**66.9%**	63.4%	**74.0%**
NSMF + Enhancement ($w = 4$)	64.3%	62.1%	71.1%
NSMF + Enhancement ($w = 5$)	63.8%	62.0%	69.9%

Table 4.1: Performance of NSMF combined with image-based structure enhancement

description of sections. The results also confirm that it is not worth using the binary segmentation mask alone for the analysis.

4.5 Chapter Summary

We have presented an image-oriented post-processing of audio similarity matrices for the enhancement of the structure visualization. Doing so, we reduce the complexity of the matrices and the discrimination of musical sections is improved. Preliminary evaluation shows that this approach consistently improves the performances of the NSMF music structure segmentation algorithm. Especially, the enhanced matrices seem to cope with over-segmentation issues. Sequence representations however confuse the algorithm and structural information is mainly lost in such cases. In the next Chapter we will thus aim at further enhancing the state representation at its root, meaning by further describing the audio signal before embedding the features in any visualization.

Chapter 5

Modeling Tonal Structures

"Wherever we are, what we hear is mostly noise.
When we ignore it, it disturbs us.
When we listen to it, we find it fascinating"
John Cage

Contents

We proposed in Chapter 3 a structure segmentation algorithm that works under the state assumption. The nature of low-level audio features however often prevents from extracting well-defined state representations and an image-based post-processing method was proposed in Chapter 4 to strengthen the state representation in similarity matrices by means of adequate filtering and exploiting pixel connectivity with morphological operators. However, image processing of similarity matrices

does not tell more about the musical content and was designed to facilitate the classification with NSMF. In this chapter we are interested in further describing the audio waveform by focusing on the harmonic attributes of sections. Therefore, a mid-level audio descriptor that reflects the harmonic context of a given audio segment is proposed. Embedding this contextual information in similarity matrices, acoustical homogeneity within sections is better characterized and the state representation is favored. Combining such matrices with the NSMF approach we improve our results for the music structure segmentation task.

5.1 Introduction

Using similarity matrices for the task of music structure segmentation, one aims at enlightening repetitive patterns or homogeneous segments within audio features distributions. Similarity matrix based music segmentation algorithms are thus usually designed to exploit the sequence or state representation of the music piece's structure [Peeters 2004a]. However, and as seen in Chapter 3, the assumptions on the audio signal, and thus the music piece, that underly proper sequence and state visualizations are rather idealistic. We reported in the state of the art review in Chapter 2 that sequence-based methods therefore make use of diverse filtering or matrix transformation techniques to strengthen stripes in similarity matrices. We have exploited in the last Chapter the same kind of approaches to enhance the state representation.

However, the main issue in the visualization of states remains in the gap between the long-term acoustical homogeneity that supposedly characterizes musical sections and the rather short-time extraction scale of low-level audio features. Take as an example popular music songs in which sections usually last about 20 seconds. Extracting low-level audio features over 250 ms analysis windows, one tends to characterize low-level structural elements such as notes and chords rather than a global acoustical or musical context of tens of seconds.

A couple of works on including contextual information in the description of time instants of the audio signal have thus arisen in the literature. In [Peeters 2002b] Peeters proposed a dynamic feature that models the spectral envelop over a short period of time, and thus describes contextual timbre information. In [Barrington 2010] Dynamic Texture is applied to model timbral and rhythmical properties of sounds. Müller et al. proposed in [Mueller 2006] a contextual measure of similarity that considers sequences of feature frames instead of single frames in the similarity matrix computation. However, no mid-level feature was proposed to describe the tonal context of musical segments.

In this chapter we therefore focus on the definition of a mid-level feature that would capture in the audio signal the longer term musical homogeneity of a section

in terms of its tonal attributes. This is motivated by the fact that a single note not only is characterized by its pitch and intensity but also by the tonal context that defines its tonal function and relative importance in a melodic line. Reflecting this tonal contextual information in a mid-level feature, i.e. extracted over an analysis window of a couple of seconds, we intend to better fit the definition of musical sections as steady musical states according to a musically meaningful property that is the tonal structure.

5.1.1 Importance of the Tonal Context

Chopin's Mazurka op.63, no.3 is a good example for illustrating the importance of the tonal context while measuring the similarity between two time instants of the audio signal. A Mazurka is a Polish folk dance whose structure is rather simple and consists of a main theme that is repeated over the piece. Though Chopin kept some of the traditional aspects of Mazurkas in his compositions, he developed a new genre by incorporating other polish traditional rhythmical elements, and by developing the harmony of the pieces. In our example, the Mazurka op.63, no.3, the piece is composed in $c^{\#}$ minor but contains a 16 bars interlude section in b^{b} minor. The score of the transition from the last 6 bars of the b^{b} minor section to the first bars of the main theme is shown in Figure 5.1.

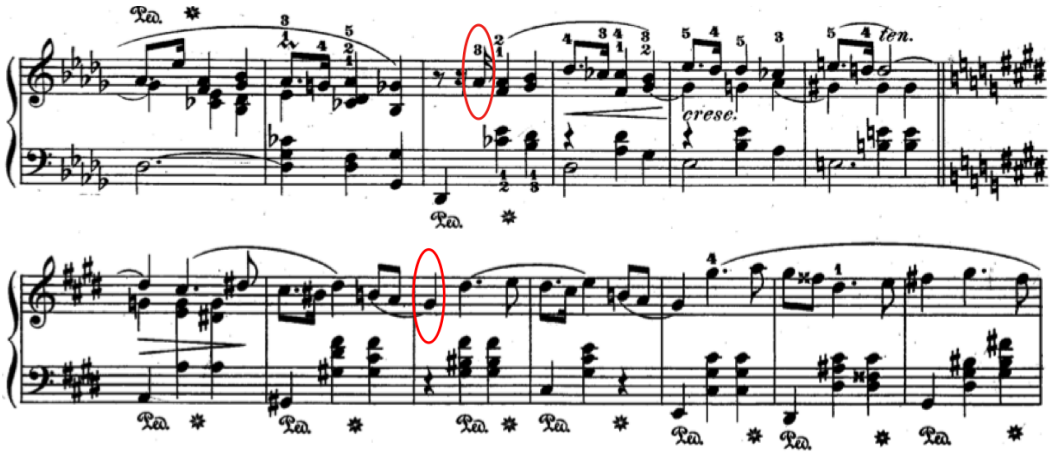


Figure 5.1: Excerpt of Chopin's Mazurka, op. 63, no.3

Two notes are highlighted in the score: an A^{b} and a $G^{\#}$. From an audio signal point of view, the two notes have the same pitch height and are purely equivalent. Consequently, the similarity between the low-level descriptors extracted over the two notes is very high. However, as part of two different sections composed in different keys, the two notes have different tonal functions and should not be considered as similar. Thus, not considering the context in which the notes are played can yield a confused structure representation in similarity matrices (see Figure 5.17).

Audio signal description tools should embed this musical context while describ-

ing an audio segment and capture information over a longer musical moment than the moment of a note. Including past- and forthcoming musical moments in the description of our two example notes, one would immediately reveal by comparing these two descriptions that they are not part of the same melodic line nor tonal context. In contrast, if information of the tonal context is embedded in the description of musical events of the same section, the regularity and self-similarity of the section will be strengthened in the similarity matrix.

5.1.2 Approach

Music cognition studies on the recognition of music genres and styles have shown that humans have a quite impressive ability to rapidly (≈400ms) distinguish music genres within the broad palette of western music styles [Gjerdingen 2008] [Krumhansl 2010]. Developing on the factors that explain this amazing ability is way beyond the scope of this thesis, but we will only note that the music attributes that influence the acquisition of this long-term memory not only concern instrumentation but also the particular scales and tonal structures that are associated to each genre. We understand here with tonal structure the hierarchical relation between tones of a musical scale. Such structures produce very specific sound colors and may discriminate music genres and styles within a genre.

One of the particularities of using specific scales and tonal structures for musical expression is that it necessarily implies a kind of regularity in the tone intervals and chord sequences that are employed in the composition. For example the sequence of tones in a major scale differs with the one in a minor scale. Also, typical chord progressions appear in classical music such as the resolution of a tension with the fifth degree chord of the tonic to its first degree chord. Such regularities play a significant role in the development of musical expectation and acquisition of an implicit musical knowledge by humans [Krumhansl 1979] [Jones 1989]. This is for example why cultural background plays a major role in experiencing music genres of different regions of the world. Europeans are much more able to detect false notes in a classical music composition than in an indian Raga. And the reciprocal is also verified.

The approach we propose in this chapter to characterize the tonal context of a musical segment consists in taking advantage of these regularities. Indeed, we claim that a hierarchy in the tones implies that a restricted set of tone intervals is dominant in a given tonal context. Including statistics about local dominant intervals of a given time instant in the audio signal, we can thus embed the tonal contextual information associated to this segment.

We use for that purpose the Multi-Probe Histograms (MPH) that were proposed in [Yu 2010] in the context of scalable audio retrieval. The idea of such histograms is to summarize sequences of chroma vectors by their most dominant local pitch class transitions. Here, local means that transitions are probed between adjacent

chroma frames of the sequence. While Multi-Probe Histograms were never applied to music content description to our knowledge, their musical interpretation hasn't been discussed in the literature. In this chapter we verify the hypothesis that MPH characterize the regularities induced by a tonal context, and we apply this result to the task of music structure segmentation.

After introducing the Multi-Probe Histograms (section 5.2), we will discuss their musical interpretation(section 5.3). Musical scales, tonal structures and illustration with the work of music cognitivists will be presented. We then validate the interpretation of MPH as tonality-related features by means of a preludes classification task. Finally we introduce the histograms as mid-level descriptors for the task of structure segmentation and show a preliminary evaluation (section 5.4).

5.2 Multi-Probe Histograms

5.2.1 Motivation

Multi-Probe Histograms (MPH) were introduced by Yu et al. in the context of scalable audio retrieval [Yu 2010]. While the computation and musical interpretation of such histograms are discussed in the next section of this chapter, we briefly introduce here the motivation of applying the work of Yu et al. to the description of tonal contexts.

The task of audio retrieval can briefly be described as measuring the similarity between a given audio query and a large audio database. One way to do so consists in measuring the similarity between the chroma vectors extracted over the audio files of the database and the chroma distribution of the query. Because computation cost is a major issue when dealing with large-scale audio databases, Yu et al. proposed to summarize the chroma sequences of each audio file of the database in a Multi-Probe Histogram (see Figure 5.2). Each file is then represented by a reduced fixed-size histogram and similarity between histograms is easily computable.

To give a simplified description of the histograms, each bin counts the frequency of a given transition between the dominant pitch classes of adjacent frames of the chroma sequence. Hence, the chroma sequence is summarized by dominant local pitch class transitions. Besides the fact that this method, combined with the use of hash tables, seemed to perform very well for the task of audio retrieval, one of the interesting conclusions of the paper was also that only few major bins of the histograms were necessary to achieve good performance. Yu et al. claim that this results from the redundancy of melodies in the audio files. But we also note that the mid-term temporal evolution of the chroma frames is not taken into account in the histograms. Melodies are thus not modeled by long-term sequences of pitch classes, but by a rather simple statistic on local pitch class transitions.

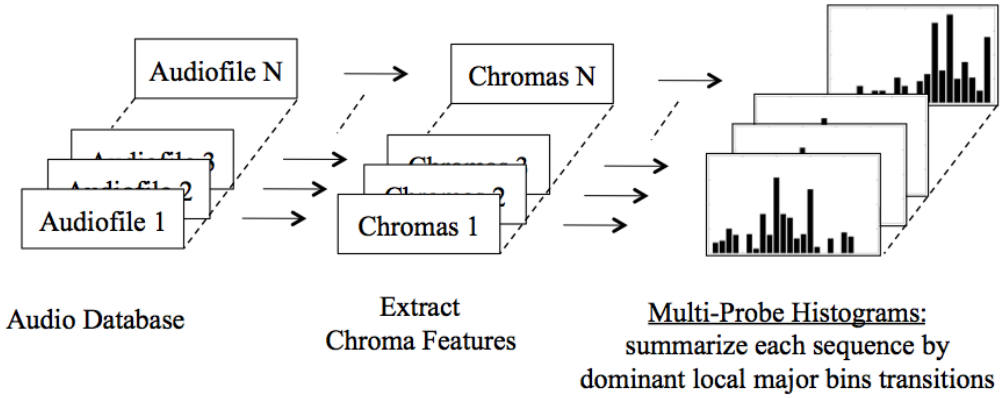


Figure 5.2: Approach to large-scale audio retrieval in [Yu 2010]

Our hypothesis is thus that the histograms capture the invariance in the pitch class transitions that is implied by the tonal structure of the audio files. As tonal structures also evolve within music pieces, we propose to use the histograms as mid-level audio features for audio content description purposes (see section 5.4).

5.2.2 Computation

Chroma features are extracted over neighboring windows of the audio signal (see section 2.2 of Chapter 2). There is thus a strong correlation between adjacent feature frames, unless a strong transition occurs in the audio signal, such as a note change, and in that case, only the first dominant bins of the chroma sequence really variate. Multi-Probe Histograms are determined by these invariances and transitions, and aim at characterizing chroma sequences by probing dominant pitch class transitions between adjacent frames.

Computation of the histograms is illustrated in Figure 5.3. A complete description of the algorithm for their extraction can be found in [Yu 2010].

In Figure 5.3, the two adjacent chroma frames that serve for the illustration correspond to a transition from a C Major chord (frame i) to a F Major chord (frame $i+1$). There are 12^2 possibilities for the maximum of energy to be transferred from a pitch class to an other from frame i to frame $i+1$. Each of these possible transitions defines a position in the Multi-Probe Histogram. In our example, the energy is logically transferred from the pitch classes of the major triad of a C to the major triad of an F. If we now consider the three pitch classes that have the maximum of energy, i.e. C-E-G for the C Major chord and F-A-C for the F Major chord, and as illustrated above, the transition from frame i to frame $i+1$ allows $3^2 = 9$ possible transfers of energy between those pitch classes. For this iteration, bins of the MPH that will be allocated a new value are defined by these 9 transitions. Considering the transition from the tone C to the tone F, a bin position in the

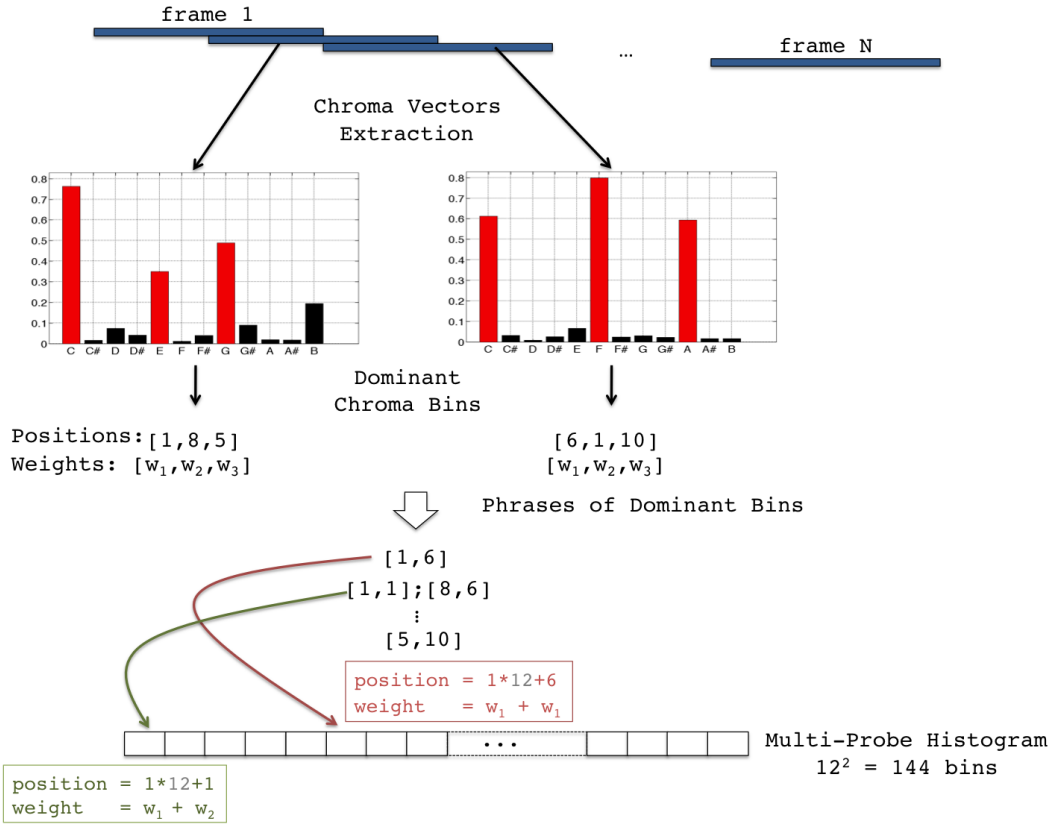


Figure 5.3: Multi-Probe Histogram computation example for a sequence of N frames

histogram is computed as in equation 5.1.

$$p = b_C * 12 + b_F \tag{5.1}$$

with b_C and b_F the positions of the pitch classes C and F in the chroma vector, respectively 1 and 6. Furthermore, each bin is affected a weight according to its relative energy in the chroma frame, from w_1 to w_3 in our example. C and F being the first maximums of energy in our example, the added weight for our histogram bin is defined as in equation 5.2.

$$w = w_1 + w_1 \tag{5.2}$$

meaning that the histogram is added the value w at its bin p. The operation is then repeated for the remaining 8 considered pitch class transitions, and iterated over the remaining frames of the chroma sequence.

The actual values of major pitch classes in the chroma vectors do not influence the histogram's value. Only the distance between dominant pitch class bin transitions does. In our example, we considered sequences of two frames, probing the combinations between the first three major bins. These two parameters are

1. Extract chroma features C
2. For each frame c_i of C
3. | Sort the 12 bins by decreasing energy
   | original positions of top K major bins are $m_{i,k}$, for $k = 1,2,\ldots,K$
4. | Assign weights $w_{i,k}$, for $k = 1,2,\ldots,K$
5. Initialize the 12^J bins of the MPH to zero
6. Compute the matrix *Bin_Phrases* of all possible combinations of K major bins in phrases of J frames
7. For *frame* $= 1, \ldots, nb\_frames(C) - J + 1$
8. | For $l = 1, \ldots, K^J$
9. | | k = *Bin_Phrases* (l , 1)
10. | | bin = $m_{\text{frame},k} \cdot 12^{J-1}$
11. | | weight = $w_{\text{frame},k}$
12. | | For $j = 2, \ldots, J$
13. | | | k = *Bin_Phrases* (l , j)
14. | | | bin = bin + $m_{\text{frame+j-1},k} \cdot 12^{J-j}$
15. | | | weight = weight + $w_{\text{frame+j-1},k}$
16. | MPH (bin) = MPH (bin) + weight
17. Normalize MPH

Figure 5.4: MPH Computation Steps

arbitrary and chosen depending on the application. Increasing the length J of the chroma frames sequence, one increases the size of the histogram 12^J. By increasing the number K of chroma bins considered per frame, the number K^J of probed chroma bins phrases increases.

The range of MPH bins depends on the length of the chroma sequence over which it is computed. When extracting histograms over the chroma sequences of whole music pieces (as in section 5.3), normalization is applied to allow for their comparison. Using MPH as a mid-level feature, the length of the chroma sequences is fixed and determined by the length of the window we choose for the analysis. Even if the range of histograms extracted over different windows is comparable, we still apply normalization for cases in which a section is repeated at a different tempo.

The computation steps of a MPH is shown in Figure 5.4.

Multi-Probe Histograms used in the remainder of this thesis are computed with:

- Sequences of $J = 2$ frames
- $K = 4$ major chroma bins considered per frame

Thus probing $2^4 = 16$ pitch class transitions in phrases of 2 frames. The size of the obtained histograms is of $12^2 = 144$ bins.

5.2.3 MPH Example

To illustrate the information conveyed in Multi-Probe Histograms, we compute the histogram with the chroma sequence extracted over a small excerpt of the Mazurka, op. 63, no.3 by Chopin. The obtained histogram is plotted in Figure 5.5. This illustrates how bins of the MPH count transitions between pitch classes of the western chromatic scale. The audio portion of the mazurka we have chosen is composed in b^b minor. Its key signature is thus composed of five flats, i.e. b^b, e^b, a^b, d^b and g^b. Pitch classes transitions revealed by the histogram seem coherent with this key signature.

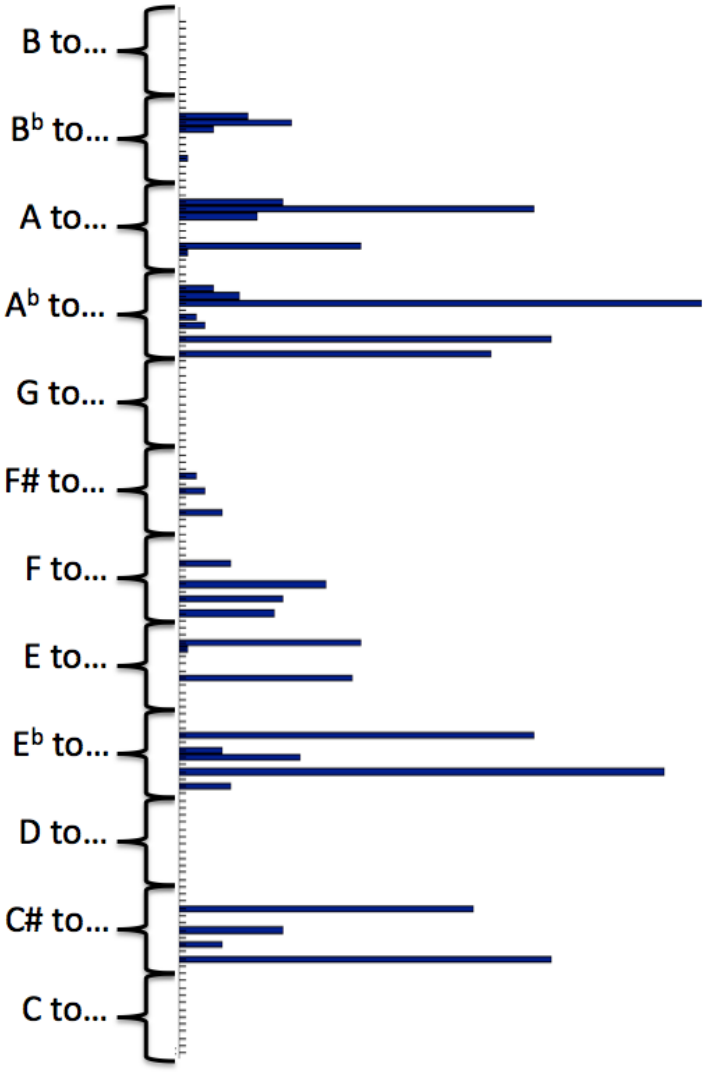


Figure 5.5: Chroma based Multi-Probe Histogram of an excerpt of the Mazurka, op. 63, no.3 composed by Frédéric Chopin

5.3 Musical Interpretation

5.3.1 Preamble

Giving a musical interpretation in the sense of music theory to a music information retrieval algorithm can be quite ambiguous. In our case, we claim that Multi-Probe Histograms can characterize the tonal structure of the studied musical segment. In music theory, a tonal structure refers to the intervals between tones of the scale, i.e. unique and distinct pitch heights. In our analysis based on chroma vectors, the pitch content is however concatenated in pitch classes that sum the contribution of fundamental tones over the whole spectrum, or at least a few octaves. Hence, octave information is lost and a transition between pitch classes in a sequence of chroma frames can under no circumstances be considered as a tone interval. Plus, chroma vectors are not completely robust to timbre variations and the approximation of pitch classes might be confused with tone intervals that share identical notes in different octaves.

Nonetheless, the music arrangement of compositions often inverse or split the tones of chords over several octaves. For example when a C major triad C4-E4-G4 is played as E4-G4-C5. Tone intervals dictated by tonal structures are thus not always played in the range of an octave. Even if very important, the octave information is thus not critical to describe tonal structures in music pieces. Without claiming that pitch class transitions approximate tone intervals, we can thus state that such intervals are strongly reflected in pitch class transitions.

5.3.2 Musical Scales and Tonal Structure

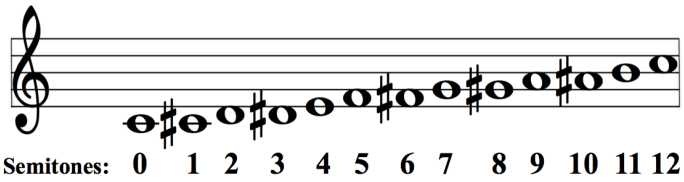
Independently of the harmonic rules that are specific to the various musical genres, western music is mostly tonal or modal. This means that it is based on musical scales that imply specific hierarchical relationships between tones that we call tonal structures. We will now briefly introduce some basic musical knowledge that will help understanding this concept and then illustrate it with the work of music cognitivists on the perception of tonal hierarchy.

A musical scale is defined as a sequence of tones of the chromatic scale (see Figure 5.6.a). This sequence is determined by the intervals between each of its tones. In the ascending and descending chromatic scales, all neighboring tones are equidistant by an interval of a semitone, i.e. the smallest musical interval in western music. If we now consider the ascending C major scale (see Figure 5.6.a), it defines a specific sequence of intervals between its tones and intervals with its tonic C that is described in Table 5.1.

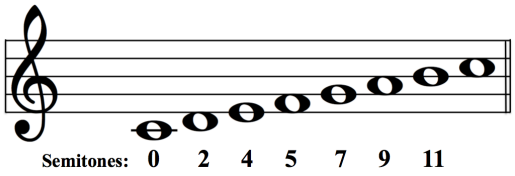
This specific sequence has a sound color that is typical of major modes and is easily recognizable. Indeed, tonalities in the major mode keep the same sequence of intervals but centered on a different tonic. The D major scale, that is altered by two flats at its key signature, follows the exact same sequence of intervals, but

Semitones	Note	Diatonic Interval
0	C	Tonic
2	D	major $2^n d$
4	E	major $3^r d$
5	F	perfect $4^t h$
7	G	perfect $5^t h$
9	A	major $6^t h$
11	B	major $7^t h$

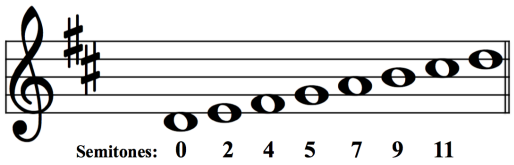
Table 5.1: Sequence of diatonic intervals of the C major scale



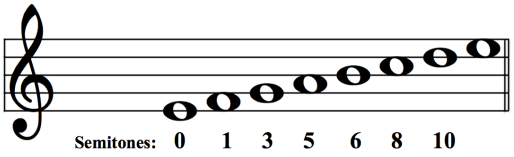
(a) Ascending chromatic scale



(b) Ascending diatonic C major scale



(c) Ascending diatonic D major scale



(d) Ascending phrygian scale centered on E

Figure 5.6: Example of musical scales

between different tones (see Figure 5.6.c). The tonic constitutes the most stable structural element of the scale and therefore leads the hierarchy. It is then followed by the two most consonant intervals that are the major fifth and the major third.

Then comes the scale tones and finally the non-scale tones. Each tone thus has a specific relation with the tonic. Consequently, a largely dominant chord in a major tonal context is the major triad of the tonic, i.e. three tones chord that is composed of the tonic (for example C), its major third (E for C) and perfect fifth (G for C). A simple melody composed in a major tonality will gravitate around these three tones.

Depending on the musical intention, composers may use various scales. For example in modal music, the phrygian scale (see Figure 5.6.c), or E mode, defines a new tonal structure in which the presence of a minor second produces a very particular sound color that recalls Spanish music.

To extent the notion of tonal structure to harmonic considerations we may cite Arnold Schoenberg:

> *"A triad standing alone is entirely indefinite in its harmonic meaning; it may be the tonic of one tonality or one degree of several others. The addition of one or more other triads can restrict its meaning to a lesser number of tonalities. A certain order promotes such a succession of chords to the function of a progression."* [Schoenberg 1969]

This means that the scale and triads of a given tonality imply a tonal hierarchy that is not strong enough to restrict its meaning to a single or a few tonal contexts. To enrich the sound color of a composition, composers indeed make use of more complex chords. The scale and its tonal hierarchy are still the basis of a tonality, but including more sounds in the chords construction, each degree of the scale produces a chord that gets more characteristic of the tonality with the number of added tones.

The sequence of tertian, or 7^th chords, derived from the C major scale that is noted in Figure 5.7 perfectly illustrates Schoenberg's statement.

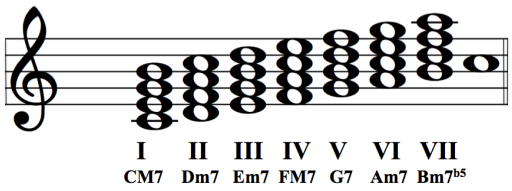


Figure 5.7: Tertian chords of the C major scale and corresponding degrees

The II, III and VI degrees all produce minor $7th$, the I and IV degrees major $7th$ but only the V degree produces a dominant $7th$, i.e. a major triad with a minor $7th$. Hence, the meaning of the particular intervals that compose a dominant $7th$ is restricted to a single tonality that is the one of the tonic of its fundamental tone. A certain order in the succession of chords is also dictated by the rules of harmony.

For example in classical music, a tension in the tonality can be resolved with a transition from the V to the I degree.

If we now inverse the previous statement, i.e. the meaning of some chords is restricted to a tonal context, we can say that a tonality produces a restricted set of tone intervals that characterize it. In that sense, studying the tonal hierarchy of musical segments seems very relevant to model their tonal context.

5.3.3 Illustration with Music Cognition Studies

Music cognition literature provides us with extensive works and studies on the perception of tone structures, verifying a strong dependency of human perception of tone intervals and chords with tonality contexts. One of the goal of these studies was to demonstrate that the definition of pitch as the psychological sensation resulting from the frequency of a note had to be added the dimension of its tonal context. The probe tone experiment was thus proposed by Krumhansl and Shepard in [Krumhansl 1979]. Mixed groups of people of diverse musical training (from non-musicians to outstanding musicians) were formed and asked to rate tone intervals depending on how well they completed a C major scale. This way, authors were able to quantify the perception of tonal hierarchies.

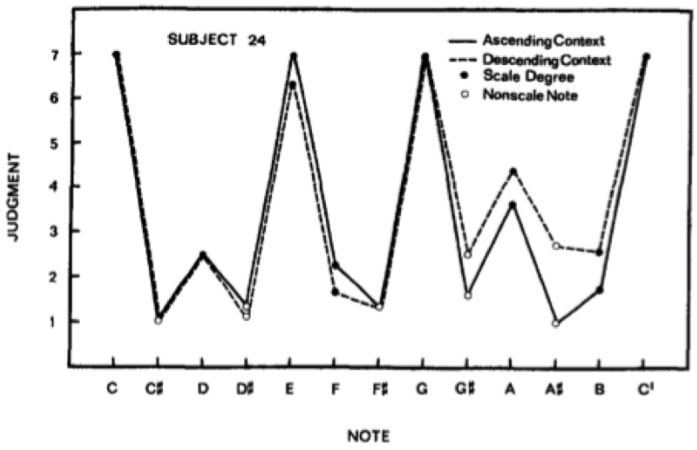


Figure 5.8: Judgment by an outstanding musician of the quality tone intervals completing a C major scale [Krumhansl 1979]

Figure 5.8 shows the judgment curve of a highly trained musician (subject had absolute pitch) that is perfectly coherent with the tonal hierarchy dictated by the music theory. This was a strong tendency in the experiment results for musically trained subjects, verifying the reality of a tonal hierarchy underlying a tonal context.

The result curves produced by non-musician subjects showed a major preference for tones close in frequency. Tones of the scale (D, E, F, G, A, B) were however significantly preferred over non-scale tones (C#,D#,F#,G#, A#), with a slight

preference for the major fifth and major third. This suggests that the physical properties of tones, i.e. closeness in frequency, are not sufficient to explain the perception of pitch heights and that the tonal context somehow influenced the subjects. One of the hypothesis to explain this is that listeners acquired an implicit musical expectation after a long-term exposure to western music and to the regularities induced by its tonal hierarchy.

The consistency of harmony hierarchies with tonal hierarchies was also studied in [Krumhansl 1991]. Harmony hierarchies are illustrated by the circle-of-fifths, in which the 12 keys of the chromatic scale in their major and minor mode are structurally linked. In this hierarchy, a tonality has the most structural significance with the tonality of its perfect fifth and fourth. Extending the probe tone experiment, Krumhansl measured the similarity between tonal hierarchies of different keys and showed a strong correlation with the circle-of-fifth. We will use this result for the experimental validation of the MPH as a tonal context descriptor with the preludes classification task in section 5.3.4.

Finally, an interesting study by Cohen in [Cohen 1991] on the establishment of a key in a music composition showed that the four notes of a music piece (in that case excerpts of the *Well-Tempered Clavier* Books) were sufficient for 75% of a group of musically trained listeners to estimate the tonic of the key. Once again, this verifies that salient tones and intervals may strongly characterize a tonal context and that this context can be locally modeled. This result justifies the use of MPH as mid-level and not only global features.

5.3.4 Preludes Classification

5.3.4.1 Introduction

As stated in [Krumhansl 1991], harmony hierarchies and hierarchical similarities between tonal hierarchies are described in the circle of fifths (Figure 5.9). A simple method for verifying the interpretation of Multi-Probe Histograms as descriptors of the tonal context therefore consists in measuring the consistency of the similarity ratings based on MPH of music pieces composed in different keys with the circle of fifths. Indeed, if Multi-Probe Histograms of chroma vectors embed information that relate to harmony, the similarity between two pieces composed in different keys measured by means of the MPH should follow the same hierarchy as the distance between keys in the circle-of fifth.

We therefore consider sets of 24 preludes composed in each of the 12 keys in major and minor modes and aim at measuring their similarity by means of MPH.

The choice of the prelude form was motivated by its relative simplicity and fair illustration of the potentialities of harmonic compositions. Indeed, preludes are usually short music pieces that consist of few repetitive melodic and rhythmic motifs and introduce a more complex piece. It also stands on its own as a music piece in some musical era.

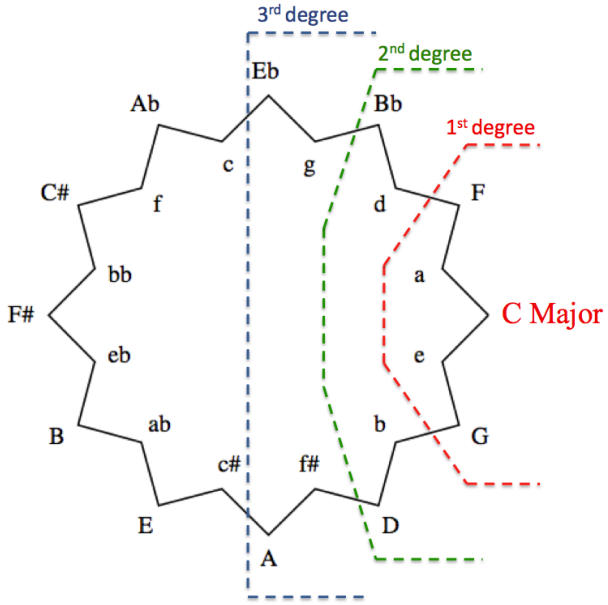


Figure 5.9: Circle of fifths and first three degrees of similarity of the C Major key

The studied preludes were taken from the following works:

a) 24 preludes of *The Well-Tempered Clavier* Book II (1744), by J.S. Bach
b) 24 preludes, Op. 28 (1839), by Frédéric Chopin
c) A collection of 24 preludes by Sergei Rachmaninov:
    1. Prelude, Op.3, No.2 (1882)
    2. 10 preludes, Op.23 (1903)
    3. 13 preludes, Op.32 (1910)

Preludes of the three composers cycle through all of the minor and major keys. Moreover, Cohen verified in [Cohen 1991] that consonance between pieces of the *The Well-Tempered Clavier* Books is consistent with the circle-of-fifth. It thus constitutes a good basis to verify the consistency of the MPH representation with the circle of fifths.

The medium length of preludes is as follows:

a) Bach: 02 minutes and 57 seconds
b) Chopin: 01 minute and 30 seconds
c) Rachmaninov: 03 minutes and 01 second

Bach's preludes analysis was done with midi files, whereas preludes of Chopin and Rachmaninov were studied using piano recordings. In the following analysis, we thus keep in mind that the chroma extraction might have been more precise with Bach's preludes than with the other preludes.

5.3.4.2 Maps of Similarity

For each set of preludes we extract a map of similarity between the midi and audio files. The extraction procedure is illustrated in Figure 5.10 and consists of the following steps:

1. Extract the chroma sequence of each Prelude (sampling rate = 10Hz)
2. Concatenate each sequence in a 144 bins Multi-Probe Histogram (K=12, J=2)
3. Measure the similarity between the 24 MPHs by means of the cosine distance

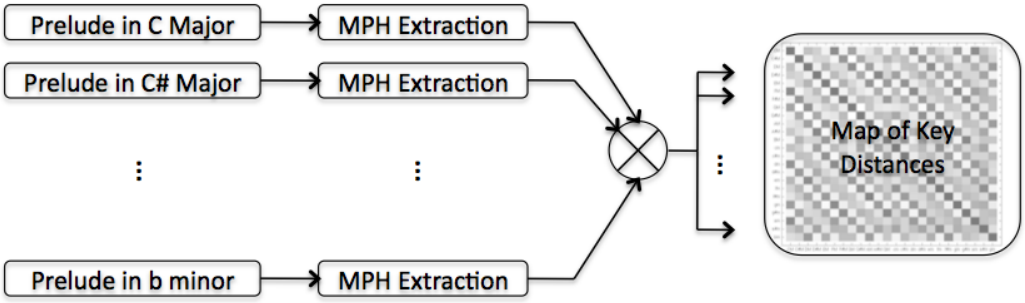


Figure 5.10: Measuring the similarity between 24 preludes extracting the MPH representation

Note that all preludes are composed for piano solo, and thus, the chroma vectors are extracted over mono instrumental pieces. This ensures that the pitch description we use is not disturbed by any timbre variations and that the analysis of tones is rather robust.

The map of similarity obtained for the preludes of the *Well-Tempered Clavier Books* is shown in Figure 5.11. In Table 5.2, the first 4 most similar pieces of each 24 preludes according to the MPH representation are listed. The complete similarity ranking between preludes of the three composers can be found in Appendix A.

From a rough visual analysis of Figure 5.11, it seems that preludes composed in keys close to each other in the circle of fifths have a high MPH similarity. On the other hand, pieces composed in keys that are highly distant are highly dissimilar in their MPH representations. Moreover, listing the four most MPH similar preludes as in Table 5.2, we find that the map of similarity of Bach's preludes, at least for keys of the same mode, is strongly consistent with the first degree of the circle of fifths. For example, the C Major prelude was found to be most similar with F Major and G Major preludes, which corresponds to the first degree of consonant major keys with the C Major in the circle of fifths. Hence, the MPH representation of pieces tends to correlate with the notion of consonance in an harmonic sense.

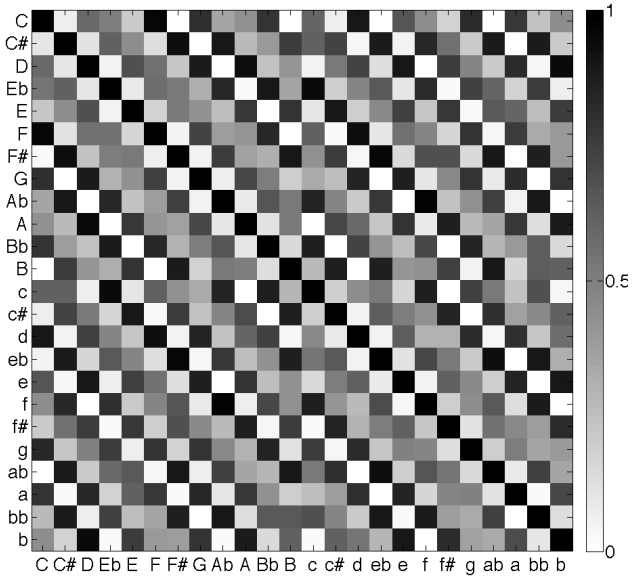


Figure 5.11: MPH similarity map of the 24 preludes of the *Well-Tempered Clavier* Books

Key	**C**	**C#**	**D**	**Eb**	**E**	**F**	**F#**	**G**	**Ab**	**A**	**Bb**	**B**
1	F	F#	b	c	c#	C	eb	D	f	D	Eb	ab
2	d	Ab	A	Bb	B	d	C#	e	bb	b	c	F#
3	g	eb	e	Ab	f#	Bb	ab	d	C#	f#	g	eb
4	G	bb	G	f	A	g	B	a	c	e	F	c#

key	**c**	**c#**	**d**	**eb**	**e**	**f**	**f#**	**g**	**ab**	**a**	**bb**	**b**
1	Eb	E	F	F#	D	Ab	A	Bb	eb	e	Ab	D
2	Bb	B	C	ab	b	bb	c#	C	B	G	eb	e
3	f	f#	G	bb	G	c	b	d	F#	D	C#	A
4	Ab	ab	g	C#	a	C#	E	F	C#	d	f	f#

Table 5.2: First 4 most similar preludes between pieces of the *Well-Tempered Clavier Books*

5.3.4.3 Measuring the consistency with the Circle of Fifths

While our maps of similarity seem musically meaningful, it is now interesting to have a deeper characterization of their consistency with the map of key distances defined by the circle-of-fifth. We therefore build a simple classification procedure for which we define the notion of similarity degree in the circle of fifth (see Figure 5.9). For example, the first degree of similarity, i.e. the more consonant relative keys, of a major key consists of the keys corresponding to the perfect fifth and

perfect fourth of its tonic in the major mode, and to the major third and minor sixth of its tonic in the minor mode.

Secondly, we distinguish the following relations between key modes:

1. major → major
2. major → minor
3. minor → minor
4. minor → major

This means that the similarity of a major key with other keys in the major mode is considered separately from the similarity of the same key with keys in the minor mode and vice versa.

According to these two criteria, i.e. degree in the circle of fifths and modes distinctions, we measure the accuracy of the maps of similarity between preludes according to the circle of fifths. For example, if we consider the first degree of the circle of fifths and the four keys measured by MPH as most consonant with the C major key, we have:

	Circle of fifth	MPH map of similarity
1.	C	C
2.	G	F
3.	F	d
4.	e	g
5.	a	G

The consistency is then measured as the percentage of correct keys besides the tonic that are present within the considered degrees, independently of the similarity rankings within these degrees. If we note A the set of major keys in the first degree of the circle of fifths, and B the set of the four most consonant preludes according to MPH, we compute the accuracy as:

$$accuracy = \frac{card(A \cap B)}{card(A)} \times 100 \tag{5.3}$$

In our example, the accuracy of our map of similarity for C major is thus of 100% major modes of the $1st$ degree and of 0%with minor modes of the $1st$ degree. The operation is then repeated for all preludes and final result is the mean accuracy.

5.3.4.4 Results and Discussion

We use a second set of maps of similarity that is computed by means of the mean value of the chroma vector sequences of preludes. To some extent, Multi-Probe Histograms can be thought as a sort of averaging of the chroma sequence. Hence, confronting it to another chroma vectors representation, we intend to enlighten the benefit of using the histograms for describing harmony.

Similarity with the 1st degree of the circle of fifths:

	Major	minor
Major	95.8%	50.0%
minor	54.1%	100%
Total	**75%**	

(a) Bach - 1^{st} degree

	Major	minor
Major	66.6%	25.0%
minor	29.1%	50.0%
Total	**38.5%**	

(b) Chopin - 1^{st} degree

	Major	minor
Major	79.1%	50.0%
minor	50.0%	62.5%
Total	**61.4%**	

(c) Rachmaninov - 1^{st} degree

Similarity up to the 2nd degree of the circle of fifths:

	Major	minor
Major	95.8%	70.8%
minor	66.6%	75.0%
Total	**77.0%**	

(d) Bach - 2^{nd} degree

	Major	minor
Major	66.6%	43.7%
minor	39.6%	52.1%
Total	**53.6%**	

(e) Chopin - 2^{nd} degree

	Major	minor
Major	70.8%	60.4%
minor	58.3%	54.2%
Total	**60.9%**	

(f) Rachmaninov - 2^{nd} degree

Similarity up to the 3rd degree of the circle of fifths:

	Major	minor
Major	97.2%	76.4%
minor	76.4%	90.2%
Total	**84.3%**	

(g) Bach - 3^{rd} degree

	Major	minor
Major	75.0%	61.1%
minor	62.5%	66.6%
Total	**64.9%**	

(h) Chopin - 3^{rd} degree

	Major	minor
Major	75.0%	68.0%
minor	68.0%	68.0%
Total	**71.5%**	

(i) Rachmaninov - 3^{rd} degree

Table 5.3: Consistency of preludes' maps of similarity with the circle of fifths measuring the similarity by means of the **Mean Chroma Vectors**

The results obtained with the mean value of chroma vectors are shown in Table 5.3 whereas the results for Multi-Probe Histograms are shown in Table 5.4.

We first notice that the results obtained for the Bach preludes are much better than those obtained for the two other sets. There is also no substantial difference between the mean chroma vectors and the MPH for that set. We explain that by two factors. First and as mentioned before, we extracted the chroma sequences of Bach preludes on midi files, in which the pitch content is more robustly extracted than in audio files. This might explain a slight, but not significant, better accuracy of the maps of similarity. Secondly, Bach preludes are essentially melodic and do not make use of dissonance in the chords' construction. Indeed, preludes sound very stable. In that case, the harmony information is mostly contained in the tones of the scale and better describing the tonal hierarchy by probing pitch class transitions, even if performing slightly better, does not seem to significantly increase the relevance of the maps of similarity. In contrast, some of the preludes of Chopin and

Similarity with the 1st degree of the circle of fifths:

	Major	minor
Major	100%	50.0%
minor	50.0%	100%
Total	**75%**	

(a) Bach - 1^{st} degree

	Major	minor
Major	75.0%	37.5%
minor	50.0%	79.1%
Total	**60.4%**	

(b) Chopin - 1^{st} degree

	Major	minor
Major	91.6%	70.8%
minor	66.6%	79.1%
Total	**77.0%**	

(c) Rachmaninoff - 1^{st} degree

Similarity up to the 2nd degree of the circle of fifths:

	Major	minor
Major	100%	68.7%
minor	64.6%	75%
Total	**78.1%**	

(d) Bach - 2^{nd} degree

	Major	minor
Major	87.5%	52.1%
minor	50.0%	66.7%
Total	**67.7%**	

(e) Chopin - 2^{nd} degree

	Major	minor
Major	83.3%	66.6%
minor	70.8%	54.2%
Total	**69.8%**	

(f) Rachmaninoff - 2^{nd} degree

Similarity up to the 3rd degree of the circle of fifths:

	Major	minor
Major	97.2%	75.0%
minor	76.4%	88.9%
Total	**84.7%**	

(g) Bach - 3^{rd} degree

	Major	minor
Major	86.1%	69.4%
minor	75.0%	77.7%
Total	**77.7%**	

(h) Chopin - 3^{rd} degree

	Major	minor
Major	83.3%	70.8%
minor	70.8%	75.0%
Total	**75.7%**	

(i) Rachmaninoff - 3^{rd} degree

Table 5.4: Consistency of preludes' maps of similarity with the circle of fifths measuring the similarity by means of Chroma vectors **Multi-Probe Histograms**

Rachmaninov are very tensed and sometimes flirt with the limits of tonal hierarchy. For these preludes, we see a significant increase of the consistency with the circle of fifths by using the Multi-Probe Histograms. Indeed, the tone scale information is not sufficient anymore, and harmony of the pieces is much better characterized by a restricted set of specific intervals. Moreover, the harmony functions of some chords that might be used to go from a tonality to another are reflected in the histograms.

We also note that the similarity between minor preludes tends to be a bit lower than the similarity between major preludes. Also, similarity between preludes of different modes is much less consistent with the circle of fifths.

The good performances obtained for the preludes classification task validate the interpretation of Multi-Probe Histograms as harmony and tonal context descriptors.

5.4 Application to Structure Discovery

We have seen how the global context of a tonality can be described with the concatenation of the chroma sequences in Multi-Probe Histograms. The tonal context however also varies within music pieces. For instant, a change of scale is quite common between the verse and chorus parts of popular music songs. In this section, we are thus interested in characterizing these scale variations within pieces by means of MPH, the final goal being to enhance the characterization of the structure information.

5.4.1 Musical Sections Modeling

To illustrate the relevance of modeling the tonal context of structural sections for their segmentation, we select a popular music song example that is *Help* by the Beatles. The song contains two main sections, a verse and a chorus, that last about 20 seconds and are each repeated three times. The verse is composed in A major and consists of the following chords sequence that is repeated twice:

$$A \quad | \; C\#m \quad | \; F\#m \quad | \; D \; G \; A$$

In contrast, the chorus section can be interpreted both in A major and D major. Indeed, its chord sequence is composed of chords of the two keys:

$$Bm \quad | \; Bm \quad /A \; | \; G \quad | \; G \quad /F\# \; | \; E^7 \quad | \; E^7 \quad | \; A^7 \quad | \; A$$

We isolate in the audio file each occurrence of these two sections and extract their corresponding chroma features sequences. A Multi-Probe Histogram is then extracted over each chroma sequence. The result is displayed in figure 5.12.

There are two very distinct patterns formed in the histograms depending on wether they have been extracted on a chorus section or on a verse section. The tonal contexts of the two sections that we have just described is thus coherently revealed by the MPHs. This shows how highly discriminative these feature can be for the task of music structure segmentation.

5.4.1.1 Properties for MSD

Now that we have illustrated how sections can be modeled by means of MPH, we study two important properties of the histograms for the task of music structure segmentation: the robustness to tempo variations and the robustness to the extraction time-scale of the feature.

Robustness to Tempo Variation A good property of a structure detection algorithm is that it should be robust to tempo variations within the music piece. This means that if a section is played twice at two different tempi, the two segments should still be detected and clustered as the same section. To show how the MPH representation of chroma features would handle such cases, we consider the first

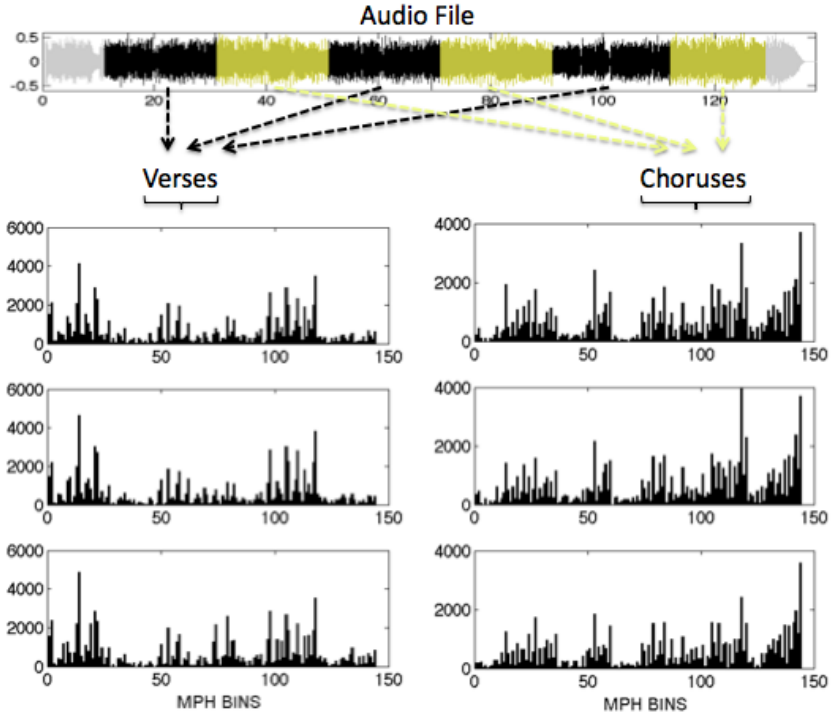
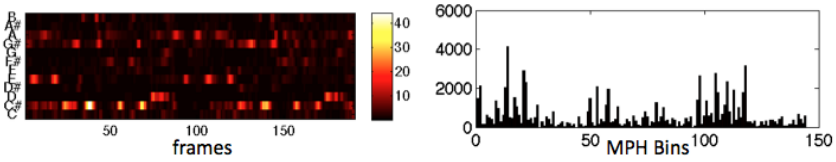


Figure 5.12: Histograms computed on chroma sequences of verses and choruses from the song Help from The Beatles

(a) Original Section (96 bpm)



(b) Section played 1.5 times faster (144 bpm)

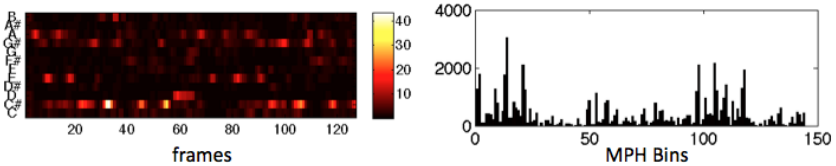


Figure 5.13: Chroma sequences and corresponding MPH's for a section played at two different tempi

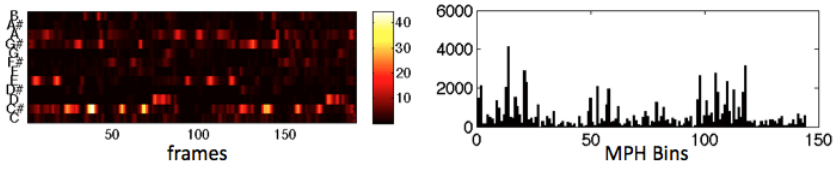
verse section of our music piece example. The chroma features are extracted over the original section, and over the same section played at 1.5 time its original speed. Note that the feature sampling rate remains unchanged. Chroma sequences and the obtained MPH's are shown in figure 5.13.

Except for the bins range, the histogram pattern for this section remains completely unchanged in the two extracted MPH. Handling tempo variations might require an alignment procedure in many MIR processing, but the fact that histograms are determined by local pitch class transitions independently of their sequence prevents from this.

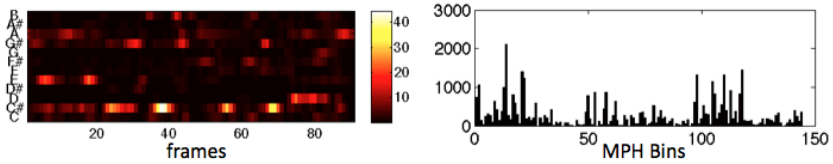
Robustness to Extraction Time-scale As seen in [Cohen 1991], the establishment of tonality in a music composition is quite fast. We are thus interested in studying whether the histogram pattern of a section remains the same if we consider the chroma sequence of the whole section, the first half, or the last third of the section. This is of great interest for music content description. For instance, the temporal segmentation step might over-segment a section but not its repetition. In that case the description should be sufficiently robust to extract salient properties even if not extracted over segments of the same length.

In figure 5.14, the chroma sequences of the three versions of the section and their corresponding MPH's are shown.

(a) Entire Section



(b) First half of the section



(c) Last third of the section

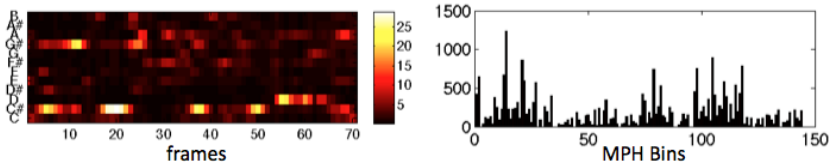


Figure 5.14: Chroma sequences and corresponding MPH's for various portions of the same section

Dominant peaks of the histogram pattern remain unchanged, confirming that the tonal context is quite homogeneous over the section. This suggests that our

approach will handle quite well over-segmentation issues.

This illustrates how musical sections can be modeled using MPH. A larger statistical evaluation of this result in the framework of Music Structure Segmentation is provided in section 5.5.2 and Chapter 6.

5.4.2 MPH as a mid-level audio feature

To allow for the description of tonal context variations within music pieces and benefit from the properties seen above, we apply the Multi-Probe Histograms as mid-level audio features. This means that the whole chroma sequence of a given audio file is now divided into overlapping mid-term chroma sequences that are all summarized in a MPH as illustrated in Figure 5.15.

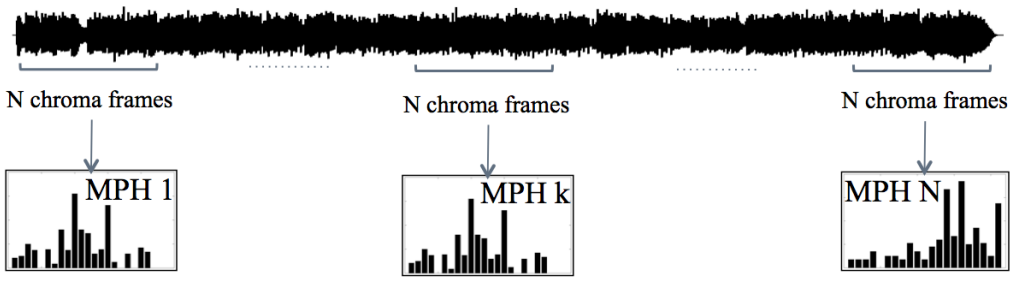


Figure 5.15: Using the Multi-Probe Histograms as a mid-level feature

Embedding the features in the audio similarity matrix, the comparison of two time instants of the audio signal now measures their tonal context similarity rather than the similarity of two isolated note or musical events.

5.4.2.1 The Mazurka Example

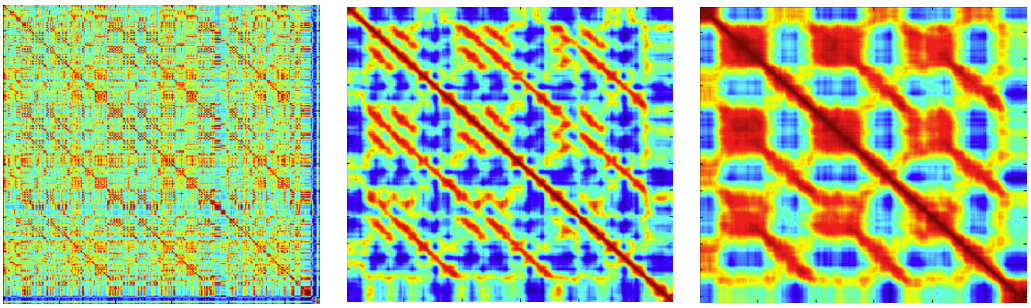
We show a concrete example in Figure 5.17. The audio excerpt is the 30 seconds excerpt of Chopin's *Mazurka, Op. 63, No. 3* whose score was already shown in Figure 5.1 and in which a tonality change from a B flat minor to a C Sharp minor occurs. In Figure 5.17.a, the standard similarity matrix as proposed in [Foote 1999] is computed on the chroma features extracted with a sampling rate of 4Hz. In Figure 5.17.b, we compute our proposed similarity matrix with MPH's computed over chroma sequences of length $N = 8$ frames, which corresponds to 2 seconds. Similarity matrices extracted with sequences of 16 and 40 frames are shown in Figures 5.17.d and 5.17.f respectively. For comparison we also compute the similarity matrices using the mean value of the chroma sequences as mid-level feature in Figures 5.17.c and 5.17.e.

The similarity matrix extracted on the original chroma vectors illustrates quite well the motivation of this chapter. While the scales of B flat minor and C Sharp minor contain similar tones, the distinction between the two sections in this visualization is very confused. In contrast, including more contextual information with

the Multi-Probe Histograms as well as with the mean chroma values significantly reduces this confusion. In our example, MPH and mean chroma values characterize the tonal context of the two sections in a similar manner. Indeed, the visual patterns within these sections are not drastically distinct. On the other hand, the dissimilarity between the two sections is much better characterized using the Multi-Probe Histograms.

5.4.2.2 The Beatles Example

If we now consider the song *Help* by the Beatles and extract similarity matrices computed over the chroma features and on histograms extracted on sequences of 16 and 40 frames, we again observe a clear enhancement of the structure representation. Indeed, boundaries between sections in the original similarity matrix are unclear. Using sequences of 16 frames, i.e. 4 seconds, for the MPH, the similarity matrix reveals that the verse is composed of a repeated chord sequence. Extending the length of sequences to 40 frames, i.e. 10 seconds, the tonal homogeneity is further characterized and the verse parts are displayed as blocks of high similarity, i.e. a state in the matrix. In contrast, the chorus part is made of a clear harmonic progression in its chords sequence as well as in the singing voice, and therefore does not have sufficient tonal self-similarity to produce a clear state in the matrix. However, its tonal dissimilarity with the verse is sufficiently characterized and the boundaries between sections remains clear.



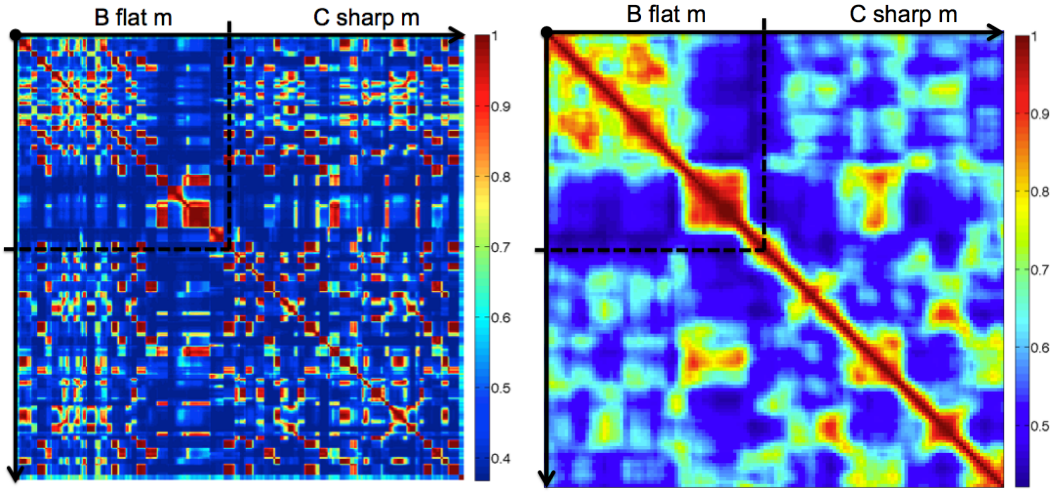
(a) Original Similarity Matrix

(b) MPH Similarity Matrix, N = 16 frames

(c) MPH Similarity Matrix, N = 40 frames

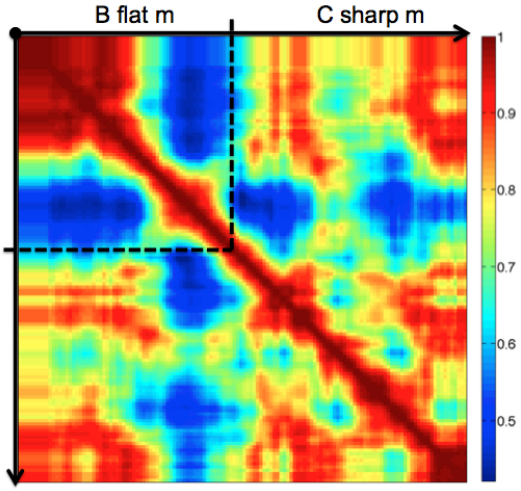
Figure 5.16: Strengthening the state representation by means of the MPH description of chroma sequences

Hence, embedding the histograms as mid-level audio features in the similarity matrices we tend to favor a representation of structure that fits the state assumption. We therefore now apply such matrices to our proposed NSMF music structure segmentation approach.

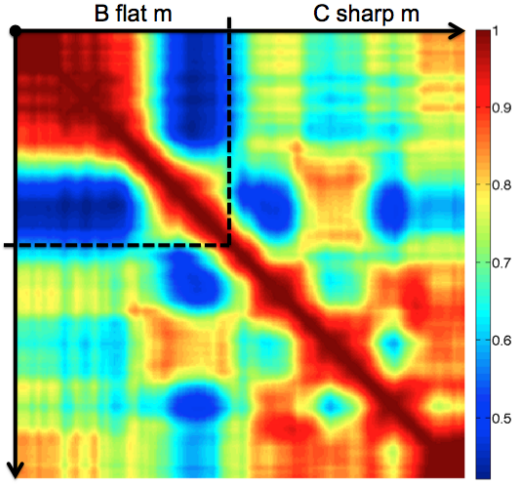


(a) Similarity Matrix with Chroma Vectors

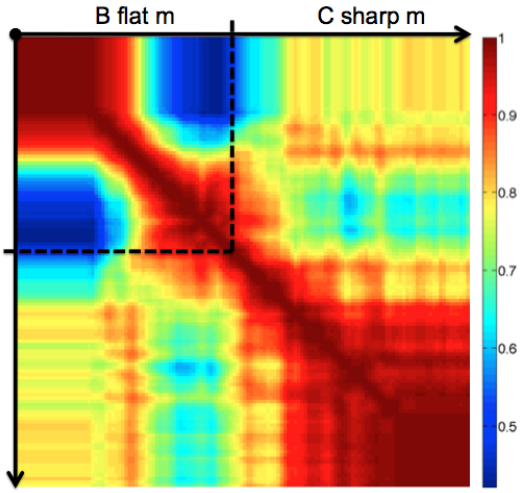
(b) MPH Similarity Matrix, N = 8 frames



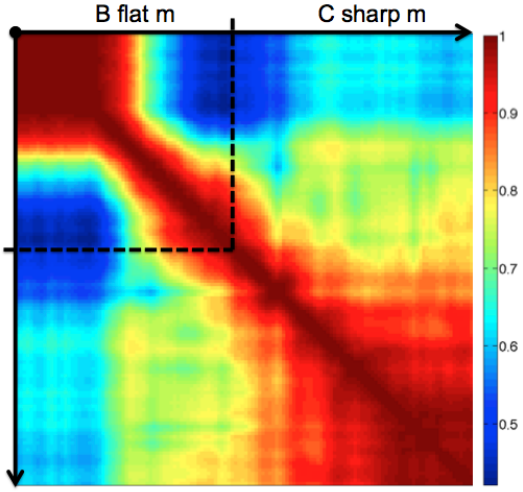
(c) Mean Chroma Vectors, N = 16 frames



(d) MPH Similarity Matrix, N = 16 frames



(e) Mean Chroma Vectors, N = 40 frames



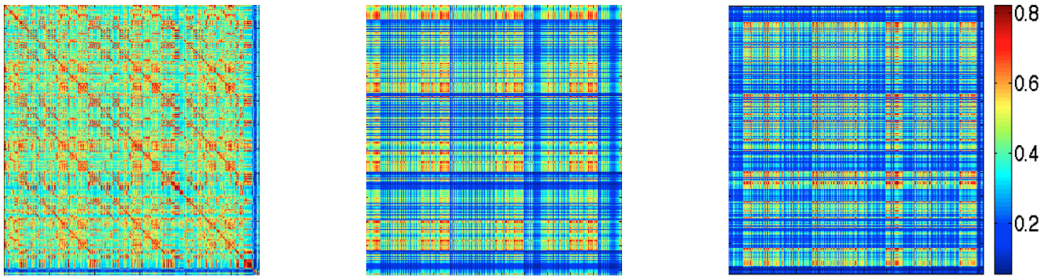
(f) MPH Similarity Matrix, N = 40 frames

Figure 5.17: Similarity matrices computed over a portion of the *Mazurka, Op. 63, No. 3*. Transition between B flat minor and C Sharp minor.

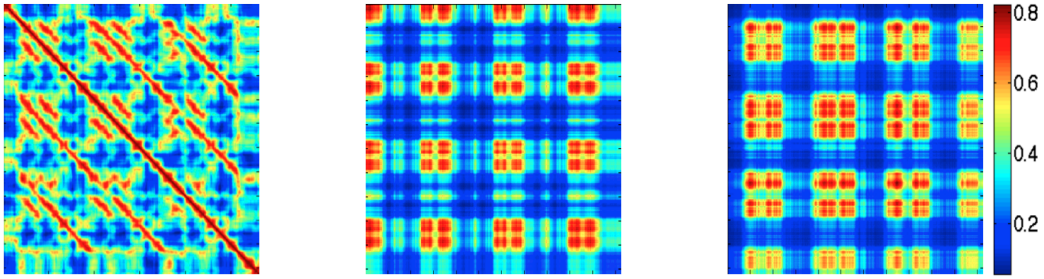
5.5 MPH and NSMF

5.5.1 Matrices Decompostion

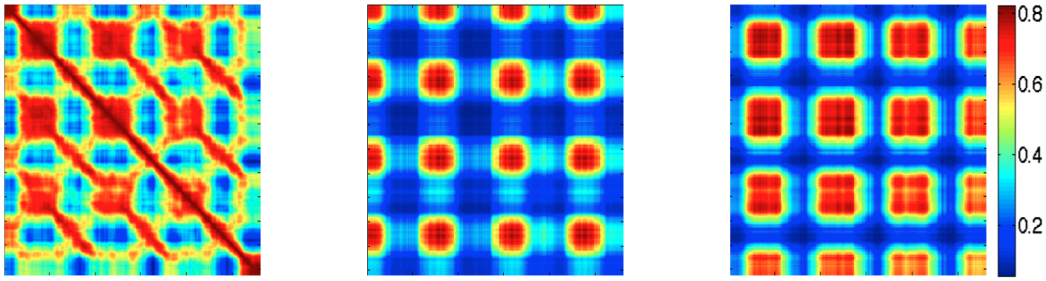
As seen in Chapter 3, the decomposition of similarity matrices by means of NMF is relevant for structural analysis only if the acoustical homogeneity of sections is characterized enough to yield sparse state representations. Prior to evaluating over a larger data set the benefit of using MPH as mid-level features for the NSMF approach, we give an illustration of the improvement in terms of sparsity of the decompositions that we obtain for the song example *Help* in Figure 5.16.



(a) Original Chroma Similarity Matrix and its 2^{nd} order symmetric NMF Decomposition



(b) MPH similarity matrix (N = 16 frames) and its 2^{nd} order symmetric NMF Decomposition



(c) MPH similarity matrix (N = 40 frames) and its 2^{nd} order symmetric NMF Decomposition

Figure 5.18: Similarity matrices and NMF decompositions for the song *Help* by the Beatles

Decomposition of the chroma similarity matrix for this song does not yield rele-

vant structural information. In that particular case, low-level features that relate to timbre information seemed to perform significantly better. However, the decompositions of the two MPH-based similarity matrices are significantly more consistent with the structure of the song, segregating the chorus and verse parts over the two dimensions of the NMF. Though states are not properly defined in the matrix that uses chroma sequences of length N = 16 frames, the decomposition seems to coherently interpret the matrix. This suggests that characterizing the homogeneity in dissimilarity between sections can cope with the lack of inner-similarity of sections.

5.5.2 Preliminary Evaluation

The combination of Multi-Probe Histograms similarity matrices with the NSMF approach is preliminarily evaluated on the *TUT Beatles* dataset that we remind consists of 175 songs by the Beatles. A further comparative evaluation on an extended dataset is proposed in Chapter 6.

We compare the performance of our system to the performance of the NSMF algorithm as described in Chapter 3 applied on chroma and timbre-related audio features and to the performance of the system proposed in [Mauch 2009] that won the MIREX[1] music structure segmentation evaluation task in 2009 for the same dataset. Description of evaluation measures can be found in Chapter 2.5. Results are shown in Table 5.5.

	MPH features with NSMF	Chroma features with NSMF	Timbre Features with NSMF	System proposed in [Mauch 2009]
F	63.3%	60.8%	61.0%	60.0 %
P	59.3%	61.5%	63.3%	56.1%
R	72.4%	64.6%	62.4%	71.0%
So	68.3%	61%	62.1%	73,9%
Su	58.8%	59.9%	60.7%	61.7%

Table 5.5: Evaluation of the proposed approach and comparison with the state-of-the-art

The general increase in the F-measure is of 2 to 3% compared to the other systems. While the algorithm seems to behave in a similar manner as in [Mauch 2009] (comparable Precision and Recall rates), the nature of the structure segmentation changes using MPHs instead of raw chroma vectors for the similarity matrix computation. Indeed, introduction of the Multi-Probe Histograms to the NSMF algorithm increases the recall rate of 8 to 10% with a reasonable loss in precision, i.e. -2% for chroma features and -4% for timbre features, meaning that our approach copes significantly better with over-segmentation issues. Over-segmentation is indeed a

[1] http://www.music-ir.org/mirex/wiki/2009:Music_Structure_Segmentation_Results

recurrent problem in structure segmentation because of the inner structure of musical sections. While this structure is reflected in the estimated segmentation, it does not match the hierarchy level of the annotated structure which is rather high-level. Using mid-level features that tend to match this definition of structure, we increase the performance.

5.6 Chapter Summary

Highlighting the lack of coincidence between the level of musical structures defined for the task of MSD and the level of description of low-level audio features, we have proposed in this chapter an efficient method to characterize the tonal context of a musical segment by concatenating its chroma features sequence in a Multi-Probe Histogram. We have shown with the preludes classification task that the histograms strongly correlate with the tonal structure implied by a tonality. Embedding this representation as mid-level feature, we were then able to compute similarity matrices in which the tonal homogeneity of structural sections is much better characterized and therefore fit the description of structural sections as states.

Preliminary evaluation of the combination of NSMF with such matrices showed very encouraging results. Structural information in the *TUT Beatles* is however characterized by harmony changes as well as by instrumentation changes. While Multi-Probe Histograms do not convey any information regarding timbre, this dataset is thus rather limited for properly evaluating the real impact of using MPH for the task of structure segmentation. In the next chapter, we will therefore run a comparative evaluation on mono instrumental classical music pieces, in which structural boundaries are only determined by melody and/or harmony variations.

Chapter 6

Comparative Evaluation

Contents

We now present a comparative evaluation of the music structure segmentation algorithms proposed in this PhD thesis. The evaluation database is described in section 6.1 and consists of 145 songs by the Beatles, a catalogue of 100 popular music songs, and 50 classical music pieces of the RWC corpus . As reminded in section 6.2, we run the NSMF algorithm as described in Chapter 3, the NSMF algorithm with enhanced similarity matrices proposed in Chapter 4 and finally the NSMF taking MPH similarity matrices as input (Chapter 5). Evaluation of the

temporal segmentation is provided in section 6.3. Comparative evaluation of the structure segmentation for the TUT Beatles, RWC Pop and RWC Classic are then presented in sections 6.4, 6.5 and 6.6 respectively.

6.1 Database Description

Music structure discovery strongly lacks from available structured annotated data for its evaluation. This is mainly due to the weak definition of the task that makes the standardization of the annotation procedure tricky. We however selected three data sets within the yet produced annotations that seem today 'quite' consensual. These were indeed used as reference annotations in the past MIREX[1] evaluation campaigns for structural segmentation. Note that annotation procedure is a rising discussion within the community and that evolution of databases should be carefully watched in the future.

6.1.1 Datasets

The first corpus we use is the the *TUT Beatles*[2] dataset that served for the preliminary evaluation in Chapters 3, 4 and 5. Annotations provided for this set were reviewed by several institutions and are quite consensual. Also, most methods in the literature being evaluated on this dataset, it allows us to situate the performance of our approach within the state of the art.

In order to provide a deeper characterization of our algorithms, we also evaluate on two sets of the RWC corpus[3]: the *RWC Pop* dataset, that consists of japanese popular music songs, and the *RWC Classic* dataset that consists of classical music pieces and allows us to focus on the harmony aspects of structure in our evaluation. Structural annotations within the RWC corpus are not ideal for our task and were initially designed for chorus detection. Therefore in some annotations of the *RWC Classic*, choruses are indicated whereas other sections are labeled as "nothing". We will keep this incomplete structural annotation in mind for the results interpretation.

We select within the *RWC Classic* dataset a subset of 28 mono-instrumental music pieces. This way we ensure that the studied musical structures are only influenced by harmony and melody variations, independently of instrumentation changes.

[1] http://www.music-ir.org/mirex/wiki/MIREX_HOME
[2] http://www.cs.tut.fi/sgn/arg/paulus/structure.html
[3] http://staff.aist.go.jp/m.goto/RWC-MDB/

Start (*s*)	End (*s*)	Label
0.0000000	1.0500000	silence
1.0500000	12.5514101	intro
12.5514101	36.0160926	verse
36.0160926	47.4230166	intro
47.4230166	70.8088987	verse
70.8088987	76.5517608	refrain
76.5517608	87.9782804	intro
87.9782804	111.2457530	verse
111.2457530	117.0478199	refrain
117.0478199	122.6722726	introH
122.6722726	145.7917332	verseS
145.7917332	151.5049930	introH
151.5049930	175.0092847	verse
175.0092847	180.8409540	refrain
180.8409540	192.2082688	intro
192.2082688	260.6133333	outro

(a) *Come Together* - TUT Beatles

Start (*sample*)	End (*sample*)	Label
0	2556	intro
2556	4389	verse A
4389	6110	verse B
6110	7775	verse C
7775	9552	chorus A
9552	9663	post-chorus
9663	11495	verse A
11495	13216	verse B
13216	14882	verse C
14882	16659	chorus A
16659	18435	bridge A
18435	20212	bridge B
20212	21989	verse C
21989	23765	chorus A
23765	25542	chorus A
25542	28497	ending

(b) RM-P004 - RWC Pop

Table 6.1: Annotation examples

To summarize, our database is composed of the three following datasets:

a) **TUT Beatles**: 175 songs by the Beatles
b) **RWC Pop** : 100 popular music songs
c) **RWC Classic** : 28 classical mono-instrumental music pieces

A complete list of pieces that compose the three datasets can be found in Appendix B. Boundaries and labels of sections are annotated for each music piece as in the examples in Table 6.1.

6.1.2 Original and Conflated Ground Truth

As already mentioned, there is no standardized definition of structure nor annotation procedure for the task of music structure segmentation. Hence, the produced annotations often match the definition of structure retained by their authors, or were made to evaluate a precise task (structure segmentation, chorus detection, repetition detection, etc.). To allow for some margin in the interpretation of these annotations, Smith proposed in [Smith 2010] to distinguish between the original and conflated ground truths. This was thought to handle some annotations in which annotators distinguish sections repetitions, i.e. *Chorus A* and *Chorus B*, even if the repeated section is not altered or modified in any sense, or when the inner structure of sections is also annotated. In the conflated segmentation, these two sections are considered identical and simply labeled as *Chorus*.

The incidence on the average number of sections, average length of sections and average number of labels for our three datasets is shown in Tables 6.2, 6.3 and 6.4 respectively.

	Original ground truth	Conflated ground truth
TUT Beatles	8.55	8.32
RWC Pop	16.11	11.22
RWC Classic Solo	8.29	5.97

Table 6.2: Average number of segments

	Original ground truth	Conflated ground truth
TUT Beatles	21.87	22.22
RWC Pop	15.51	23.88
RWC Classic Solo	29.74	41.21

Table 6.3: Average length of segments

	Original ground truth	Conflated ground truth
TUT Beatles	5.28	5.12
RWC Pop	9.15	5.81
RWC Classic Solo	5.70	3.76

Table 6.4: Average number of labels

Deviations between the original and conflated annotations for the *TUT Beatles* dataset are not sufficiently significant for distinguishing them in our evaluation. In contrast, it strongly modifies the nature of the annotated structure for the RWC datasets, reducing of about one third the average number of segments and average number of labels, and improving the average length of segments.

6.2 Algorithms

Algorithms evaluated in this chapter were all presented in the chapters of this thesis and are labelled in the evaluation as *Algo 1* for the NSMF algorithm, *Algo 2* for the NSMF plus matrix enhancement procedure, and finally *Algo 3*, *Algo 3 bis*, and *Algo 4* for the Multi-Probe Histograms based approaches. A brief description of these algorithms is reminded in this section.

6.2.1 Algo 1

The algorithm labeled as *Algo 1* is the NSMF algorithm proposed in Chapter 3. Similarity matrices computed on either low-level timbral features, i.e MFCC and Spectral Moments, or chroma features are decomposed by means of Non-negative Matrix Factorization. The decomposed basis vectors are then used as intermediate features for the structure hierarchical clustering. The rank of decomposition for the NMF is set to 9.

6.2.2 Algo 2

The algorithm labeled as *Algo 2* is the extended version of the NSMF algorithm where the similarity matrices are previously processed by means of image processing techniques as introduced in Chapter 4. Matrices are first low-pass filtered with anisotropic diffusion, binarization with Otsu thresholding method is applied and morphological finally use pixel connectivity to enhance the structural states in the matrix.

6.2.3 Algo 3 and Algo 3bis

In the algorithms labeled as *Algo 3* and *Algo 3bis* introduced in Chapter 5, NSMF segmentation is applied on similarity matrices computed by means of Multi-Probe Histograms applied as mid-level features. *Algo 3* and *Algo 3bis* differ in the length of the chroma sequences N for the computation of histograms. For *Algo 3*, N was chosen as 16 frames, i.e. 4 seconds, while sequences of 40 frames, i.e. 10 seconds, were used in *Algo 3bis*.

6.2.4 Algo 4

The algorithm *Algo 4* also makes use of MPHs but not of NSMF. MPHs are computed between boundaries of the segments found by means of the novelty function. Distance between the MPHs extracted on segments is then computed and agglomerative hierarchical clustering is applied to generate the structural segmentation.

6.3 Temporal Segmentation Evaluation

In the NSMF algorithm we apply for the temporal segmentation of music pieces the audio novelty score approach introduced in [Foote 2000] and described in Chapter 2 and Chapter 3. With this method, change points in the audio signal are detected in the similarity matrix by means of a 2D checkerboard that models an ideal shaped boundary. While the three versions of NSMF we evaluate in this chapter are based on different types of similarity matrices, we evaluate how the audio novelty approach performs for each of them. In Table 6.5, Tables 6.6 and 6.7 and Tables 6.8 and 6.9, evaluation by means of the second order F-Measure, Precision and Recall (*cf* section 2.5) is provided for the three datasets individually.

Algo 1 is run on timbre-related low-level audio features for the *TUT Beatles* and *RWC Pop* datasets and on chroma features for the *RWC Classic* dataset. In *Algo 2*, we apply the audio novelty score to the same matrices but processed by our matrix enhancement procedure proposed in Chapter 4. And finally, MPH similarity matrices extracted with sequences of N = 16 frames, i.e. 4 seconds, were used in *Algo 3*. Segmentation is evaluated for a boundary deviation tolerance that varies from 0.5 seconds to 3 seconds.

6.3.1 Results

TUT Beatles Dataset

	F-measure @0.5s	F-measure @3s	Precision @0.5s	Precision @3s	Recall @0.5s	Recall @3s
Algo 1	17.0%	54.5%	14.0%	45.0%	18.2%	59.2%
Algo 2	18.0%	59.6%	13.5%	43.0%	21.0%	67.8%
Algo 3	10.0%	43.8%	6.6%	31.8%	10.0%	49.7%

Table 6.5: Segmentation evaluation for the ***TUT Beatles*** dataset with the **original annotation** as reference

RWC Pop Dataset

	F-measure @0.5s	F-measure @3s	Precision @0.5s	Precision @3s	Recall @0.5s	Recall @3s
Algo 1	7.8%	49.2%	9.2%	58.3%	7.7%	48.3%
Algo 2	7.2%	45.2%	8.9%	55.6%	7.1%	44.2%
Algo 3	7.1%	43.6%	6.9%	43.7%	7.2%	44.5%

Table 6.6: Segmentation evaluation for the ***RWC Pop*** dataset with the **original annotation** as reference

	F-measure @0.5s	F-measure @3s	Precision @0.5s	Precision @3s	Recall @0.5s	Recall @3s
Algo 1	8.7%	54.3%	7.3%	45.9%	9.4%	58.4%
Algo 2	8.2%	50.2%	7.1%	44.0%	8.8%	53.6%
Algo 3	7.0%	42.2%	4.9%	29.9%	8.1%	48.3%

Table 6.7: Segmentation evaluation for the ***RWC Pop*** dataset with the **conflated annotation** as reference

RWC Classic Dataset

	F-measure @0.5s	F-measure @3s	Precision @0.5s	Precision @3s	Recall @0.5s	Recall @3s
Algo 1	3.5%	34%	2.2%	21.8%	4.5%	45.4%
Algo 2	4.9%	32%	3.7%	20.2%	5.9%	41.8%
Algo 3	7.0%	33.2%	4.8%	22.3%	8.7%	42.2%

Table 6.8: Segmentation evaluation for the ***RWC Classic*** dataset with the **original annotation** as reference

	F-measure @0.5s	F-measure @3s	Precision @0.5s	Precision @3s	Recall @0.5s	Recall @3s
Algo 1	3.1%	29.0%	1.6%	14.5%	4.7%	46.0%
Algo 2	3.7%	26.2%	2.0%	12.8%	5.1%	40.9%
Algo 3	5.0%	30.8%	2.5%	15.4%	7.6%	46.6%

Table 6.9: Segmentation evaluation for the ***RWC Classic*** dataset with the **conflated annotation** as reference

6.3.2 Discussion

The segmentation evaluation obviously reflects that the boundaries retrieved by means of the audio novelty score are not precisely scaled with the annotated boundaries. The highest performance is obtained on the *TUT Beatles* dataset with a F-measure of 59.6% for *Algo 2*.

For the *TUT Beatles* and *RWC Classic* datasets, the recall rates are obviously higher than the precision rates. This means that the number of false positive retrieved boundaries is significantly higher than the number of false negative, suggesting that the algorithm tends to over-segment the music pieces for these datasets.

It is interesting to note the change in the nature of our segmentation with regard to the conflated and original boundaries for the *RWC Pop* dataset. While the

recall rates decrease from the original to the conflated evaluation, precision for the original segmentation is significantly higher, i.e. 45.9% for the conflated annotation and 58.3% for the original segmentation with *Algo 1* for *RWC Pop*. This means that after removing boundaries in the conflated segmentation, the number of false negative boundaries decreased while the number of false positive increased. Hence, the scale of the estimated boundaries seems to rather fit the original segmentation for this dataset. Moreover, precision rates are higher than recall rates when evaluating with the original annotation, suggesting that the boundaries deviate from the annotation but do not necessarily strongly over-segment the music pieces. This is illustrated in Figure 6.1 with the segmentation result for the file *RM-P029* of RWC Pop.

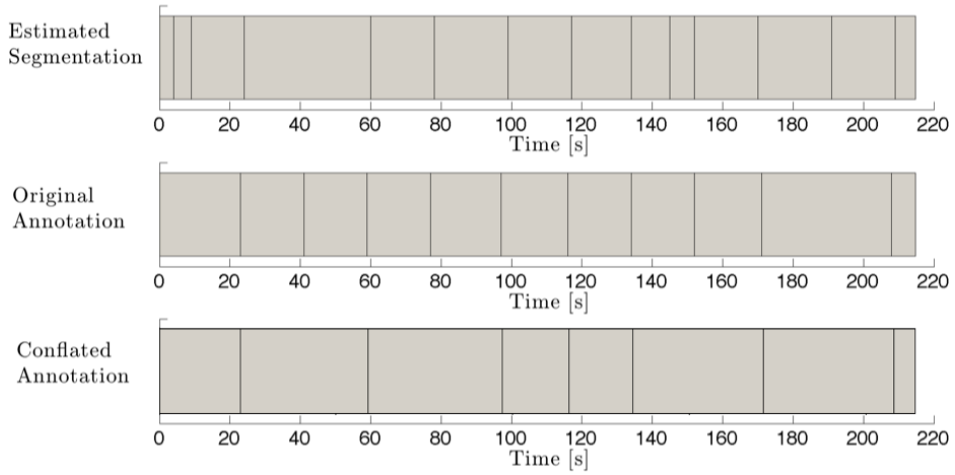


Figure 6.1: Segmentation for the file RM-P029 of the RWC Pop dataset

For the *TUT Beatles* and *RWC Pop*, the performance obtained with boundaries retrieved on similarity matrices computed on Multi-Probe Histograms (*Algo 3*) is always behind the two other algorithms. As a matter of fact, introducing contextual information in the audio description induces a smoothing effect on the similarity blocks edges and one looses precision in the description of change points in the audio signal. As a result, we can not see any sharp edges between states in MPH-based similarity matrices and the audio novelty approach is thus not very adapted. In the forthcoming evaluation, the temporal segmentation of classical similarity matrices will be used for the evaluation of the MPH-based algorithm for these two sets.

Ground Truth vs. Estimated Segmentation To allow for the evaluation of the clustering part independently of the effect of the deviation between estimated and annotated boundaries, labeling performance of the proposed algorithms is evaluated in the remainder of this section using both the estimated and annotated boundaries as temporal segmentation.

6.4 TUT Beatles Data Set

Evaluation of our algorithms on the *TUT Beatles* data set was already provided in each chapter. We however summarize these results in Table 6.10. The number of clusters to form for the algorithms is set to 4.

	F-Measure	Precision	Recall	So	Su
Algo 1 - timbre	61.0%	63.3%	62.4%	62.1%	60.7%
Algo 1 - chroma	60.8%	61.5%	64.6%	61.0%	59.9%
Algo 1 - fusion	62.1%	63.6%	64.5%	62.0%	60.0%
Algo 2 - timbre	62.9%	60.2%	70.0%	66.5%	59.6%
Algo 2 - chroma	62.3%	59.1%	70.0%	66.2%	58.6%
Algo 2 - fusion	62.8%	59.4%	70.8%	67.1%	58.9%
Algo 3	63.3%	59.3%	72.4%	68.3%	58.8%

Table 6.10: Evaluation on the ***TUT Beatles*** data set with **estimated temporal segmentation**

Despite the rather low performance of the temporal segmentation, the NSMF based approaches proposed in this thesis achieve a F-measure of up to 63%, showing slightly better performances than the state of the art (see section 5.5.2 of Chapter 5). Over-segmentation of the estimated boundaries thus seems to be compensated at the clustering step. Introduction of matrix filtering and tonal context modeling in algorithms 2 and 3 allows to characterize a higher level of structure and therefore compensate even more temporal over-segmentation. This is reflected in the strong increase in the recall rates for *Algo 2* and *Algo 3*, accompanied by a slight reduction of precision rates. Interestingly, structural information for this dataset is contained in both instrumentation and harmony. Performance of Algorithms 1 and 2 using chroma and timbre features are indeed very comparable. Moreover, the MPH similarity matrices achieve the best performance suggesting that tonal variations between verse and chorus are quite recurrent in the Beatles' compositions.

	F-Measure	Precision	Recall	So	Su
Algo 1 - timbre	78.4%	82.1%	78.3%	78.0%	87.0%
Algo 1 - chroma	76.6%	81.2%	75.7%	78.0%	86.5%
Algo 1 - fusion	77.8%	82.3%	77.0%	78.2%	86.9%
Algo 2 - timbre	80.6%	78.2%	86.7%	87.8%	81.8%
Algo 2 - chroma	81.2%	77.2%	90.3%	90.8%	81.5%
Algo 2 - fusion	81.5%	78.4%	88.0%	89.4%	82.3%
Algo 3	83.3%	79.8%	91.5%	92.2%	84.2%

Table 6.11: Evaluation on the ***TUT Beatles*** data set with **annotated temporal segmentation**

In order to compare how the algorithms behave when the temporal segmentation fits the annotation, i.e. meaning that the clustering step is separately evaluated, we show the evaluation of our algorithms using the annotated temporal segmentation in Table 6.11. The performance hierarchy between the algorithms is kept and the MPH-based approach achieves the best F-measure. As in the previous evaluation, *Algo 1* shows higher precision rates and lower recall rates than the two other algorithms. Enhancing the state representation in the similarity matrices, the NSMF algorithm thus seems to have a slight tendency to under-segment songs.

Providing the algorithms with the boundaries information, the global performance is logically significantly higher than with the estimated boundaries. Once again, this highlights the fact that the segmentation evaluation is highly dependent on the accuracy of the boundary retrieval with regard with the annotation. There is however no absolute truth in annotating musical structures and thus no absolute evaluation of the task.

6.5 RWC Pop

In this section we report the evaluation of the first three algorithms on the *RWC Pop* data set. As structural boundaries are mainly induced by instrumentation changes in the songs of the dataset, further algorithms based on Multi-Probe Histograms (Algo 3 bis and 4) are not evaluated with this set. Algorithms 1 and 2 are used with timbre-related low-level features.

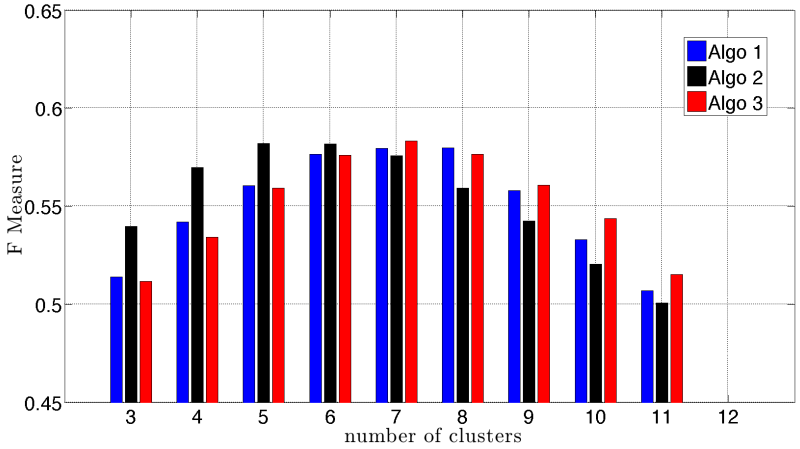
In order to analyze the effect of the chosen number of clusters on the algorithms' performance, we run several evaluations for different number of clusters. Both conflated and original segmentations are used as reference for the performance evaluation. Description of the used metrics can be found in section 2.5 of Chapter 2.

We first run the evaluation using the automatic temporal segmentation by means of the audio novelty score. Histograms representation of the obtained F-Measures is plotted in Figure 6.2.a for the conflated segmentation and in Figure 6.4.a for the original segmentation. Recall versus precision rates are plotted in Figure 6.2.b for the conflated segmentation and in Figure 6.4.b. for the original segmentation.

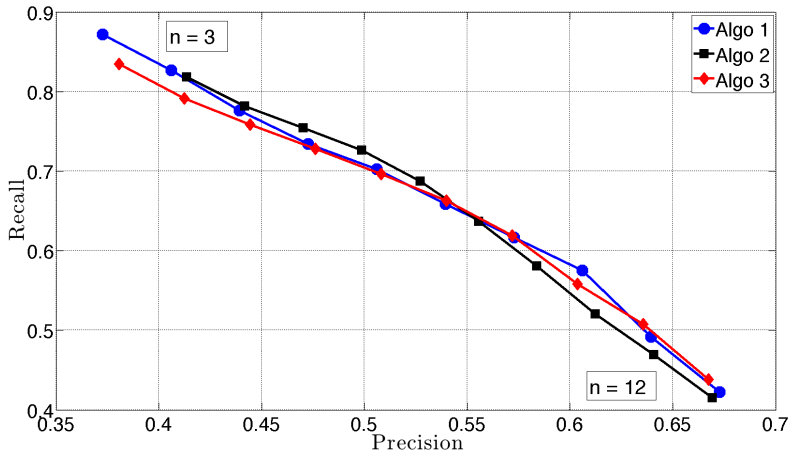
Secondly, the evaluation is run using the annotated boundaries. As previously noted this allows to compare the algorithms independently of the performance of the temporal segmentation. Histogram representation of the obtained F-Measures is plotted in Figure 6.3.a for the conflated segmentation and in Figure 6.5.a for the original segmentation. Recall versus precision rates are plotted in Figure 6.3.b for the conflated segmentation and in Figure 6.5.b. for the original segmentation.

For each evaluation, performance for the number of clusters that yields the best results is detailed in tables 6.12, 6.13, 6.14 and 6.15.

6.5.1 Evaluation with Conflated Groundtruth



(a) F-Measures

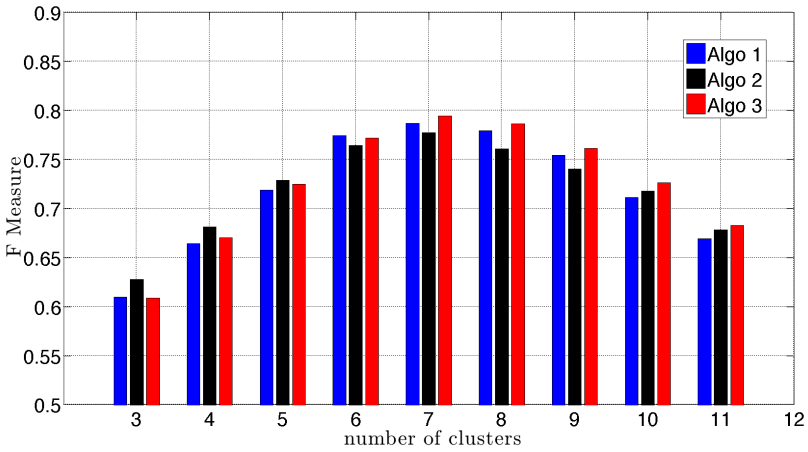


(b) Recall vs. Precision, n = number of clusters

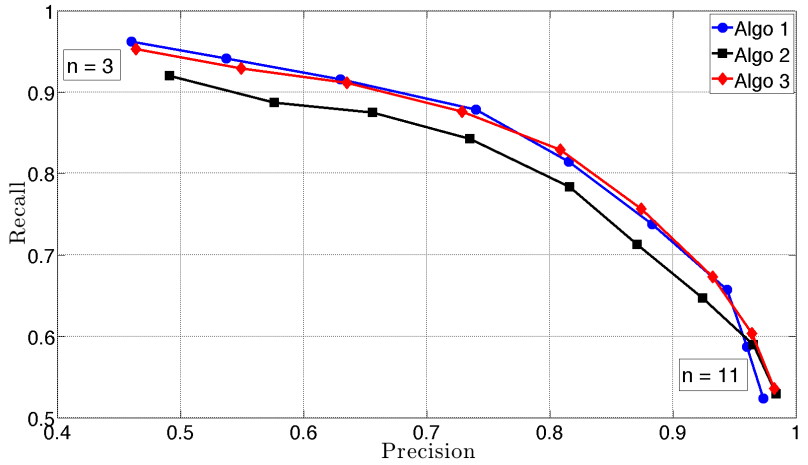
Figure 6.2: Evaluation on the ***RWC Pop*** with **automatic temporal segmentation** and **conflated ground truth**

	F-Measure	Precision	Recall	So	Su
Algo 1 - timbre	57.7%	51.4%	69.6%	70.1%	55.6%
Algo 2 - timbre	58.2%	54.7%	65.3%	67.6%	58.6%
Algo 3	57.6%	51.5%	69.0%	70.0%	55.9%

Table 6.12: ***RWC Pop*** - Estimated Segmentation - Conflated Ground Truth - Number of clusters n = 6



(a) F-Measures



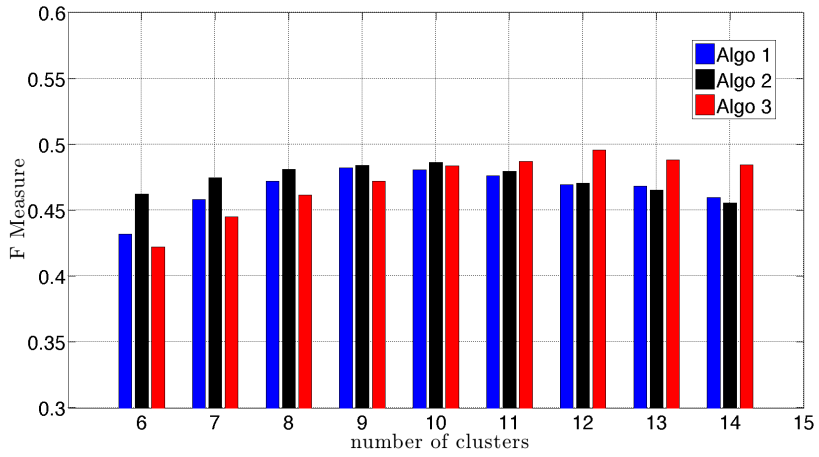
(b) Recall vs. Precision, n = number of clusters

Figure 6.3: Evaluation on the ***RWC Pop*** with **annotated temporal segmentation** and **conflated ground truth**

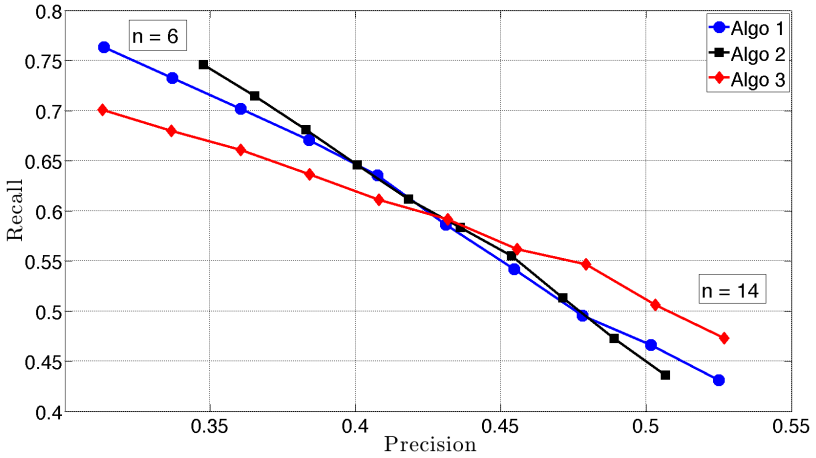
	F-Measure	Precision	Recall	So	Su
Algo 1 - timbre	78.7%	81.5%	81.4%	86.2%	87.2%
Algo 2 - timbre	77.8%	81.7%	78.3%	84.1%	86.8%
Algo 3	79.5%	80.9%	82.9%	87.2%	85.8%

Table 6.13: ***RWC Pop*** - Ground Truth Segmentation - Conflated Ground Truth - Number of clusters n = 7

6.5.2 Evaluation with Original Groundtruth



(a) F-Measures

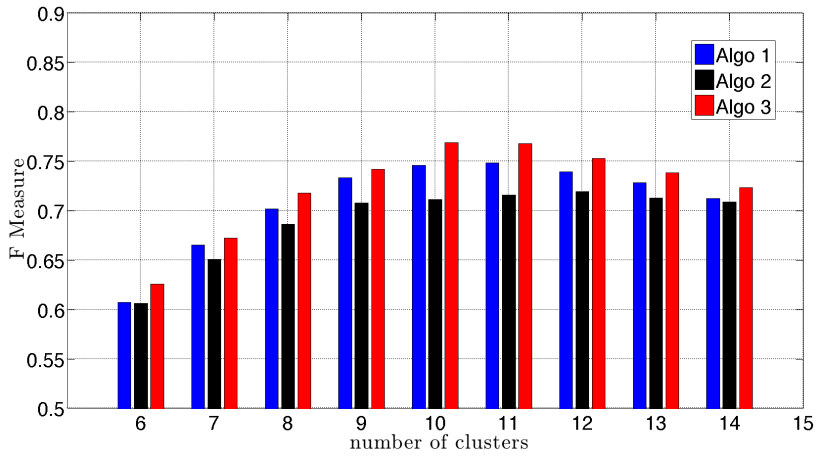


(b) Recall vs. Precision, n = number of clusters

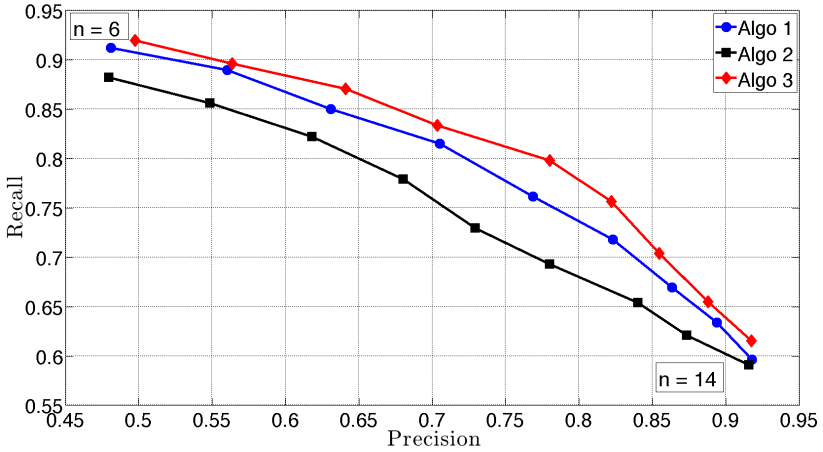
Figure 6.4: Evaluation on the ***RWC Pop*** with **automatic temporal segmentation** and **original ground truth**

	F-Measure	Precision	Recall	So	Su
Algo 1 - timbre	48.1%	43.6%	57.5%	70.1%	61.8%
Algo 2 - timbre	48.6%	45.3%	55.6%	69.2%	62.7%
Algo 3	48.4%	43.4%	58.9%	69.5%	60.6%

Table 6.14: ***RWC Pop*** - Estimated Segmentation - Original Ground Truth - Number of clusters n = 10



(a) F-Measures



(b) Recall vs. Precision, n = number of clusters

Figure 6.5: Evaluation on the ***RWC Pop*** with **annotated temporal segmentation** and **original ground truth**

	F-Measure	Precision	Recall	So	Su
Algo 1 - timbre	74.6%	76.9%	76.2%	87.2%	88.2%
Algo 2 - timbre	71.2%	72.9%	72.9%	85.2%	85.9%
Algo 3	76.9%	78.0%	79.8%	88.7%	88.8%

Table 6.15: ***RWC Pop*** - Ground Truth Segmentation - Original Ground Truth - Number of clusters n = 10

6.5.3 Discussion

6.5.3.1 Differences between Conflated and Original Groundtruth

Looking at the pairwise measures, the NSMF algorithms perform significantly better when evaluated with the conflated groundtruth. Indeed, the F-Measure difference reaches 9.6% for the image enhanced matrices (*Algo 2*). While the boundary retrieval evaluation indicated that the songs were a bit over-segmented with regard with the conflated segmentation for this set, the algorithms seem to have compensated by combining segments of the same conflated label at the clustering step as already suggested in the evaluation with the *TUT Beatles* dataset. As a matter of fact, performance of the system using the annotated boundaries is also higher when evaluated with the conflated segmentation and this result holds for all algorithms. Evaluation with the original segmentation implies a loss of 10% in both the recall and precision rates, once again suggesting that versions of the same section indicated in the original segmentation were not much discriminated by the algorithms.

Surprisingly, the over-segmentation and under-segmentation scores indicate contradictory results with the pairwise measures. Indeed, while the So scores remain relatively constant for each algorithm in both evaluations, the Su scores strongly increase when evaluating with the original segmentation. Using the annotated boundaries, we observe the same phenomena: precision and recall rates decrease from conflated to original evaluation while the So and Su scores remain in the same range and even slightly increase. Looking at the definition of the under- and over-segmentation scores in section 2.5, the only explanation we may propose for that is that the scores could be slighlty dependent on the number of labels. Indeed, conditional entropies are normalized with the logarithm two of the the number of annotated and estimated labels.

It is to be noted that the F-measures for this dataset are slightly under the one obtained with the *TUT Beatles* set. Listening to the songs of the RWC set we can however say that they are rather smooth in their structural changes. For instance background musical accompaniment often remains unchanged and played along the whole songs, making the acoustical modeling of sections tricky. Also, the actual interpretation of such evaluations is alway prone to the quality of the annotations.

6.5.3.2 Algorithms Comparison

All variants of the NSMF algorithms achieve rather comparable performances for this dataset and have the same relationship with the number of clusters. Setting a smaller number of clusters, *Algo 2* performs sensibly better than the other algorithms. Indeed, with the image-based matrix processing we simplify the structure representation in the similarity matrices and thus reduce the number of visualized states. The benefit of this technique is thus logically higher using a lower number of clusters. Moreover, when *Algo 2* is imposed the boundaries of the original segmentation that has smaller segments' size, it then performs less better than

the other algorithms. The effect of using this approach is however lower than observed with the *TUT Beatles* dataset and the method remains dependent on the pre-characterization of states with the low-level audio features. Interestingly, *Algo 2* tends to over-segment the pieces more than *Algo 1*. The contrary result was observed in the previous evaluation. Segmentation masks obtained on this dataset may thus have over-segmented the similarity matrices.

Though the harmonic changes in this dataset seemed less significant than in the Beatles set, the MPH-based approach (*Algo 3*) performs quite well and even outperforms the other algorithms with the original groundtruth using the annotated boundaries. Songs of the dataset do not always contain significant tonal changes in the musical accompaniment though. Averaging this content by means of MPH, *Algo 3* tends to under-segment the songs a bit more than the other algorithms as indicated by the evaluation. Note that for the evaluations reported in Tables 6.2 and 6.4 the boundaries retrieved on classical timbre-related similarity matrices were used. In general we observe a large variability in the performance of the system. Standard deviation of the F-Measures is indeed of about 10% and of 14% for the precision and recall rates.

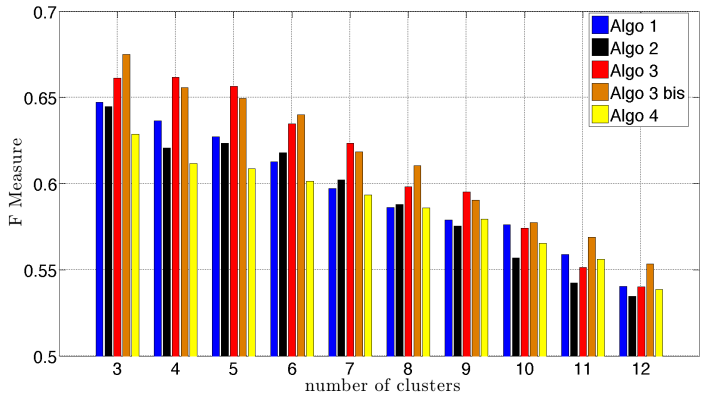
6.6 RWC Classic

The RWC Classic database is composed of 50 classical music pieces from that cover various music era (see details in Appendix B). We want to evaluate with this data set the benefit of modeling local harmonic properties by means of Multi-Probe Histograms. Therefore we consider a subset of the RWC Classic database of 28 mono-instrumental pieces, with a majority of pieces composed composed for piano solo. We ensure this way that the extraction of chroma features is robust and do not convey any timbre modulation information. All algorithms are evaluated on this subset of the RWC Classic. As timbre contain no relevant information for this set, algorithms 1 and 2 are evaluated using chroma features.

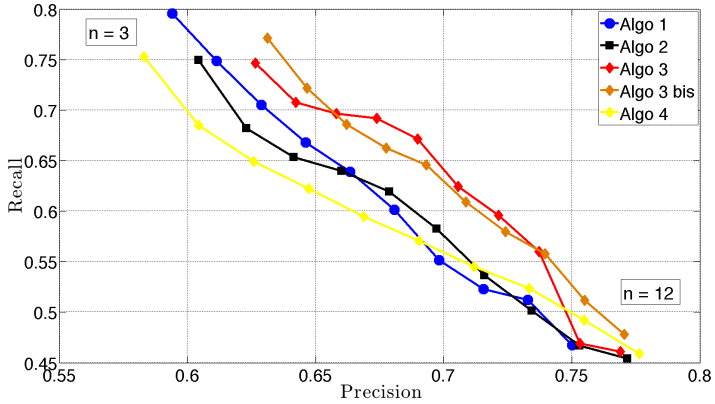
Just as in the last section, we first run the evaluation using the automatic temporal segmentation by means of the audio novelty score. Histograms representation of the obtained F-Measures is plotted in Figure 6.6.a for the conflated segmentation and in Figure 6.8.a for the original segmentation. Recall versus precision rates are plotted in Figure 6.6.b for the conflated segmentation and in Figure 6.8.b. for the original segmentation. Secondly, the evaluation is run using the annotated boundaries. As previously noted this allows to compare the algorithms independently of the performance of the temporal segmentation. Histogram representation of the obtained F-Measures is plotted in Figure 6.7.a for the conflated segmentation and in Figure 6.9.a for the original segmentation. Recall versus precision rates are plotted in Figure 6.7.b for the conflated segmentation and in Figure 6.9.b. for the original

segmentation. For each evaluation, performance for the number of clusters that yields the best results is detailed in tables 6.16, 6.17, 6.18 and 6.19.

6.6.1 Evaluation with Conflated Groundtruth



(a) F-Measures

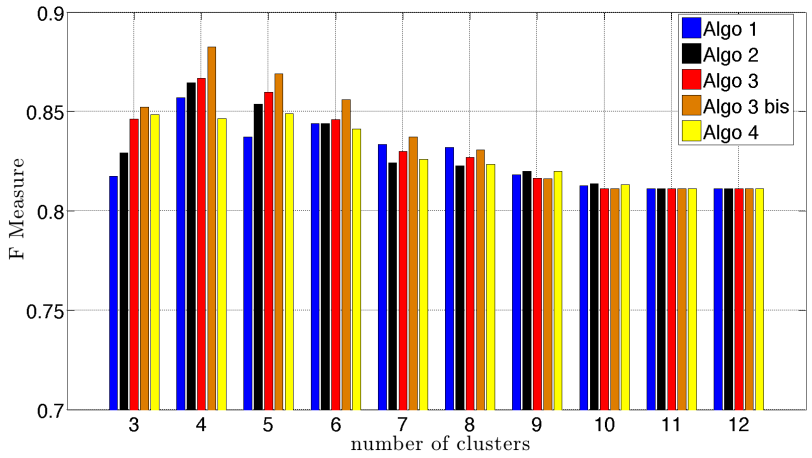


(b) Recall vs. Precision, n = number of clusters

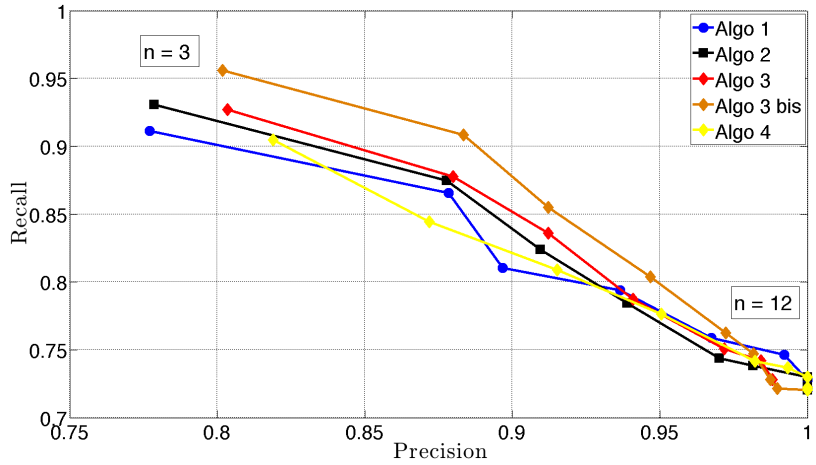
Figure 6.6: Evaluation on the ***RWC Classic*** with **automatic temporal segmentation** and **conflated ground truth**

	F-Measure	Precision	Recall	So	Su
Algo 1 - chroma	64.7%	59.4%	79.6%	76.3%	53.0%
Algo 2 - chroma	64.5%	60.4%	75.0%	69.3%	56.8%
Algo 3	66.2%	62.7%	74.6%	66.2%	56.3%
Algo 3 bis	67.5%	63.2%	77.1%	69.0%	57.3%
Algo 4	62.8 %	58.3%	75.2%	68.6%	54.1%

Table 6.16: ***RWC Classic*** - Estimated Segmentation - Conflated Ground Truth - Number of clusters n = 3



(a) F-Measures



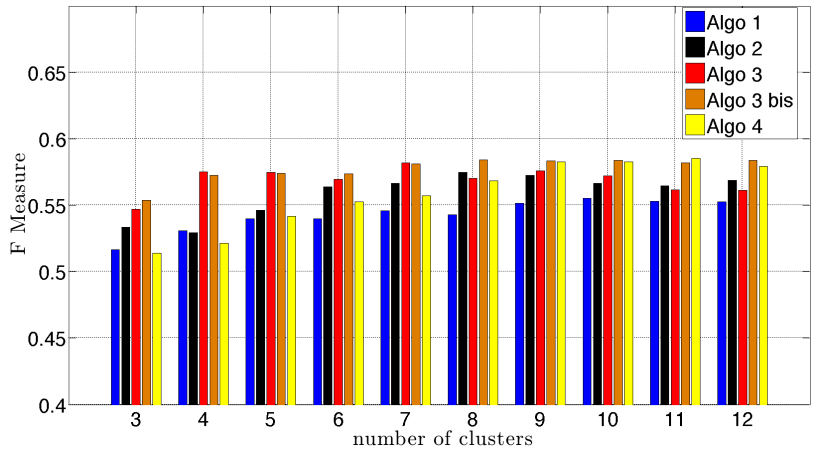
(b) Recall vs. Precision, n = number of clusters

Figure 6.7: Evaluation on the ***RWC Classic*** with **annotated temporal segmentation** and **conflated ground truth**

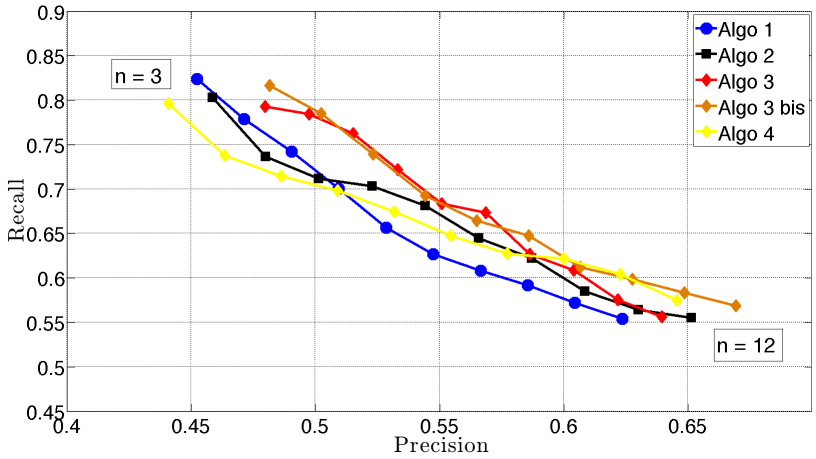
	F-Measure	Precision	Recall	So	Su
Algo 1 - chroma	85.7%	87.8%	86.5%	86.3%	88.4%
Algo 2 - chroma	86.4%	87.8%	87.5%	87.2%	88.4%
Algo 3	86.7%	88.0%	87.8%	88.0%	89.0 %
Algo 3 bis	88.3%	88.4%	90.9%	89.8%	69.2%
Algo 4	84.6%	87.2%	84.4%	84.9%	88.0%

Table 6.17: ***RWC Classic*** - Ground Truth Segmentation - Conflated Ground Truth - Number of clusters n = 4

6.6.2 Evaluation with Original Groundtruth



(a) F-Measures

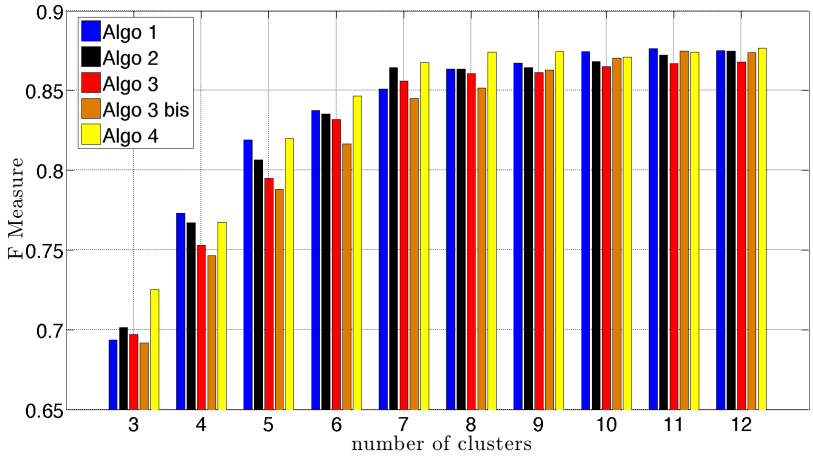


(b) Recall vs. Precision, n = number of clusters

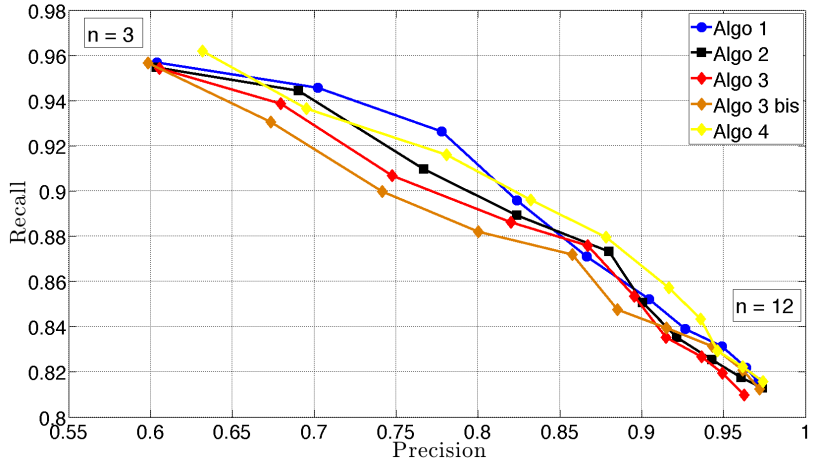
Figure 6.8: Evaluation on the ***RWC Classic*** with **automatic temporal segmentation** and **original ground truth**

	F-Measure	Precision	Recall	So	Su
Algo 1 - chroma	55.2%	57.0%	60.6%	70.2%	67.5%
Algo 2 - chroma	57.2%	59.3%	60.9%	69.7%	68.1%
Algo 3	57.0%	59.7%	62.1%	70.7%	70.1%
Algo 3 bis	58.3%	60.3%	61.8%	70.1%	70.7%
Algo 4	58.3 %	61.1%	61.7%	70.3%	71.2%

Table 6.18: ***RWC Classic*** - Estimated Segmentation - Original Ground Truth - Number of clusters n = 9



(a) F-Measures



(b) Recall vs. Precision, n = number of clusters

Figure 6.9: Evaluation on the ***RWC Classic*** with **groundtruth boundaries** and **original ground truth**

	F-Measure	Precision	Recall	So	Su
Algo 1 - chroma	86.7%	92.7%	83.9%	90.1%	96.4%
Algo 2 - chroma	86.4%	92.2%	83.5%	90.0%	96.1%
Algo 3	86.2%	91.5%	83.5%	89.8%	95.5%
Algo 3 bis	86.3%	91.5%	84.0%	90.0%	95.8%
Algo 4	87.5%	93.6%	84.3%	90.4%	96.9%

Table 6.19: ***RWC Classic*** - Ground Truth Segmentation - Original Ground Truth - Number of clusters n = 9

6.6.3 Discussion

In the temporal segmentation evaluation reported in section 6.3 we had shown that the boundary retrieval worked significantly worse on the *RWC Classic* than with the other sets, i.e. maximal F-Measure of 34%. However, the results reported in this section show that the NSMF algorithms were able to cope with this lack of precision in the segmentation and clustering performances in the range of the one obtained on the other sets are obtained. Evaluation by means of the annotated boundaries confirms this good structure clustering performance with a F-measure of up to 88.3% for *Algo 3 bis* with the conflated groundtruth.

Though the statistical relevance of this evaluation is not optimal because of its rather small size, it constitutes a good basis to compare the algorithms with each other. The first result is that the combination of the MPH similarity matrices with the NSMF algorithm (Algorithms 3 and 3 bis) obviously improves the performance of the structure segmentation. The nature of the dataset confirms that this approach is thus relevant for the modeling of structural changes contained in harmonic information. Moreover, computing the Multi-Probe Histograms with chroma sequences of length 10 seconds instead of 4 seconds allows the best performance for the dataset and does not seem to imply any strong under-segmentation effect. Clustering the segments based on their MPH representations only as in *Algo 4* did not show any advantage when using the estimated boundaries. This is however not the case when using the annotated boundaries (Tables 6.7 and 6.9) where *Algo 4* achieves more comparable labeling performances with *Algo 3* and *3 bis*. The underperformance of the approach in Tables 6.6 and 6.8 might thus be explained by the fact that missed boundaries in the estimated segments may confuse the MPH's and affect the detection of repetitions.

The rather good performance of *Algo 1* for the dataset may be attributed to the fact that structural information in the similarity matrices was partly displayed as states even though chroma features were used. The benefit of *Algo 2* is however not systematic and one can not generalized on its relevance with chroma features here. The performance of the NSMF algorithms is also very variable and dependent on the nature of the music pieces. States are not always visualized and for some music pieces, the structural information in chroma features similarity matrices is drowned in clouds of highly similar chroma features. This also explains why *Algo 1* seems to under-segment the music pieces more than the other algorithms, i.e high recall rates and lower precision rates.

Just as in the previous evaluation, the evaluation with the conflated groundtruth of the *RWC Classic* dataset shows significantly higher pairwise measures than with the original groundtruth. Fine variations in the occurrences of a same structural section that were annotated in the original groundtruth were thus averaged in the algorithm and are not conveyed in the final structural segmentation when using the estimated boundaries. Evaluation where the algorithms are provided with the

segments information however show very comparable performances with regard with the conflated and original groundtruths. A part of this performance loss can thus be attributed to the boundary retrieval step.

More generally, the standard deviations in F-measures for all algorithms is of about 12% for the MPH-based algorithms and 8% for algorithms 1 and 2. Note that the annotations of the dataset are partly incomplete and that some estimated segmentations, though being musically meaningful, did not have much in common with the annotated labels. Interpretation of the obtained performances is thus also prone to the quality of these annotations. Nevertheless, there is still room for improvement in the analysis of the similarity matrices. Indeed, while the structural information is often clearly visualized, the boundary retrieval step still misses too much information and affects the performance of the NSMF clustering. Improving the temporal segmentation will be crucial for further research in music structure segmentation.

6.7 Complexity of the Algorithms

An issue that has not been discussed in this thesis is the computation cost of our methods. Indeed, similarity matrices have a quadratic complexity, and methods based on their calculation are not exactly computationally efficient. However, reducing the audio features sample rate to 4Hz and using the diagonal properties of the matrices can reduce this cost. Our method also performs a NMF decomposition for which the complexity is difficult to estimate and is still the subject of many publications. This is however an iterative process and the computation time of the NMF thus depends on its rapidity of convergence. These and the audio features extraction steps are the most computationally expensive steps of the system. Indeed the increase in calculation induced by the Mutli-probe Histograms computation and the image filtering are very reasonable. To give an idea of how time consuming is the method, computation times for the estimation of the structure of two songs of different lengths in matlab with a recent computer are compared in Table 6.20.

Method	Song 1 Computation Time [s]	Song 2 Computation Time [s]
Algo 1	20.98	97,90
Algo 2	22.02	97.19
Algo 3	21.90	97.36

Table 6.20: Computation times for the structural segmentation in matlab of two songs. Song 1 lasts 2m02s and Song 2 lasts 4m49s.

One can observe a strong increase in the computation times between the two songs. Since the integration in a user software of our methods is beyond the scope of this thesis, we have not been looking for more computationally efficient methods for the estimation of the NMF and audio features extraction. This could however be done in the future.

6.8 Chapter Summary

In this chapter we reported the evaluation on three different datasets of the NSMF algorithms for music structure segmentation presented in this thesis. The first result is that though the boundary retrieval step is not very precise with regard to the annotated boundaries, the algorithm seems to partly compensate this underperformance of the temporal segmentation at the clustering step. However, segments are taken as a whole in the NSMF clustering and lack of precision in the boundaries still affects the labeling. Evaluation of the labeling alone indeed shows significantly better results. Further developments of the approach will have to address this issue.

In general, the post-processing of similarity matrices with image filtering did have a good impact on the performances of the system on popular music. This impact is less significant with classical music where structure tends to be less represented as states in the chroma similarity matrices. The visualization obtained by means of the Mutli-Probe Histograms and the evaluation of the structure segmentation with these matrices did in contrast show very promising results and confirmed their strong potential for the segmentation of structural information contained in harmonic variations.

Chapter 7

Conclusion

Throughout this document, we have introduced the problem of music structure segmentation and presented new approaches for its solution. We will now give a summary of these contributions and conclude with a discussion on the results and perspectives drawn by this work.

7.1 Summary of the PhD Thesis

In this PhD Thesis we have proposed a system for the task of music structure segmentation that is based on the analysis of audio similarity matrices. The system was first built upon the extraction of a mid-level description of the structural information of music pieces by means of the non-negative matrix factorization of their similarity matrices. Indeed, we showed that supposing that the description of the audio signal by means of low-level audio features could sufficiently characterize the acoustical homogeneity that defines structural states, the dimensions of such a factorization are correlated with the structural sections of the piece. This result was successfully applied to music structure segmentation with the NSMF algorithm, i.e. Non-negative Similarity Matrix Factorization. Evaluation on Popular music shows that the system achieves performances comparable to the state-of-the-art systems. However, the variety of musical contexts that one encounters in practice often prevents from properly describing structural states with classical audio features, lowering the precision of the structural segmentation. We thus proposed a further development of the NSMF algorithm with the enhancement of states visualizations in similarity matrices by means of matrix filtering techniques inspired from image segmentation research. Indeed, we find an analogy between the states formed by sections in similarity matrices taken as intensity images and the segmentation of foreground and background objects in images. We are then able by means of adequate filtering to pre-segment states in the matrices and strengthen their representation for the NSMF algorithm.

While the evaluation of this approach showed that the performance of the system could be decently improved, the algorithm remained dependent on the original properties of the audio signal and could not face every musical situation. In fact, the low-level audio features are the only abstraction of the audio signal used and

the benefit of any post-processing of the similarity matrices is tied to the quality of this description for the final task. In the remainder of the thesis, we thus came back to the original description of the audio signal in order to improve the description of the particular acoustic properties of states, especially in terms of acoustical homogeneity. We therefore focused on the description of the tonal context of music pieces. Indeed, most of music is composed within the frame of music theory that necessarily implies some regularities in the tone intervals that compose chords and melodies. Such regularities define tonal structures in which tones have particular hierarchical relations. The richness of music then resides in the ability of composers to play with the rules and modulate such tonal contexts. Our hypothesis was thus that a single structural state often does not contain any strong variations of its tonal context, whereas a musical intention could be expressed by such variations between sections.

Taking advantage of the regularities inherent to tonal contexts, we used Mutli-Probe Histograms for their description. Such histograms allow to highlight dominant local pitch class intervals within chroma sequences and we showed that they constitute a good descriptor of harmonic contexts by means of the preludes classification task. Embedding the histograms in mid-level audio features, we were able to compute similarity matrices in which the state representation was considerably strengthened. Applying the NSMF algorithm on such similarity matrices, we could significantly improve the structural segmentation of music characterized by rich harmonic variations.

Finally, a comparative evaluation of the different versions of the NSMF algorithm was proposed. In order to reflect different musical situations, the evaluation dataset was composed of popular and classical music.

7.2 Discussion and Outlook

We have shown in the description and evaluation of the NSMF algorithms in this document as well as in the related publications that the analysis of similarity matrices by means of NMF constitutes a powerful approach to the estimation of musical structures. The rather good performances obtained with the NSMF algorithm applied on classical similarity matrices suggest that though it is based on the state hypothesis, the method is robust at detecting structures even if noisily represented in the visualization. The approach consisting in post-processing the matrices by means of image processing to strengthen their structural representation also provided good results in the continuity of the first version of NSMF. Parameters of such a filtering are however not generalizable to each particular matrix and the approach presents a risk of under- or over- segmentation of the matrices before their

processing by NMF. Including contextual information about the tonal structure of segments shows very promising results for further development of music structure segmentation algorithms. It's impact on the performance of the system as reported in the evaluation has indeed shown itself to be significant. It is to be noted that annotations of the corpuses are far from being idealistic and a deeper analysis shows that the algorithm sometimes revealed meaningful musical segments that were not annotated. Independently of the performance evaluation and looking at the visualizations obtained by means the MPH mid-level features, the approach clearly strengthens the structure representation and further developments and evaluation of the system should further highlight this good result.

To some extent, the hierarchy between results obtained with the different methods proposed in this thesis correlates with the past evaluation campaigns within MIREX for the task of structural audio segmentation. Indeed, for the MIREX 2009 dataset that is composed of the *TUT Beatles* dataset, the performance of the method proposed by Mauch in [Mauch 2009] hasn't yet been matched in the evaluation campaigns. Interestingly the method highlights the interaction between the chords and structure description in music and the benefit of joint-estimation of these levels. As seen in Chapter 2, other methods for music structural segmentation proposed in the state of the art mostly derive structural information from low-level features by means of more or less complex machine learning algorithms that are not aware of any musical constraints or information that could characterize a higher level of understanding of the structural information. The measure of similarity is then reduced to a poor musical significance and performance of systems seems to be limited. While results are satisfactory for music pieces characterized by strong timbre or harmony modulations, we believe that going further in the description of structural changes in a musical way constitutes a major trend for the task.

To discuss some possible further developments of the algorithm, there should be some room for improvement in the processing of the NMF output. Up to now, the mid-level representation that is derived from the similarity matrices is clustered by means of a hierarchical approach. Looking at the patterns formed in the NMSF space, it would be interesting to include some temporal analysis and modeling of the trajectories that are often formed by structural sections. Comparison of segments could be sensibly improved this way. We also noted that the NSMF approach has been further employed for structure segmentation in [Chen 2011] since its introduction in [Kaiser 2010]. The approach includes a preliminary classification of timbral and harmonic features that yields a score matrix that is to some extent comparable to a similarity matrix on which NMF is then applied to classify the structure. An interesting extension of the authors consists in using a sparseness measure of the obtained decomposition for selecting the optimal rank for the NMF with an approach that is however computationally expensive.

Though the labeling evaluation is tied to the precision of the temporal segmentation with regard to the annotation, the huge difference between the performances achieved using either the estimated or annotated boundaries also suggests that there is considerable room for improvements at the boundary retrieval step. The audio novelty score has proved reasonable and has been extensively used in structure segmentation research but it would interesting to explore other possibilities. The latest contributions in audio analysis in the context of information geometry shows for example promising results for the audio segmentation task. Also, though the temporal modeling with the Multi-Probe Histograms strengthens the discrimination between the structural sections, it smoothens the boundaries between segments. In this case, the novelty score approach is thus not adapted and alternative approaches should be proposed. For instance, the NMF decomposition of such matrices of such matrices provides sharper boundaries between segments than in the original matrix and could be used for the segmentation.

At each step of the development of the NSMF algorithm we have tried in the analysis of music pieces to come closer to the definition of structure as a succession of states. This is however a rather simplistic assumption on music for its annotation that can not be representative of the whole musical landscape. A musical structure is indeed sometimes finer than a sequence of homogeneous segments. As a matter of fact the standard deviation in the F-Measures obtained in the evaluation is rather large, and while we achieve almost perfect segmentation for some music pieces, there are some situations where the algorithm produced segmentations that are not even close to the annotation. This raises the issue of the lack of consistent evaluation procedures for the task of music structure segmentation that has been pointed out throughout this document. There is no absolute truth in structure and interpretation of the evaluation of our algorithms in terms of absolute performance is hardly feasible. It seems to us that one of the future directions in music structure segmentation research should thus be the introduction of more knowledge of the different aspects of music theory and practice that infer structure. Modeling tonal structures by means of the Multi-Probe Histograms showed for example quite convincing results for revealing structural changes induced by harmonic variations. This approach could be further extended with a finer analysis of the chord sequences and especially of their degrees with regard to a tonality. The tempo information and rhythmic patterns could be also used, not to scale the features extraction step, but because they express in their variation a musical intention that often leads to a structural boundary. All these elements are used by musicians in a playing context and allow them to situate themselves in the structure of a musical performance. This suggests that there is more structural information in local portions of the audio signal than actually used by structure segmentation algorithms yet. Of course, in order for this to be correctly evaluated, it should also be included in the annotation.

This work also leads to the conclusion that we probably shouldn't try to find a generic solution for the task of music structure segmentation. Listening to classical and jazz music are two different experiences and algorithms should thus also be able to differentiate such situations. This was illustrated in our evaluation where the benefit of using Multi-Probe Histograms for the analysis of tonal contexts was strengthened when studying classical music. Moreover, some music genres have very characteristic structures and harmonic progressions. The classical form in Jazz compositions of a 32 bar AABA structure that uses the particular chord progression of "Rhythm Changes" by George Gershwin is a good example of such genre particularities. This structure and sequence of chord degrees are indeed used in countless Jazz pieces and it could be used for their structural analysis. Also, this reinforces the idea of having a finer analysis of the chord degrees in the audio signal, highlighting their harmonic function.

To conclude this work on more general considerations, music structure segmentation is a young discipline that raises more and more interest in the Music Information Retrieval research community. It is of course of great interest in the context for a growing demand of audio indexing and accessibility to musical contents. But it also questions our perception of music and raises the issue of measuring similarity in a given musical context. We believe that segmentation is not the only application of such an analysis and that research in this field will definitely have an impact on future computer interactions in musical performances.

Appendix A

Ranking of Preludes

In this appendix we detail the ranking of preludes obtained by measuring their similarity by means of Multi-Probe Histograms. The circle of fifths is shown in Figure A.1 for the reader to compare the relevance of the rankings with their theoretic harmonic similarity. Rankings for the Bach, Chopin and Rachmaninov preludes are given in Tables A.1, A.2 and A.3 respectively.

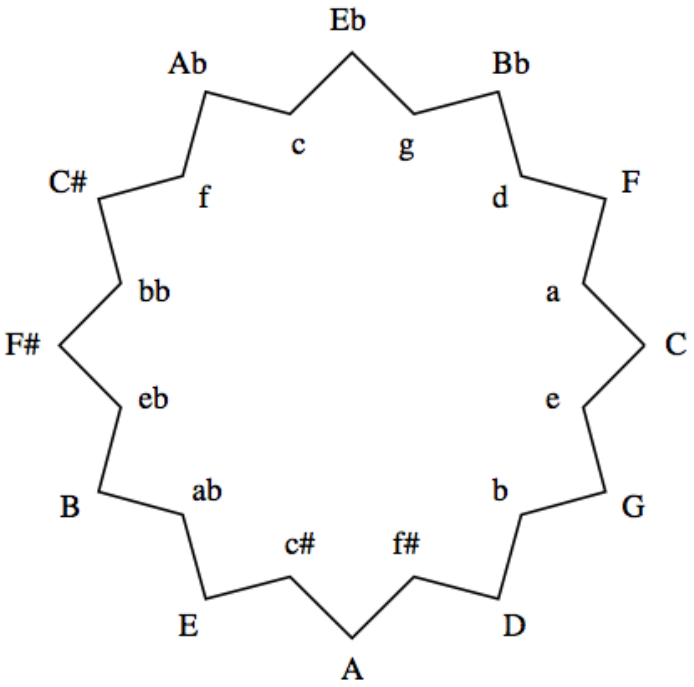


Figure A.1: Circle of Fifths

Key	C	C#	D	Eb	E	F	F#	G	Ab	A	Bb	B
1	F	F#	b	c	c#	C	eb	D	f	D	Eb	ab
2	d	Ab	A	Bb	B	d	C#	e	bb	b	c	F#
3	g	eb	e	Ab	f#	Bb	ab	d	C#	f#	g	eb
4	G	bb	G	f	A	g	B	a	c	e	F	c#
5	Bb	ab	a	g	b	a	bb	C	Eb	a	C	E
6	a	f	f#	bb	e	G	c#	g	eb	E	d	C#
7	e	B	d	eb	D	c	Ab	b	F#	G	f	f#
8	c	c#	E	C#	ab	Eb	f#	F	ab	c#	Ab	bb
9	D	Eb	C	ab	a	e	f	A	Bb	d	bb	b
10	Eb	c	F	F	F#	D	E	Bb	B	B	G	A
11	f	f#	c#	C	eb	f	Eb	f#	c#	C	eb	Ab
12	b	E	g	F#	C#	A	c	E	g	F	a	f
13	A	Bb	B	d	G	b	b	c	F	F#	C#	e
14	Ab	A	Bb	B	bb	Ab	A	Eb	C	ab	F#	D
15	bb	g	F#	G	Ab	bb	Bb	c#	f#	g	ab	Eb
16	E	b	ab	a	d	f#	D	B	E	C#	e	c
17	f#	F	eb	c#	f	E	g	f	d	eb	D	a
18	C#	D	C#	e	C	C#	e	ab	A	bb	b	G
19	c#	C	Eb	D	F	eb	F	Ab	a	Bb	B	Bb
20	eb	d	c	E	c	F#	d	F#	G	Ab	A	g
21	F#	e	bb	b	Eb	ab	G	eb	b	f	f#	d
22	ab	a	Ab	f#	g	c#	a	bb	e	Eb	E	F
23	B	G	f	A	Bb	B	C	C#	D	c	c#	C

key	c	c#	d	eb	e	f	f#	g	ab	a	bb	b
1	Eb	E	F	F#	D	Ab	A	Bb	eb	e	Ab	D
2	Bb	B	C	ab	b	bb	c#	C	B	G	eb	e
3	f	f#	G	bb	G	c	b	d	F#	D	C#	A
4	Ab	ab	g	C#	a	C#	E	F	C#	d	f	f#
5	g	F#	a	B	A	Eb	D	G	c#	C	F#	G
6	bb	C#	Bb	Ab	E	Bb	B	Eb	Ab	A	Eb	E
7	F	A	D	f	C	eb	F#	c	bb	F	ab	a
8	C#	eb	e	Eb	d	F#	e	a	f	b	c	c#
9	C	b	A	c#	f#	ab	C#	D	E	E	Bb	B
10	eb	D	b	c	F	F	ab	e	Eb	g	B	d
11	ab	e	Eb	f#	c#	g	eb	f	f#	f#	c#	C
12	d	Ab	c	E	g	C	a	Ab	c	Bb	f#	F
13	F#	bb	f#	Bb	B	B	G	b	A	c#	g	F#
14	G	f	f	b	Bb	c#	bb	bb	b	c	F	g
15	B	a	Ab	A	ab	d	d	A	Bb	Eb	E	ab
16	a	G	bb	g	c	E	Ab	eb	D	B	C	eb
17	c#	c	E	D	F#	f#	f	C#	e	f	d	C#
18	e	Eb	eb	F	eb	a	C	ab	g	ab	A	Bb
19	E	d	c#	e	Eb	G	F	F#	a	Ab	b	bb
20	D	C	C#	d	C#	A	g	f#	G	bb	D	Eb
21	b	g	F#	G	f	e	Eb	B	F	F#	a	c
22	f#	F	B	C	bb	b	Bb	E	d	C#	G	Ab
23	A	Bb	ab	a	Ab	D	c	c#	C	eb	e	f

Table A.1: Relation between Preludes of the *Well-Tepered Clavier Books* in terms of similarity. Similarity is computed by means of Multi-Probe Histograms

Key	C	C#	D	Eb	E	F	F#	G	Ab	A	Bb	B
1	G	Ab	A	Bb	e	d	C#	C	C#	D	Eb	ab
2	F	c#	b	g	a	C	Ab	a	F#	f#	g	F#
3	d	F#	f#	Ab	ab	f	B	d	ab	b	bb	eb
4	c	ab	G	f	B	g	eb	e	Eb	G	eb	c#
5	e	bb	a	bb	A	G	ab	D	bb	e	f	C#
6	a	B	e	eb	c#	c	bb	F	f	a	Ab	f#
7	g	f	d	c	b	Bb	c#	b	c#	c#	F	Ab
8	f	Eb	E	C#	F#	a	Eb	A	c	E	F#	E
9	b	f#	C	F#	eb	Eb	f#	c	g	d	c	bb
10	D	eb	c#	ab	G	bb	f	E	Bb	ab	C#	b
11	A	c	F	F	Ab	e	Bb	g	B	C	B	Eb
12	Bb	Bb	B	B	D	D	E	f#	eb	B	ab	Bb
13	Eb	g	eb	f#	C#	eb	g	f	f#	bb	d	A
14	E	E	ab	C	f#	A	c	Bb	E	C#	f#	D
15	Ab	A	g	d	Bb	Ab	b	ab	F	F	C	e
16	ab	F	Bb	E	d	b	D	c#	A	Ab	E	c
17	bb	b	F#	c#	C	f#	A	Eb	C	Bb	D	a
18	c#	d	c	D	Eb	C#	e	eb	e	eb	c#	f
19	C#	e	Eb	b	g	E	F	Ab	b	F#	a	g
20	f#	D	bb	A	bb	F#	a	B	d	c	G	G
21	eb	C	Ab	G	c	c#	C	bb	D	g	e	C
22	F#	a	f	a	f	ab	d	C#	a	Eb	A	d
23	B	G	C#	e	F	B	G	F#	G	f	b	F

key	c	c#	d	eb	e	f	f#	g	ab	a	bb	b
1	g	C#	F	Bb	a	g	D	Bb	B	e	Bb	D
2	f	ab	G	bb	G	Eb	c#	f	Ab	G	Ab	G
3	C	Ab	C	F#	E	c	A	Eb	C#	E	C#	A
4	Eb	B	a	B	C	Ab	B	c	c#	d	eb	e
5	Ab	f#	g	Eb	A	F	eb	F	F#	A	Eb	f#
6	F	F#	e	f#	b	bb	C#	Ab	E	C	f	a
7	C#	A	A	Ab	d	Bb	F#	bb	Eb	D	F#	E
8	Bb	bb	f	C#	D	C#	b	C	c	b	g	B
9	ab	E	D	f	ab	F#	bb	d	b	F	c#	ab
10	G	D	c	g	F	C	ab	eb	f#	g	B	C
11	d	eb	Bb	ab	c#	eb	Ab	C#	eb	ab	f#	c
12	bb	c	E	E	c	d	Eb	G	e	c	c	c#
13	b	e	Eb	c#	g	ab	Bb	ab	bb	Bb	ab	eb
14	c#	b	bb	D	B	c#	E	F#	Bb	eb	F	F#
15	e	f	f#	F	Bb	f#	f	a	A	B	A	d
16	F#	Eb	eb	b	eb	G	g	D	g	c#	d	F
17	eb	Bb	b	c	Ab	B	d	e	f	f#	E	Bb
18	a	g	c#	a	f#	E	G	c#	a	f	C	Ab
19	B	a	Ab	d	f	e	c	f#	D	Eb	D	C#
20	f#	d	C#	e	F#	b	a	E	G	Ab	e	Eb
21	E	G	ab	A	C#	D	F	B	C	bb	b	f
22	D	C	F#	G	bb	a	e	A	d	F#	a	bb
23	A	F	B	C	Eb	A	C	b	F	C#	G	g

Table A.2: Relation between Chopin's Preludes in terms of similarity. Similarity is computed by means of Multi-Probe Histograms

Key	C	C#	D	Eb	E	F	F#	G	Ab	A	Bb	B
1	c	Ab	b	c	c#	bb	eb	e	Eb	D	Eb	ab
2	a	F#	A	Ab	B	Bb	C#	b	C#	b	d	E
3	f	f	G	Bb	A	a	B	D	c	E	F	c#
4	F	bb	e	eb	ab	d	bb	A	bb	c#	g	F#
5	e	c#	g	bb	C#	f	f#	g	f	e	bb	eb
6	Eb	eb	f#	f	b	C	f	a	eb	G	eb	C#
7	Ab	ab	d	g	F#	f#	Ab	C	F#	f#	c	A
8	bb	B	a	C#	f#	Ab	ab	d	Bb	B	Ab	b
9	Bb	c	E	C	D	c	c	f#	ab	a	a	Ab
10	G	E	c#	ab	eb	Eb	c#	B	C	ab	C	f
11	d	Eb	B	F#	f	C#	E	c	F	d	f	c
12	g	f#	ab	F	e	g	Eb	E	B	g	F#	D
13	f#	F	eb	d	Ab	F#	Bb	ab	c#	F#	C#	f#
14	b	Bb	F#	B	G	eb	F	Bb	E	C#	f#	Eb
15	C#	C	C	a	bb	e	b	eb	d	C	G	bb
16	eb	A	Bb	G	a	c#	A	Eb	f#	F	ab	e
17	A	d	F	c#	c	A	D	c#	g	eb	e	G
18	D	b	C#	e	C	D	g	F	a	f	D	Bb
19	c#	g	c	f#	F	G	C	F#	A	bb	B	C
20	F#	a	Eb	E	Eb	E	d	f	e	Bb	b	F
21	ab	D	bb	b	d	b	e	Ab	G	Ab	c#	g
22	E	G	f	D	g	B	G	C#	b	c	A	a
23	B	e	Ab	A	Bb	ab	a	bb	D	Eb	E	d

key	c	c#	d	eb	e	f	f#	g	ab	a	bb	b
1	Eb	E	g	F#	G	bb	F#	d	B	d	f	G
2	Ab	B	Bb	C#	b	Ab	F	Bb	E	F	Ab	D
3	f	A	a	ab	D	C#	A	D	eb	C	F	e
4	C	C#	F	B	A	c	D	G	c#	e	C#	A
5	eb	ab	D	c	a	F	c#	Eb	C#	g	F#	E
6	Bb	F#	C	Eb	C	C	bb	a	F#	Bb	Bb	B
7	bb	f#	Eb	Bb	g	F#	C#	e	Ab	G	Eb	f#
8	C#	eb	G	Ab	f#	Eb	f	F	Eb	D	c	g
9	F#	Ab	bb	f	d	eb	b	C	c	A	C	ab
10	ab	f	f#	bb	E	f#	E	b	A	b	f#	c#
11	F	D	e	f#	B	Bb	eb	eb	f	f#	eb	a
12	g	b	A	E	c	c#	d	c	b	bb	d	C
13	B	bb	Ab	c#	F	B	g	f#	G	Eb	c#	F#
14	f#	F	c	g	Bb	ab	C	A	D	f	a	eb
15	d	a	eb	F	ab	E	e	bb	bb	c	B	d
16	G	c	f	b	c#	a	a	Ab	Bb	c#	g	c
17	a	C	b	d	Eb	d	B	F#	e	E	E	f
18	e	e	C#	C	eb	b	G	ab	C	Ab	ab	F
19	c#	G	F#	D	F#	e	Bb	C#	g	eb	A	C#
20	b	Eb	c#	A	f	A	c	f	f#	C#	e	Bb
21	E	d	ab	G	bb	g	Ab	B	a	ab	D	Eb
22	D	Bb	E	e	Ab	G	ab	c#	F	B	b	bb
23	A	g	B	a	C#	D	Eb	E	d	F#	G	Ab

Table A.3: Relation between Rachmaninov's Preludes in terms of similarity. Similarity is computed by means of Multi-Probe Histograms

Appendix B

Description of Songs Databases

Contents

A brief description of the evaluation datasets used in this thesis is given in this Appendix.

B.1 TUT Beatles

The *TUT Bealtes* dataset is composed of the songs of the 12 following albums with their release date:

1. 1963 - *Please Please Me*
2. 1963 - *With the Bealtes*
3. 1964 - *A Hard Day's Night*
4. 1964 - *Beatles For Sale*
5. 1965 - *Help!*
6. 1965 - *Rubber Soul*
7. 1965 - *Revolver*
8. 1967 - *Sgt. Pepper's Lonely Heart Club Band*
9. 1967 - *Magical Mystery Tour*
10. 1968 - *The Beatles*
11. 1969 - *Abbey Road*
12. 1970 - *Let It Be*

B.2 RWC Popular Music

Piece No.	**Title**	**Artist** (Vocal)	**Length**	**Tempo**	**Instrumentation**	**Drum Information**
No. 1	Eien no replica	Kazuo Nishi	3:29	135	Gt	Drum sequences
No. 2	Magic in your eyes	Hiromi Yoshii	3:42	100	Gt & Bs	Drum sequences
No. 3	HORO	MIT	3:15	111		Drum sequences
No. 4	Spice of Life	Hisayoshi Kazato	4:02	86	Gt	Drum sequences
No. 5	Koino Ver.2.4	Eves	3:48	135	Gt	Drum sequences
No. 6	Funky Life	Oriken	3:26	120	Gt	Drum sequences
No. 7	PROLOGUE	Tomomi Ogata	4:58	122	Gt	Drum sequences
No. 8	Jinsei konnamono	fevers	3:12	127	Gt & Bs	Drum sequences
No. 9	Doukoku	Kazuo Nishi	4:37	70	Gt & Bs & Dr	Live drums
No. 10	Getting Over	Brakes	3:35	125	Gt	Drum sequences
No. 11	Ienai	Hisayoshi Kazato	4:27	90	Gt	Drum sequences
No. 12	KAGE-ROU	Kazuo Nishi	3:24	120	Gt & Bs	Drum sequences
No. 13	Catch ball	Konbu	3:39	103	Gt	Drum sequences
No. 14	Karehairo no Twilight	Rin	3:54	88	Gt & Bs	Drum sequences
No. 15	old fashioned	Katsuyuki Ozawa	2:42	132	Gt	Drum sequences
No. 16	Game of Love	Hiromi Yoshii	4:22	122	Gt & Bs	Drum loops
No. 17	Anata to aete	Hiromi Yoshii	4:01	97	Gt	Drum sequences
No. 18	True Heart	Tomomi Ogata	4:14	112	Gt	Drum sequences
No. 19	COOL Motion	Hisayoshi Kazato	4:49	130	Gt	Drum sequences
No. 20	Tokimeki no syunkan	Eri Ichikawa	4:10	134		Drum sequences

Table B.1: List of Songs in the RWC Popular Music Dataset

Piece No.	**Title**	**Artist** (Vocal)	**Length**	**Tempo**	**Instrumentation**	**Drum Information**
No. 21	Feeling In My Heart	Rin	4:28	98	Gt	Drum sequences
No. 22	Koi ni ochiru jikan ni	Kousatsu & Nishi	3:29	135	Gt	Drum sequences
No. 23	SHAKE	MAPS	3:21	132	Gt	Drum sequences
No. 24	it's all right	Hisayoshi Kazato	4:00	130		Drum sequences
No. 25	tell me	Tomomi Ogata	4:16	103	Gt	Drum sequences
No. 26	aozora sanpo michi	Tomoko Nitta	3:27	158	Gt & Bs	Drum sequences
No. 27	stay	Shingo Katsuta	5:18	124	Gt	Drum sequences
No. 28	Fly away	Tomomi Ogata	4:10	109	Gt	Drum sequences
No. 29	One Two STEP	Kazuo Nishi	3:35	103	Gt	Drum sequences
No. 30	syounen no omoi	Mitsuru Tanimoto	3:16	104		Drum sequences
No. 31	Moving Round and Round	Yuuichi Nagayama	4:10	129		Drum loops
No. 32	what could I do for you	Masaki Kuehara	4:12	125	Gt	Drum sequences
No. 33	DREAM MAGIC	Hiromi Yoshii	4:47	108	Gt & Bs	Drum sequences
No. 34	Hitoyo no yume	Hiromi Yoshii	3:27	93	Gt & Bs & Dr	Live drums
No. 35	Midarana kami no moushigo	Hiromi Yoshii	3:13	170	Gt & Bs & Dr	Live drums
No. 36	over and over	Kazuo Nishi	5:16	135	Gt & Bs & Dr	Live drums
No. 37	Replica	Yoshinori Hatae	3:59	184	Gt & Bs & Dr	Live drums
No. 38	1999	Kousuke Morimoto	4:35	125	Gt & Bs & Dr	Live drums
No. 39	SPUL	Kousuke Morimoto	4:48	73	Gt & Bs & HH	Drum sequences
No. 40	promise	Kazuo Nishi	3:46	122	Gt & Bs	Drum sequences

Table B.2: List of Songs in the RWC Popular Music Dataset

Piece No.	**Title**	**Artist** (Vocal)	**Length**	**Tempo**	**Instrumentation**	**Drum Information**
No. 41	Non Stop Driving	Katsuyuki Ozawa	2:50	200	Gt	Drum sequences
No. 42	Fly to the moon	Kousuke Morimoto	4:08	125	Gt & Bs	Drum sequences
No. 43	Centimeter no kodoku	Kazuo Nishi	3:24	163	Gt	Drum sequences
No. 44	REAL na 5 hun	Kousuke Morimoto	4:06	124	Gt & Bs	Drum sequences
No. 45	Hajimari	Kousuke Morimoto	3:42	77	Gt & Bs & Dr	Live drums
No. 46	Senro wa tsuzukuyo	Kousuke Morimoto	3:19	168	Gt	Drum sequences
No. 47	Deaetakara	Satoshi Kumasaka	3:30	94	Gt & Bs & Dr &Pf	Live drums
No. 48	Syodoubutsu	Hiroshi Sekiya	4:29	86	Gt	Drum sequences
No. 49	Sekai no mikata	Hiroshi Sekiya	4:35	100	Gt	Drum sequences
No. 50	Mrs. Maril	Rin	3:15	114	Gt & Bs	Drum sequences
No. 51	Modoranai natsu	Hiroshi Sekiya	6:07	104	Gt ? 2 & Bs & Dr & Pf	Live drums
No. 52	Haru ga kurukara	Tomomi Ogata	3:45	140	Gt	Drum sequences
No. 53	Ashita wa	Rin	3:39	132	Gt & Bs	Drum sequences
No. 54	Harukana omoi	Rin	3:42	125	Gt & Bs	Drum sequences
No. 55	First Love	Akiko Kaburagi	4:09	74	Gt ? 2 & Bs & Dr	Live drums
No. 56	I've got a mail	Masashi Hashimoto	5:22	74	Gt & Bs	Drum loops
No. 57	Stay with me	Masashi Hashimoto	4:27	70		Drum sequences
No. 58	Silver shoes	Rin	3:45	118		Drum loops
No. 59	Tenshi no utatane	Rin	3:25	98	Gt & Tp	Drum loops
No. 60	Kumorizora	Yuzu Iijima	4:03	148	Gt ? 2 & Bs	Drum loops

Table B.3: List of Songs in the RWC Popular Music Dataset

Piece No.	**Title**	**Artist** (Vocal)	**Length**	**Tempo**	**Instrumentation**	**Drum Information**
No. 61	FOR YOU	Kazuo Nishi	4:43	121	Gt	Drum sequences
No. 62	Be with me Now	Rin	3:11	81	Gt	Drum sequences
No. 63	Power of mind	Reiko Sato	4:11	126	Gt & Bs & Dr	Live drums
No. 64	So Long	Kousuke Morimoto	4:52	100	Gt & Bs	Drum sequences
No. 65	Tanabata	Makiko Hattori	3:52	76	Gt & Bs &	Dr Live drums
No. 66	Mousugu natsu ga kuru	M & Y	5:20	76	Gt & HH & Pf	Drum sequences
No. 67	Tokei no hayasa wa	Makiko Hattori	4:10	88	Gt & Bs & Dr	Live drums
No. 68	Nichiyoubi	Makiko Hattori	4:54	92	Gt & Bs & Dr	Live drums
No. 69	Gin no sora	Hiromi Yoshii	6:00	62	Gt	Drum sequences
No. 70	Miageta sora wa	Tamako Matsuzaka	4:06	104	Gt & Bs & Dr & Pf	Live drums
No. 71	Tsuki no youni	Hiromi Yoshii	4:46	70	Gt & Pf	Without drums
No. 72	Heart to Hurt	Kousuke Morimoto	3:21	76	Pf & Vc	Without drums
No. 73	Miss Maria	Kazuo Nishi	3:16	144	Pf & Vc	Without drums
No. 74	Kimi no iro	Kousuke Morimoto	3:14	94	Gt	Without drums
No. 75	Toui machi e	Hiromi Yoshii	3:21	108	Pf	Without drums
No. 76	Chikai	Kousuke Morimoto	3:50	70	Gt	Without drums
No. 77	Aishiteru	Makiko Hattori	3:56	120	Gt & Pf	Without drums
No. 78	Kumo	Masaki Kuehara	4:21	75	Gt	Without drums
No. 79	Together	Tomomi Ogata	4:28	92		Without drums
No. 80	Sagashimono	Tomomi Ogata	3:39	80		Without drums

Table B.4: List of Songs in the RWC Popular Music Dataset

Piece No.	**Title**	**Artist** (Vocal)	**Length**	**Tempo**	**Instrumentation**	**Drum Information**
No. 81	How Deep Is Your Love?	Donna Burke	3:50	90	Gt	Drum sequences
No. 82	Once in a life time	Shinya Iguchi	5:37	88	Gt & Dr	Live drums
No. 83	Doing That Thing	Jeff Manning	3:37	140	Gt	Drum sequences
No. 84	Someday	Shinya Iguchi	4:40	138	Gt & Dr	Live drums
No. 85	Waiting for the moment	Jeff Manning	3:30	98	Gt	Drum sequences
No. 86	Angel Baby	Betty	4:17	80	Gt	Drum sequences
No. 87	I think of you	Jeff Manning	4:51	90	Gt	Drum sequences
No. 88	Woman Like You	Shinya Iguchi	4:09	120	Gt & Dr	Live drums
No. 89	Life Is What You Make It To Be	Donna Burke	3:43	134	Gt	Drum sequences
No. 90	Don't Say Good bye	Shinya Iguchi	4:37	127	Gt & Dr	Live drums
No. 91	Change Of Heart	Donna Burke	3:43	124	Gt	Drum sequences
No. 92	I'll be there for you	Betty	3:40	134		Drum sequences
No. 93	Sweet Dreams	Donna Burke	4:18	90	Gt	Drum sequences
No. 94	Life	Betty	3:43	78	Gt	Drum sequences
No. 95	Feel	Jeff Manning	3:52	161	Gt & Bs & Dr	Live drums
No. 96	Weekend	Betty	4:40	130		Drum loops
No. 97	Don't Lie To Me	Donna Burke	4:03	91	Gt	Drum sequences
No. 98	31 BLUES	Jeff Manning	3:31	107	Gt & Bs &	Dr Live drums
No. 99	Once and For All	Shinya Iguchi	5:17	73	Gt	Without drums
No. 100	No Regrets	Shinya Iguchi	4:53	80	Gt	Drum loops

Table B.5: List of Songs in the RWC Popular Music Dataset

B.3 RWC Classical Music

Piece No.	**Title**	Composer	**Artist**	**Length**
No. 22	Hungarian Dance no.5 in F# minor	Brahms, Johannes	Furuichi & Kikuchi	2:25
No. 23	Suite Ma Mère l"Oye (Mother Goose)	Ravel, Maurice	Furuichi & Kikuchi	
	Pavane de la Belle au Bois Dormant			1:20
	Petit Poucet			2:45
	Laideronnette, Impératrice des Pagodes			3:25
	Les Entretiens de la Belle et de la Bête			4:09
	Le Jardin Féerique			2:50
No. 24	The Anna Magdalena Bach Notebooks	Bach, Johann Sebastian	Atsuko Watanabe	
	Menuet in G major			1:26
	Menuet in G minor			1:29
	Musette in D major			0:52
No. 25	The Well-Tempered Clavier, Book 1	Bach, Johann Sebastian	Atsuko Watanabe	
	no.1 in C major BWV.846. Prelude			2:03
	Fugue			2:11
	no.2 in C minor BWV.847. Prelude			1:31
	ibid. Fugue			1:54
No. 26	Piano Sonata in A major, K.331/300i. 1st mvmt.	Mozart, Wolfgang Amadeus	Norio Shimizu	10:00
No. 27	Variations on "Ah Vous Dirai-je Maman", K.265/300e	Mozart, Wolfgang Amadeus	Yuriko Furuichi	7:27
No. 28	Piano Sonata no.23 in F minor, op.57 "Appasionata". 1st mvmt.	Beethoven, Ludwig van	Sarasa Watanabe	9:50
No. 29	"Träumerei" from Suite *Kinderszenen*, op.15	Schumann, Robert	Norio Shimizu	2:25
No. 30	Nocturne no.2 in Eb major, op.9-2	Chopin, Frédéric	Yuriko Furuichi	4:02
No. 31	Etude in E major, op.10-3	Chopin, Frédéric	Miki Narai	4:16
No. 32	Etude in F minor, op.25-2	Chopin, Frédéric	Yuriko Furuichi	1:49

Table B.6: List of Pieces in the RWC Classical Music Dataset

Piece No.	Title	Composer	Artist	Length
No. 33	Polonaise no.6 in Ab major "Héroïque" op.53	Chopin, Frédéric	Mutsuko Tajima	6:45
No. 34	"La Campanella" from *Grandes Études de Paganini*	Liszt, Franz	Naoko Matsuoka	5:09
No. 35	Three Gymnopédies	Satie, Erik	Norio Shimizu	
	no.1			3:49
	no.2			3:01
	no.3			2:46
No. 36	6 Sonatas and Partitas for Unaccompanied Violin no.6	Bach, Johann Sebastian	Michi Mizutori	3:32
No. 37	Violin Sonata no.5 in F major, op.24 "Spring"	Beethoven, Ludwig	van Michi Mizutori	1:17
No. 38	24 Caprices, op.1 no.24 in A minor	Paganini, Niccolò	Michi Mizutori	5:21
No. 39	Violin Sonata in A major. 4th mvmt.	Franck, César	Michi Mizutori	6:44
No. 40	Meditation from *Thaïs*	Massenet, Jules	Michi Mizutori	5:06
No. 41	Suite for Unaccompanied Cello no.1 in G major	Bach, Johann Sebastian	Hitoshi Miyazawa	2:39
No. 42	"Le Cygne" from Suite *Le Carnaval des Animaux*	Saint-Saëns, Camille	Hitoshi Miyazawa	2:31
No. 43	Sicilienne op.78	Fauré, Gabriel	Hidehiko Watase	4:09
No. 44	"The Flight of the Bumble Bee"	Rimski-Korsakov, Nikolay	Hidehiko Watase	1:03
No. 45	Air "Ombra mai fu" (Largo) from "Serse"	Handel, George Frideric	Rihoko Yanagisawa	2:49
No. 46	"Der Hölle Rache kocht in meinem Herzen" (*Die Zauberflöte*)	Mozart, Wolfgang Amadeus	Makiko Saito	3:08
No. 47	"Ständchen" (Serenade) from *Schwanengesang*	Schubert, Franz Peter	Asuka Ueda	3:46
No. 48	"Der Lindenbaum" from *Winterreise*, op.89/D.911	Schubert, Franz Peter	Tomihiko Kotaka	4:26
No. 49	Air "La donna è mobile qual piuma al vento" from *Rigoletto*	Verdi, Giuseppe	Ryosuke Hagino	2:11
No. 50	Micaela"s Air "Je dis que rien ne m'épouvante" from *Carmen*	Bizet, Georges	Hiromi Tohyama	5:05

Table B.7: List of Pieces in the RWC Classical Music Dataset

List of Figures

List of Tables

Bibliography

[Abdallah 2005] Samer Abdallah, Katy Noland, Mark Sandler, Michael Casey and Christophe Rhodes. *Theory and evaluation of a Bayesian music structure extractor*. In Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR), 2005. (Cited on page 22.)

[Alatan 1998] A.A. Alatan, L. Onural, M. Wollborn, R. Mech, E. Tuncel and T. Sikora. *Image sequence analysis for emerging interactive multimedia services-the European COST 211 framework*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 8, no. 7, pages 802 –813, November 1998. (Cited on page 54.)

[Arvanitidou 2009] M.G. Arvanitidou, A. Glantz, A. Krutz, T. Sikora, M. Mrak and A. Kondoz. *Global motion estimation using variable block sizes and its application to object segmentation*. In Proceedings of the international Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), pages 173–176, May 2009. (Cited on page 55.)

[Aucouturier 2005] Jean-Julien Aucouturier, François Pachet and M. Sandler. *"The way it Sounds": timbre models for analysis and retrieval of music signals*. IEEE Transactions on Multimedia, vol. 7, no. 6, pages 1028–1035, 2005. (Cited on pages 22 and 25.)

[Barrington 2010] Luke Barrington, Antoni B. Chan and Gert Lanckriet. *Modeling music as a dynamic texture*. IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 3, 2010. (Cited on pages 25 and 72.)

[Bartsch 2005] Mark A. Bartsch and Gregory H. Wakefield. *Audio thumbnailing of popular music using chroma-based representations*. IEEE Transactions on Multimedia, vol. 7, no. 1, pages 96–104, 2005. (Cited on pages 11, 22 and 25.)

[Brown 1991] Judith Brown. *Calculation of a constant Q spectral transform*. Journal of the Acoustical Society of America, vol. 89, no. 1, pages 425–434, January 1991. (Cited on page 15.)

[Bruderer 2006] Michael J. Bruderer, Martin F. McKinney and Armin Kohlrausch. *Structural boundary perception in popular music*. In Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR), pages 198–201, 2006. (Cited on pages 11 and 18.)

[Catral 2004] M. Catral, Lixing Han, Michael Neumann and R. J. Plemmons. *On reduced rank nonnegative matrix factorizations for symmetric matrices*. In

Special Issue on Positivity in Linear Algebra, pages 107–126, 2004. (Cited on page 34.)

[Chai 2005] Wei Chai. *Automated analysis of musical structure.* PhD thesis, Massachusetts Institute of Technology, School of Architecture and Planning, Program in Media Arts and Sciences, 2005. (Cited on page 25.)

[Chen 2011] Ruofeng Chen and Ming Li. *Music Structural Segmentation by Combining Harmonic and Timbral Information.* In Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR), 2011. (Cited on pages 43 and 127.)

[Cohen 1991] Annabel J. Cohen. *Tonality and Perception: Musical scales primed by excerpts from the Well-Tempered Clavier of J.S.Bach.* Psychological Research, vol. 53, no. 4, pages 305–314, 1991. (Cited on pages 84, 85 and 93.)

[Cont 2011] Arshia Cont, Shlomo Dubnov and Gerard Assayag. *On the Information Geometry of Audio Streams with Applications to Similarity Computing.* IEEE Transactions on Audio, Speech and Language Processing, vol. 19, no. 4, May 2011. (Cited on page 24.)

[Cooper 2002] Matthew L. Cooper and Jonathan Foote. *Summarizing video using non-negative similarity matrix factorization.* In Proceedings of the IEEE Workshop on Multimedia Signal Processing, pages 25–28. IEEE Signal Processing Society, 2002. (Cited on pages 17, 21, 32 and 35.)

[Cooper 2003] M. Cooper and J. Foote. *Summarizing Popular Music via Structural Similarity Analysis.* In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2003. (Cited on pages 21 and 25.)

[Dannenberg 2002] Roger B. Dannenberg and Ning Hu. *Pattern discovery techniques for music audio.* In Proceedings of the International Conference on Music Information Retrieval, pages 63–70, 2002. (Cited on pages 23 and 25.)

[Deng 2001] Y. Deng and B.S. Manjunath. *Unsupervised segmentation of color-texture regions in images and video.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 8, pages 800 –810, August 2001. (Cited on page 54.)

[Ding 2005] Chris Ding, Xiaofeng He and Horst D. Simon. *On the equivalence of nonnegative matrix factorization and spectral clustering.* In Proceedings of the SIAM International Conference on Data Mining (SMD), pages 606–610, 2005. (Cited on page 34.)

[Dubnov 2011] Shlomo Dubnov, Gérard Assayag and Arshia Cont. *Audio Oracle analysis of Musical Information Rate*. In Proccedings of the IEEE International Conference on Semantic Computing (ICSC), 2011. (Cited on page 24.)

[Eronen 2007] A. Eronen. *Chorus detection with combined use of MFCC and chroma features and image processing filters*. In Proceedings of the 10th International Conference on Digital Audio Effects (DAFx), 2007. (Cited on pages 11 and 25.)

[Fan 2001] Jianping Fan, D.K.Y. Yau, A.K. Elmagarmid and W.G. Aref. *Automatic image segmentation by integrating color-edge extraction and seeded region growing*. IEEE Transactions on Image Processing, vol. 10, no. 10, pages 1454 –1466, October 2001. (Cited on page 54.)

[Foote 1999] Jonathan Foote. *Visualizing music and audio using self-similarity*. In Proceedings of the ACM Multimedia, pages 77–80, 1999. (Cited on pages 3, 7, 17, 25 and 94.)

[Foote 2000] Jonathan Foote. *Automatic Audio Segmentation using a Measure of Audio Novelty*. In Proceedings of the IEEE International Conference on Multimedia and Expo, 2000. (Cited on pages 19, 25 and 106.)

[Gjerdingen 2008] Robert O. Gjerdingen and David Perrott. *Scanning the Dial: The Rapid Recognition of Music Genres*. Journal of New Music Research, vol. 37, no. 2, pages 93–100, 2008. (Cited on page 74.)

[Goto 2003] Masataka Goto. *Chorus-Section Detecting Method for Musical Audio Signals*. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2003. (Cited on pages 3, 18, 22 and 25.)

[Imai 1983] S. Imai. *Cepstral analysis synthesis on the mel frequency scale*. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1983. (Cited on page 12.)

[Jacobson 2001] 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. (Cited on page 42.)

[Jehan 2005] Tristan Jehan. *Creating Music by Listening*. PhD thesis, Massachusetts Institute of Technology, 2005. (Cited on pages 21 and 25.)

[Jensen 2007] Kristoffer Jensen. *Multiple Scale Music Segmentation Using Rhythm, Timbre, and Harmony*. EURASIP Journal on Advances in Signal Processing, vol. 2007, 2007. (Cited on page 25.)

[Jones 1989] Mari Riess Jones and Marilyn Boltz. *Dynamic attending and responses to time.* Psychological Review, vol. 96, pages 459–491, 1989. (Cited on page 74.)

[Kaiser 2010] Florian Kaiser and Thomas Sikora. *Music Structure Discovery in Popular Music using Non-negative Matrix Factorization.* In Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), aug 2010. (Cited on pages 4, 21 and 127.)

[Kaiser 2011a] Florian Kaiser, Marina Georgia Arvanitidou and Thomas Sikora. *Audio Similarity Matrices Enhancement in an Image Processing Framework.* In Proceedings of the 9th International Workshop on Content-Based Multimedia Indexing (CBMI), Madrid, Spain, June 2011. (Cited on pages 4 and 22.)

[Kaiser 2011b] Florian Kaiser and Thomas Sikora. *Multi-Probe Histograms: A Mid-Level Harmonic Feature for Music Structure Segmentation.* In Proceedings of the 14th International Conference on Digital Audio Effects (DAFx), Paris, France, September 2011. (Cited on page 5.)

[Kamphorst 1987] J.-P. Eckmann S. Oliffson Kamphorst and D. Ruelle. *Recurrence Plots of Dynamical Systems.* Europhysics Letters, vol. 4, no. 9, 1987. (Cited on page 17.)

[Kim 2005] Hyoung-Gook Kim, Nicolas Moreau and Thomas Sikora. Mpeg-7 audio and beyond: Audio content indexing and retrieval. John Wiley & Sons, 2005. (Cited on page 11.)

[Krumhansl 1979] Carol L. Krumhansl and Roger N. Shepard. *Quantification of the Hierarchy of Tonal Functions within a Diatonic Context.* Journal of Experimental Psychology: Human Perception and Performance, vol. 5, no. 4, pages 579–594, 1979. (Cited on pages 74 and 83.)

[Krumhansl 1991] Carol L. Krumhansl. *Music Psychology: Tonal Structures in Perception and Memory.* Annual Reviews Psychology, vol. 43, pages 277–303, 1991. (Cited on page 84.)

[Krumhansl 2010] Carol L. Krumhansl. *Plink: "Thin Slices" of Music.* Music Perception, vol. 27, no. 5, pages 337–354, June 2010. (Cited on page 74.)

[Krutz 2007] A. Krutz, M. Kunter, M.l Mandal, M. Frater and T. Sikora. *Motion-based Object Segmentation using Sprites and Anisotropic Diffusion.* In Proceedings of the International Workshop on Image Analysis for Multimedia Interactive Services, June 2007. (Cited on pages 55, 56 and 57.)

[Krutz 2009] A. Krutz, A. Glantz, T. Borgmann, M. Frater and T. Sikora. *Motion-based object segmentation using local background sprites.* In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1221–1224, April 2009. (Cited on page 55.)

[Krutz 2010] Andreas Krutz. *From Sprites to Global Motion Temporal Filtering.* PhD thesis, Technische Universität Berlin, 2010. (Cited on page 56.)

[Kunter 2009] Matthias Kunter, Sebastian Knorr, Andreas Krutz and Thomas Sikora. *Unsupervised object segmentation for 2D to 3D conversion.* In Proceedings of the IS&T/SPIE Electronic Imaging, January 2009. (Cited on page 55.)

[Lee 1999] Daniel D. Lee and H. Sebastian Seung. *Learning the parts of objects by non-negative matrix factorization.* Nature, vol. 401, no. 6755, pages 788–791, October 1999. (Cited on pages 32 and 33.)

[Lee 2000] Daniel D. Lee and H. Sebastian Seung. *Algorithms for Non-negative Matrix Factorization.* In Proceedings of NIPS, July 21 2000. (Cited on pages 33 and 34.)

[Levy 2008] M. Levy and M. Sandler. *Structural Segmentation of Musical Audio by Constrained Clustering.* IEEE Transactions on Audio, Speech & Language Processing, vol. 16, no. 2, pages 318–326, 2008. (Cited on pages 11, 22, 25 and 27.)

[Logan 2000] B. Logan and S. Chu. *Music summarization using key phrases.* In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), ICASSP '00, pages II 749–752, 2000. (Cited on pages 21 and 25.)

[Lu 2004] Lie Lu, Muyuan Wang and Hong-Jiang Zhang. *Repeating pattern discovery and structure analysis from acoustic music data.* In Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval, MIR '04, pages 275–282, 2004. (Cited on pages 22 and 25.)

[Lukashevich 2008] Hanna M. Lukashevich. *Towards Quantitative Measures of Evaluating Song Segmentation.* In Juan Pablo Bello, Elaine Chew and Douglas Turnbull, editeurs, ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, September 14-18, 2008, pages 375–380, 2008. (Cited on page 28.)

[Ma 1997] W.Y. Ma and B.S. Manjunath. *Edge flow: A framework of boundary detection and image segmentation.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 744 –749, June 1997. (Cited on page 54.)

[Maddage 2006] Namunu C. Maddage. *Automatic Structure Detection for Popular Music*. IEEE MultiMedia, vol. 13, pages 65–77, 2006. (Cited on pages 25 and 26.)

[Marolt 2006] Matija Marolt. *A Mid-level Melody-based Representation for Calculating Audio Similarity*. In Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR), October 2006. (Cited on page 25.)

[Mauch 2009] Matthias Mauch, Katy Noland and Simon Dixon. *Using Musical Structure To Eenhance Automatic Chord Transcription*. In ISMIR, 2009. (Cited on pages 25, 49, 98 and 127.)

[Mueller 2006] Meinard Mueller and Frank Kurth. *Enhancing similarity matrices for music audio analysis*. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2006. (Cited on pages 22 and 72.)

[Müller 2007a] Meinard Müller. Information retrieval for music and motion. Springer Verlag, 2007. (Cited on page 16.)

[Müller 2007b] Meinard Müller and Frank Kurth. *Towards Structural Analysis of Audio Recordings in the Presence of Musical Variations*. EURASIP Journal on Advances in Signal Processing, vol. 2007, no. 1, 2007. (Cited on page 25.)

[Müller 2011] Meinard Müller, Frank Kurth and Michael Clausen. *Chroma Toolbox: Pitch, Chroma, CENS, CRP*. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR), 2011. (Cited on page 16.)

[Obin 2006] Nicolas Obin. Apprentissage de la corrélation de la f0 et de l'enveloppe spectrale: Application à la transposition de la voix parlée. Master's thesis, ATIAM, 2006. (Cited on page 13.)

[Ong 2007] B. Ong. *Structural Analysis and Segmentation of Music Signals*. PhD thesis, Music Technology Group - Universitat Pompeu Fabra, 2007. (Cited on pages 22, 25, 60 and 61.)

[Otsu 1979] Nobuyuki Otsu. *A Threshold Selection Method from Gray-Level Histograms*. IEEE Transactions on Systems, Man and Cybernetics, vol. 9, no. 1, pages 62 –66, 1979. (Cited on pages 22 and 57.)

[Paulus 2006] Jouni Paulus and Anssi Klapuri. *Music structure analysis by finding repeated parts*. In Proceedings of the 1st ACM workshop on Audio and music computing multimedia, AMCMM '06, pages 59–68, 2006. (Cited on page 25.)

[Paulus 2009] 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, vol. 17, no. 6, pages 1159–1170, 2009. (Cited on page 25.)

[Paulus 2010] Jouni Paulus, Meinard Müller and Anssi Klapuri. *Audio-based Music Structure Analysis.* In Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), 2010. (Cited on pages 19, 20 and 25.)

[Peeters 2002a] Geoffroy Peeters. *Automatically Selecting Signal Descriptors for Sound Classification.* In Proceedings of the International Computer Music Conference (ICMC), 2002. (Cited on page 44.)

[Peeters 2002b] Geoffroy Peeters. *Toward automatic music audio summary generation from signal analysis.* In Proceedings of the International Conference on Music Information Retrieval (ISMIR), pages 94–100, 2002. (Cited on pages 15, 23 and 72.)

[Peeters 2004a] Geoffroy Peeters. Deriving musical structures from signal analysis for music audio summary generation: "sequence" and "state" approach, volume 2771 of *Lecture notes in Computer Science*, pages 143–166. Springer, 2004. (Cited on pages 20, 22, 25 and 72.)

[Peeters 2004b] Geoffroy Peeters. *A Large Set of Audio Features for Sound Description in the CUIDADO Project.* Rapport technique, CUIDADO I.S.T., 2004. (Cited on page 11.)

[Peeters 2007] Geoffroy Peeters. *Sequence Representation of Music Structure using Higher-Order Similarity Matrix and Maximum-Likelihood Approach.* In Proceedings of the International Conference on Music Information Retrieval (ISMIR), 2007. (Cited on pages 22 and 25.)

[Peiszer 2007] E. Peiszer. Automatic audio segmentation: Segment boundary and structure detection in popular music. Master's thesis, Vienna University of Technology, 2007. (Cited on page 21.)

[Perona 1990] P. Perona and J. Malik. *Scale-space and edge detection using anisotropic diffusion.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 7, pages 629 –639, July 1990. (Cited on pages 54 and 57.)

[Pollak 2000] I. Pollak, A.S. Willsky and H. Krim. *Image segmentation and edge enhancement with stabilized inverse diffusion equations.* IEEE Transactions on Image Processing, vol. 9, no. 2, pages 256 –266, February 2000. (Cited on page 54.)

[Rhodes 2007] Christophe Rhodes and Michael Casey. *Algorithms for Determining and Labelling Approximate Hierarchical Self-Similarity.* In Proceedings of the International Conference on Music Information Retrieval (ISMIR), pages 41–46, 2007. (Cited on pages 23 and 25.)

[Salembier 1999] P. Salembier and F. Marques. *Region-based representations of image and video: segmentation tools for multimedia services.* IEEE Transactions on Circuits and Systems for Video Technology, vol. 9, no. 8, pages 1147 –1169, December 1999. (Cited on page 54.)

[Schoenberg 1969] Arnold Schoenberg. Structural functions of harmony. W. W. Norton & Company, 1969. (Cited on page 82.)

[Sezgin 2004] Mehmet Sezgin and Bulent Sankur. *Survey over image thresholding techniques and quantitative performance evaluation.* Journal of Electronic Imaging, vol. 13, no. 1, pages 146–168, 2004. (Cited on page 54.)

[Shi 2000] Jianbo Shi and J. Malik. *Normalized cuts and image segmentation.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pages 888 –905, August 2000. (Cited on page 54.)

[Shiu 2005] Y. Shiu, H. Jeong and C. C. J. Kuo. *Musical structure analysis using similarity matrix and dynamic programming.* In Proceedings of SPIE - Multimedia Systems and Applications VII, volume 6015, pages 398–409, 2005. (Cited on page 25.)

[Smith 2010] Jordan B. L. Smith. A comparison and evaluation of approaches to the automatic formal analysis of musical audio. Master's thesis, McGill University, 2010. (Cited on pages 20 and 103.)

[Soille 2003] P. Soille. Morphological image analysis: Principles and applications. Springer-Verlag New York, Inc., 2003. (Cited on page 59.)

[Turnbull 2007] Douglas Turnbull, Gert Lanckriet, Elias Pampalk and Masataka Goto. *A Supervised Approach for Detecting Boundaries in Music using Difference Features and Boosting.* Proceedings of the International Conference on Music Information Retrieval (ISMIR), September 2007. (Cited on page 25.)

[Wegener 2008] Sebastian Wegener, Martin Haller, Juan José Burred, Thomas Sikora, Slim Essid and Gaël Richard. *On the Robustness of Audio Features for Musical Instrument Classification.* In Proceedings of the 16th European Signal Processing Conference (EUSIPCO), 2008. (Cited on page 12.)

[Wellhausen 2003] Jens Wellhausen and Michael Höynck. *Audio Thumbnailing Using MPEG-7 Low Level Audio Descriptors.* In Proceedings of SPIE International Symposium ITCom on Internet Multimedia Management Systems

IV, volume 5242, pages 63–73, Orlando, FL, USA, September 2003. (Cited on pages 22 and 25.)

[Yu 2010] Y. Yu, M. Crucianu, V. Oria and E. Damiani. *Combining Multi-Probe Histogram and Order-Statistics Based LSH for Scalable Audio Content Retrieval*. In ACM Multimedia, 2010. (Cited on pages 74, 75 and 76.)