Automatic Melody Segmentation

Marcelo Enrique Rodríguez López

ISBN 978-90-393-6594-6.

Typeset in LaTeX.

Printed in the Netherlands by Koninklijke Wöhrmann, http://nl.cpibooks.com/.

The research presented in this dissertation has been funded by the Netherlands Organization for Scientific Research NWO-VIDI, grant 276-35-001.

Automatic Melody Segmentation

Automatische Melodie Segmentatie
(with a summary in English)

PROEFSCHRIFT

Ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op maandag 20 juni 2016 des middags te 12.45 uur

door

Marcelo Enrique Rodríguez López

geboren op 2 Maart 1981 te Calama, Chili

Promotor: Prof.dr. R.C. Veltkamp
Copromotor: Dr. A. Volk

Contents

Chapter 1

Introduction

1.1 Motivation

Music is a universal cultural trait. In each culture music plays a number of roles, for example in religious ceremonies (Sylvan 2002), in entertainment (Frith 1998), or in forging a social identity (Christenson and Roberts 1998). Music has the power to induce intense emotional and physiological responses in humans (Harrison and Loui 2014; Woelfer and Lee 2012), and has been shown to serve as a powerful stimulus for evoking memories (Kirke et al. 2015). Music's cultural and social importance makes its analysis relevant for a number of scientific, commercial, artistic, and technological disciplines, ranging from sociology, cognitive science, and neuroscience, to the film, game, and music industry.

As stated by Wiering and Veltkamp (2005) *"Music's most important manifestation, sound generated during performance, is volatile. It can be captured, though imperfectly, in two ways, as sound recording and as music notation"*. Modern digital technology has made easy and cheap to capture, store, and distribute music. Because of these technological developments, at present recorded and notated music is available digitally, in large quantities, creating a need for automatic ways to analyse its content.

1.2 Focus

The work presented in this dissertation focuses on modelling a specific type of music analysis, namely, *segmentation*. In the field of Musicology, segmentation refers to a score analysis technique, whereby notated pieces or music passages are divided into 'units' referred to as sections, periods, phrases, and so on. Segmentation analysis is a widespread practice among musicians: performers use it to help them memorise pieces (Rusbridger 2013; Cienniwa 2014), music theorists and historians use it to compare works (Caplin 1998; Roberts 2001), music students use it to understand the compositional strategies of a given composer or genre (La Rue 1970; Cook 1994). In the field of Music Psychology it is posited that a similar type of analysis is performed by our auditory system when constructing mental representations of music. In fact, most theories consider segmentation to be a core listening mechanism, fundamental to the way humans recognise, categorise, and memorise music (Lerdahl and Jackendoff 1983; Narmour 1992, 1990; Hanninen 2001). In this dissertation we are interested in modelling segmentation via computer simulation. This puts our research at the intersection between the discipline of *Computational Modelling of Music Cognition* and that of *Music Information Retrieval*. Below we briefly describe each discipline in turn, and in the following section we list applications of a segmentation analysis for each.

Music Information Retrieval (MIR)

MIR has been defined as a *"research endeavour that strives to develop innovative ... searching schemes, novel interfaces, and evolving networked delivery mechanisms in an effort to make the world's vast store of music accessible to all"* (Downie 2003). MIR is a discipline whose boundaries have been under constant expansion since its conception.[1] For comprehensive descriptions we refer to (Downie 2003; Orio 2006; Casey et al. 2008; Müller 2011). For discussions on challenges we refer to (Byrd and Crawford 2002; Futrelle and Downie 2003; Wiering 2006; Serra et al. 2013; Sturm 2014).

Computational Modelling of Music Cognition (CMMC)

CMMC (Desain et al. 1998) is a discipline lying on the border between artificial intelligence and psychology. It is concerned with building computer models of human

[1] In fact, in recent years the MIR discipline has grown outside of the confines of what can be thought of as 'retrieval' (Herrera et al. 2009), and so the community has updated research agendas and roadmaps, which have proposed to change the name to Music information *Research* (Serra et al. 2013).

cognitive processes, based on an analogy between the human mind and computer programs. The brain/mind and computer are viewed as general-purpose symbol-manipulation systems, capable of supporting software processes (generally no analogy is drawn at a hardware level). For a discussion on the present state and challenges of CMMC refer to (Pearce and Rohrmeier 2012).

1.3 Applications

In MIR, applications of the information gained by segmenting musical input are diverse, for instance:

- Using the start or end location of segments as markers to aid music (and multimedia) navigation, editing, and synchronisation (Rubin et al. 2013; Gohlke et al. 2010; Swaminathan and Doddihal 2007; Lee and Cremer 2008).
- Using segments as building blocks for automatic or computer-assisted composition and improvisation systems (Cope 1992; Rowe 1992; Bigo and Conklin 2015; Loeckx 2015).
- Using segments as means to index music files for fast and accurate search and browsing of music collections (Downie and Nelson 2000; Chang and Jiau 2004; Orio and Neve 2005; Chang and Jiau 2011; Sridhar et al. 2010; Sankalp et al. 2016).
- Sampling music pieces by selecting prominent or memorable segments to perform musicological analyses (Serrà et al. 2012; Mauch et al. 2015; Jensen and Hebert 2015).
- Rendering segments visible to guide and orientate players in music games (Biamonte 2010), or to improve music discovery systems (McFee 2015).

In CMMC computer models of segmentation are taken as a way to test theories of segmentation proposed in the fields of music theory and music cognition, and ultimately as a way to gain understanding into how music is perceived by humans. Some landmark experiments are (Deliège 1987; Bruderer 2008; Clarke and Krumhansl 1990; Pearce et al. 2010b).

1.4 Scope

Type of music: In this dissertation we focus on the segmentation of *melody*, which is *"readily separable from other musical constructs (such as harmony) and is thus subject with minimal damage to reductionist science, and it is clearly present in the vast majority of the world's musics, in excitingly varied forms"* (Wiggins and Forth 2015, p. 129).

Music file format: Melodies are assumed to be mentally represented as a sequence of discrete sonic events in a way that is comparable to conventional Western notation. Thus, in this dissertation the input to a segmentation analysis is represented in *symbolic* form (Harris et al. 1991), i.e. any computer readable format where melodic events correspond roughly to notes as notated on a score (e.g. `MIDI` or `**kern`).

Task: Research in music segmentation modelling has been conducted by subdividing the segmentation problem into a number of different tasks – see Figure 1.1 for an illustration. In this dissertation we focus on the task of boundary detection. We limit our scope to segments resembling the music theoretic concepts of *figure*, *phrase*, and *section*. Special attention is given to the study of phrases, due to their fundamental role in musical analysis within the field of Musicology – see (Schoenberg 1967; La Rue 1970; Stein 1979; Rothstein 1989; Caplin 1998).

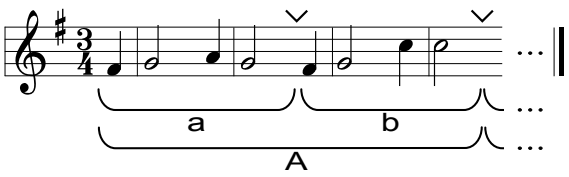

Figure 1.1: Example segmentation analysis of a melody. At present, in MIR segmentation consists of two tasks: segment boundary detection and segment labelling. Segment boundary detection is the task of automatically locating the time instants separating contiguous segments (downward arrows in the figure). Segment labelling is the task of tagging segments with an equivalence class label, to identify the sets of segments with similar musical characteristics (a,b,A in the figure).

1.5 Challenges

Machine melody segmentation has been an active topic of research for more than three decades. While significant advances have been made during that period, results

of recent evaluation studies suggest that a fully automatic solution to the problem is still out of reach (Rodríguez-López and Volk 2012). The three main reasons that make melody segmentation challenging are:

1. *The problem of segmentation is difficult to formalise.* Segments and segmentation are often discussed and described by analogy to 'form analysis' in music theory, where terms such as 'figure', 'phrase', or 'section' are used. However, these terms are generally ambiguous, being often used interchangeably by different theorists and researchers. Likewise, segmentation criteria are often defined using terms such as 'proximity', 'novelty', or 'homogeneity', by analogy to visual segment perception (Gestalt principles). It is commonly unclear what these terms mean within a musical context. These issues make difficult to motivate and formalise segmentation, complicating the categorisation, comparison, and evaluation of machine segmenters.
2. *Evaluating machine melody segmenters is non-trivial.* A number of non-trivial problems arise when trying to assess what constitutes a 'valid' segmentation of a melody. Generally, the approach taken is to have listeners manually segment a collection of melodies, and then compare these segmentations to automatically obtained segmentations. However, existing test collections lack stylistic diversity and often contain only a single manual segmentation. The lack of stylistic diversity complicates having an idea of how performance results generalise. The low number of manual segmentations per melody makes it difficult to define how to score different segmentations, complicating the design of appropriate quantitative measures of performance.
3. *Segment perception is multifaceted and context-dependent.* Listening studies have shown that, even for short music fragments, there are often multiple factors influencing segment perception. Many of these factors are likely to suggest different segmentations to a listener. Moreover, experimental findings also suggest that the perceived salience and importance of any factor seen to influence segmentation seems to be highly dependent on the local and global musical context in which it occurs. This makes segmentation a formidable problem to tackle, requiring the development of multi-variate and multi-scale systems, which are also likely to require distributed/parallel processing strategies and coordination control mechanisms.

1.6 Contributions

Our contributions tackle each of the challenges described in §1.5.

Contributions tackling challenge 1

- In depth discussion of music theory and music cognition concepts relating to melody segmentation.

- Introduced a novel taxonomy of music segmentation cues, which allows a more coherent organisation of automatic segmenters in respect to their goals rather than the technology or modelling technique employed, and facilitates the motivation and description of segmentation criteria.

Contributions tackling challenge 2

- Identified issues with the current evaluation framework of automatic segmentation, and proposed solutions, most of which are used to evaluate segmenters in this dissertation.

- Development of a corpus of 125 jazz theme melodies for benchmarking machine segmenters. Each melody in the corpus has been annotated with segment boundaries by three human listeners.

Contributions tackling challenge 3

- Introduced, implemented, and evaluated three single-cue machine segmenters: a repetition-based segmenter, a contrast-based segmenter, and a template-based segmenter.

- Introduced, implemented, and evaluated a system that combines single-cue segmenters, using context-aware strategies.

Publications

This dissertation is based on the following publications:

P1 Rodríguez-López, M. and Volk, A. (2012). Automatic Segmentation of Symbolic Music Encodings: A Survey. *Technical Report UU-CS-2012-015*. Utrecht University.

P2 Rodríguez-López, M. and Volk, A. (2012). Melodic Segmentation Using the Jensen-Shannon Divergence, in *Proc. of the 11th International Conference on Machine Learning and Applications (ICMLA)*, pp. 351–356.

P3 Rodríguez-López, M. and Volk, A. (2013). Symbolic Segmentation: A Corpus-Based Analysis of Melodic Phrases, in *Proc. of the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pp. 381–388.

P4 Rodríguez-López, M., Volk, A. and de Haas, W.B. (2014). Comparing Repetition-Based Melody Segmentation Models, in *Proc. of the 9th Conference on Interdisciplinary Musicology (CIM)*, 2014, pp. 143–148.

P5 Rodríguez-López, M., Bountouridis, D. and Volk, A. (2014). Multi-strategy Segmentation of Melodies, in *Proc. of the 15th Conference of the International Society for Music Information Retrieval (ISMIR)*, 2014, pp. 207–212.

P6 Rodríguez-López, M. and Volk, A. (2015). Location Constraints for Repetition-Based Segmentation of Melodies, in *Proc. of the 5th International Conference on Mathematics and Computation in Music (MCM)*, pp. 73–84.

P7 Rodríguez-López, M. and Volk, A. (2015). On the Evaluation of Automatic Segment Boundary Detection, in *Proc. of the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pp. 234–246.

P8 Rodríguez-López, M., Bountouridis, D. and Volk, A. (2015). Novel Music Segmentation Interface and the Jazz Tune Collection, in *Proc. of the 5th Folk Music Analysis Workshop (FMA)*, pp. 79–85.

P9 Rodríguez-López, M. and Volk, A. (2015). Selective Acquisition Techniques for Enculturation-Based Melodic Phrase Segmentation, in *Proc. of the 16th Conference of the International Society for Music Information Retrieval (ISMIR)*. pp. 218–224.

1.7 Dissertation Overview

This dissertation is organised by mirroring our outline of contributions, i.e. first the problem is formalised, subsequently evaluation strageties are proposed, then four segmenters are introduced and tested, and finally conclusions are drawn. Below a brief summary of each chapter is presented.

In Chapter 2 we formalise the problem of melody segmentation. In doing so we find two main obstacles: unclear terminology and unclear goals. We hence introduce a conceptual framework, aimed to guide the development of machine segmenters. The framework is grounded on cognitive theories of music listening, and is composed of a conceptual model and a taxonomy. The conceptual model consists of working definitions for what a segmenter is (as a cognitive mechanism) and how it operates. The taxonomy classifies both the processing mechanisms (subcomponents) and information (cues) needed for the segmenter to operate. Moreover, we provide working definitions of segments and segment types, and define computational modelling tasks.

The conceptual framework is used to classify existing segmenters, identify niches of novel research, and motivate/guide the development of melody segmenters introduced in this dissertation.

In Chapter 3 we critically review the evaluation chain of automatic melody segmenters. At present automatic segmentations are evaluated by comparing them to manual human annotated segmentations (a *direct* scenario). We identify three important limitations of this evaluation scenario: first, available segment-annotated databases lack stylistic diversity; second, currently used evaluation measures give no partial score to nearly missing a boundary; third, due to the low number of boundary annotations per melody, it is impossible to estimate how to penalise an insertion or full miss. Our contributions to tackle these limitations are threefold: we present a new benchmark corpus consisting of 125 jazz melodies which helps broadening the stylistic diversity of annotated corpora; we survey measures proposed in the field of text segmentation that can give partial scores to near misses; we propose an approach to help extending the annotations of existing corpora to allow better penalisation of insertions and full misses. Additionally, in this chapter we construct the melodic database used to test the automatic segmenters proposed in this dissertation. The corpus consists of 125 vocal and 125 instrumental folk melodies, as well as 125 jazz melodies. We refer to this corpus as the `FJ375`.

In Chapter 4 we tackle the problem of *repetition*-based melody segmentation. Repetition-based segmentation relies on identifying and selecting repetitions of melodic fragments, and then using the start or end points of selected repetitions as segment boundaries. A known limitation of automatic melody repetition identification is that the number of repetitions detected is generally much larger than the number of repetitions actually recognised by human listeners. Robust methods to select segmentation-determinative repetitions is thus crucial to the performance of repetition-based segmentation models. Repetition selection is most often modelled by enforcing constraints based on the frequency, length, and temporal overlap of/between detected repetitions. We propose and quantify constraints based on the location of repetitions relative to (a) each other, (b) the whole melody, and (c) temporal gaps. To test our selection constraints, we incorporate them in a state-of-the-art repetition-based segmenter. The original and constraint-extended versions of the segmenter are used to segment the FJ375 melodies. Our results show the constraint-extended version of the segmenter achieves a statistically significant improvement over it's original version, suggesting that location is an important aspect of how human listeners might be recognising segmentation-determinative repetitions.

In Chapter 5 we tackle the problem of *contrast*-based melody segmentation. Contrast-based segmenters attempt to identify boundaries as points of change in the attributes

describing a melody. One of the main limitations of existing contrast-based segmenters that they rely on manual setting of parameter which are crucial to the their performance. These parameters are: (a) selecting an appropriate window size (amount of temporal context) to detect meaningful contrasts, (b) selecting the size of the melodic figures needed for detecting meaningful contrasts, and (c) selecting the melodic representation where meaningful contrasts can be detected. We propose and evaluate a statistical model of contrast detection that can automatically select and tune the aforementioned parameters. We test the model in the FJ375 corpus. Our results show our contrast-based segmenter achieves a statistically significant improvement over the selected baselines, suggesting our parameter automation techniques are better fit to model how human listeners identify segmentation-determinative contrasts.

In Chapter 6 we tackle the problem of *template*-based melody segmentation. Template-based segmentation investigates the role of melodic schemata, acquired through listening experience, in melody segmentation. We concentrate on the role that melodic enculturation has in the segmentation of a melody of the same and other styles. (By enculturation we mean having internalised melodic figures characteristic of a style.) One of the main limitations of existing template-based segmenters is that they model previous listening experience by storing information indiscriminately into memory (memory refers to a model of long-term memory that embodies an artificial listener's previous listening experience.) We argue that selective (rather than indiscriminate) information acquisition is necessary to simulate enculturation. We hence propose and investigate two techniques for selective acquisition learning. To compare the segmentations produced by enculturated segmenters using selective and non-selective acquisition techniques, we perform a melody classification experiment involving melodies of different cultures, where the segments are used as classification features. Our results show that the segments produced by our selective learning segmenters substantially improve classification accuracy when compared to segments produced by using a non-selective learning segmenter, two local segmentation methods, and two naïve baselines.

In Chapter 7 we tackle the problem of segmentation cue *combination*. Multi-cue segmenters consist of two or more segmenters that model a single segmentation cue, and put their effort in devising strategies to combine the output of these segmenters. We formulate multiple cue segmentation as an optimisation problem, and introduce a cost function that penalises segmentations by considering cues related to boundaries, segments, and the complete segmentation. Our segmenter differs from existing multi-cue segmenters in three respects. First, it is more complete, in that it has a wider coverage of cues. Second, it has a higher degree of autonomy, in that it has dedicated modules to estimate all needed cue information. Third, it relies less on

hardcoded parameters. An added feature of our segmenter is the interpretability of its mechanisms, made possible by its modular approach and cue-defined cost function. We evaluate our segmenter on the `FJ375` corpus. To have a comparison point we also evaluated a state-of-the-art multi-cue segmenter and two naïve baseline segmenters on the same corpus. Results show that our segmenter achieves statistically significant $\overline{F1}$ improvements of ~8% in respect to the state-of-the-art, and of over 20% in respect to the baselines. Our results also show clear benefits of using multiple sources of information for segmentation, which supports the hypothesis of human segmentation mechanisms as being composed of multi-scale, multi-cue, parallel processing modules.

In Chapter 8 we present the conclusions of this dissertation. We discuss our main findings and their implications, and give an outlook for future work.

This dissertation also contains three Appendices. Appendix A presents a summary of formulas used to compute melodic attribute sequences from symbolic input. Appendix B presents a summary of cognitive theories of music segmentation. Appendix C presents a summary of empirical studies of segmentation.

Chapter 2

Formalising the Problem of Melody Segmentation

In this chapter we formalise the problem of melody segmentation.

Chapter Contributions

We introduce a conceptual framework to guide the development of machine segmenters, grounded on cognitive theories of music listening.

The framework is composed of a conceptual model and a taxonomy. Moreover, we provide working definitions of segments and segment types, and define computational modelling tasks. The conceptual framework is used to classify existing segmenters, identify niches of novel research, and motivate/guide the development of melody segmenters introduced in this dissertation.

This chapter is based on (Rodríguez-López and Volk 2012).

2.1 Introduction

In this chapter we set out to formalise the problem of melody segmentation.

Challenges. Automatic melody segmentation has been an active topic of research in MIR and Music Cognition for more than three decades. In (Rodríguez-López and Volk 2012) we survey more than 30 machine segmenters proposed after 1980, the majority of which are meant to segment melody/monophony. In our survey we note two issues that make discussing and comparing machine segmenters difficult:

1. *Terminology is often unclear.* Researchers generally refer to segments using terminology of music theory, e.g. 'phrase', 'subphrase', and 'motive'. However, these terms are often left unspecified or used interchangeably. Similarly, the terms used to denote the cues that are modelled, e.g. 'novelty' or 'discontinuity', are left unspecified, taking implicit, ad-hoc meanings, e.g. 'novelty := abrupt changes in timbre', 'discontinuity := pitch jumps'.

2. *Existing segmenter classification schemes are unclear and uninformative.* Machine segmenters are often described and categorised in respect to technical aspects. For instance, whether they are 'rule based', 'memory based', 'knowledge driven', or 'data driven'. This technically motivated distinction tends to obscure the goals of the segmenters, and on occasion it makes segmenters with the same overall goal appear as incompatible.

These terminological issues put in evidence the lack of a cognitive framework, which negatively affect both segmenter development and evaluation.

Contributions. To address the issues outlined above we introduce a cognitively-grounded conceptual framework to guide the development of machine segmenters. The framework is composed of a conceptual model and a taxonomy. The conceptual model consists of working definitions for what a segmenter is (as a cognitive mechanism) and how it operates. The taxonomy classifies both the processing mechanisms (subcomponents) and information (cues) needed for the segmenter to operate. Moreover, we provide working definitions of segments and segment types, and define computational modelling tasks. The conceptual framework is used to classify existing segmenters, identify niches of novel research, and motivate/guide the development of melody segmenters introduced in this dissertation.

Chapter Structure. This chapter is organised as follows. In §2.2 we introduce a conceptual model of segmentation. In §2.3 we introduce a taxonomy of cues. In §2.4

we provide working definitions for the type of music to be segmented (melodies) and the type of segments to be identified (phrases, and so on). In §2.5 we describe and classify existing melody segmenters. Finally, in §2.6 we present our conclusions.

2.2 Conceptual Model

In this section we introduce our conceptual model. The model presents working definitions for segmentation, segment, segment structure, and segment cue. Most importantly, we introduce the concept of *segmenter mechanism* and how it operates.[2]

2.2.1 Music Segmentation as a Cognitive Process

Our conceptual model is based on cognitive theories of music listening that include segmentation (Lerdahl and Jackendoff 1983; Narmour 1992, 1990; Hanninen 2001; Deliège 2001; Ockelford 2004; Ahlbäck 2004; Wiggins and Forth 2015) – refer to Appendix B.1 for a review. These theories all share the idea that the human mind transforms continuous auditory input into sequences of 'musical events', at multiple time scales. That is, the consensus is that when we experience an extension of time, say a minute of music, we do so based on events lasting fractions of a second, a few seconds, and tens of seconds.

To give more ground to this idea, and at the same time relate segmentation to other cognitive processes, we take some time to briefly introduce what is arguably the most influential theory of music listening: Lerdahl and Jackendoff's Generative Theory of Tonal Music (GTTM). In this theory segmentation is considered one of the four main structuring processes of music, the other three being *metric induction*, *time-span reduction*, and *prolongational reduction*.[3] In Figure 2.1 we illustrate a mockup analysis of melody using these cognitive structuring processes. According to the GTTM the analysis of a piece can be seen as the construction of a tree, where the root of the tree (uppermost level) represents the entire piece, the intermediate branches represent the result of analyses of hierarchy between nodes, and the terminal nodes, the 'leaves',

[2] Throughout this dissertation the term 'segmenter' is used in three contexts: (1) *machine* segmenter, (2) *human* segmenter, and (3) segmenter *mechanism*. The first refers to a computational model of segmentation. The second to a person performing manual segmentation. The third to a mental process. To allow disambiguation we are consistent with the use of the accompanying term. On occasion we drop the 'machine' noun for brevity, but it is only on cases where the paragraph offers a clear context.

[3] It must be noted that in the GTTM segmentation is referred to as *grouping*. In other theories segmentation is referred to as *chunking*. In this dissertation we consider all of these terms to be analogous.

represent notes as notated on a score. In Figure 2.1 (left) we mention the types of structural description resulting with each analysis. The arrows indicate the interrelationships between the depicted types of structural description. The **segmentation analysis** results in a nested set of segments (represented by horizontal curly brackets), ordered so that each group of notes is enclosed in a larger group of notes. The **metric analysis** results in a grid of strong/weak accent positions, hierarchically ordered as either subdivision or multiples of a central pulse or 'beat'. The **time-span reduction** analysis uses the metrical and grouping analyses, and as a result retains tree nodes considered more important in respect to rhythmic stability. Finally, the **prolongational reduction** analysis continues the categorization of nodes in the tree, this time in respect to tension/relaxation (by incorporating tonal knowledge).

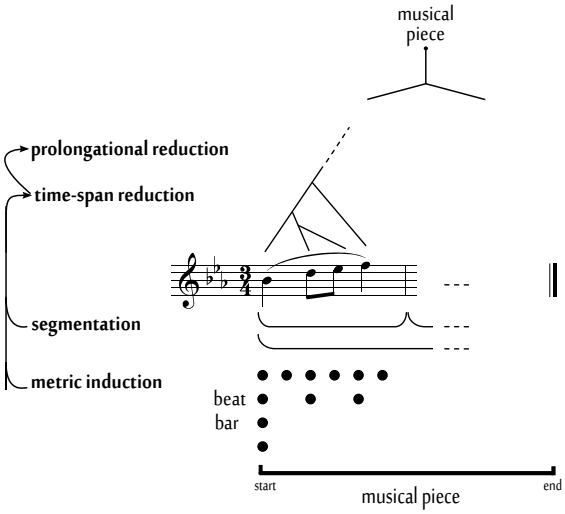

Figure 2.1: Music analysis depicting GTTM structuring processes (in bold). The analysis illustrates the types of mental representation resulting from each analysis and considers dependencies among processes as hypothesised in the GTTM.

Working Definitions

In what follows we establish working definitions for *segment*, *segment structure*, and *segmentation*. (It must be noted that in music separating the concepts of segment and segment structure is a catch-22 problem: since listeners do not normally hear segments in isolation, the aspects that define a segment are intrinsically linked to those of the structure they form, and it is hence impossible to define one without referencing the other.)

Segment. *A segment is a unit in a segment structure. Musical segments have two fundamental properties. First, they are bounded in time. Second, they are comparable.*

Segment Structure. *A segment structure is a mental representation of music. It consists of segments organised into either groups, chains, or holarchies.*

Segmentation. *A Segmentation is the process by which a segment structure is abstracted from auditory input.*

In our definition of structure we mention three likely types: groups, chain, and holarchies. A brief description of these types of structure is in place. Group structures represent situations where listeners relate segments, but fail to encode temporal location. Chain structures represent situations where listeners relate segments, encode temporal information, but only at one predominant time span. Holarchical structures represent situations where listeners perceive music as embedded chains of segments at multiple time scales, so that briefer ones are either approximately or exactly contained within larger ones. Holarchical structures are thought to be the most commonly perceived ones.

It seems reasonable to expect that the type of structure a listener constructs is dependent on a number of factors related to both the listener and the music, to name a few: listening mode (attentive or passive), music listening experience, familiarity with the style or genre of the piece, number of instruments, musical texture, and overall length of the piece. In exceptional cases the segment structure can be extremely detailed. W. A. Mozart, who was allegedly able to transcribe a whole mass from memory after just one listen (Gardner 2008, p. 55), would have probably been able to construct deep and highly optimised holarchic segment structures. The opposite extreme is an unengaged casual listener, which might fail to make associations in the music she/he is listening to, so that her/his experience might very well be a series of musical moments, with little relation to one another (hence closer to a group structure).

In this dissertation we assume an attentive listener able to maintain a segment structure that allows her/him to actively switch between a global view and local view of the structure. This would mean that, for instance, if a listener acquainted with the pop music genre is listening to a generic song, she/he would be expected to have an approximate sense of 'where' in the structure she/he is, e.g. 'the second or third repetition of the second chorus'.

2.2.2 A Segmenter as a Mechanism of Cognition

We assume that, when engaged in music listening, human listeners create and maintain a constantly-evolving segment structure. We posit that this structure is the result of a parallel processing mechanism, consisting of a cue detection system and a combination system – see Figure 2.2 for an illustration.

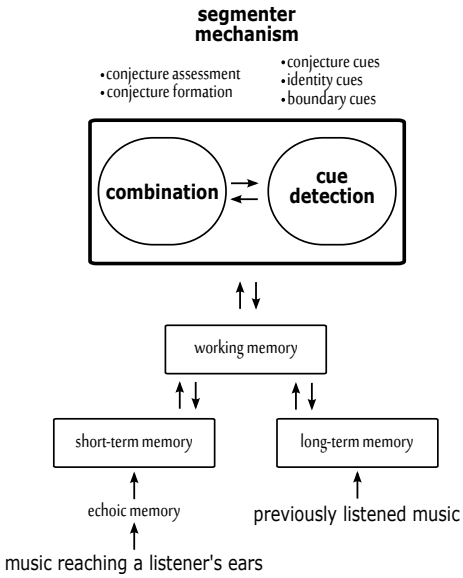

Figure 2.2: Segmentation during ongoing listening.

Cues inform different aspects of segments, e.g. where they start or end (*boundary cues*), or what their internal characteristics are, and how do they relate to other segments (*identity cues*). Cues also inform desirable/unwanted aspects of a segment structure, e.g. whether it fits a known structure, or if it has desirable conditions of symmetry (*conjecture cues*). Cues then have a dual functionality: in some cases they *suggest* segment or segment structure hypotheses (boundary/identity cues), and in others they *constraint* hypotheses (identity/conjecture cues).

The combinator consists of two processing modules: conjecture formation and conjecture assessment. The former is in charge of taking detected cues and generate a space of possible segment structure hypotheses. The latter is in charge of assessing which hypothesis(ses) might be more useful to the listeners, e.g. in terms of cognitive economy: which structure leads to the most parsimonious mental description of the music, or in terms of attention: which structure gives us the best description of that

which the listener wishes to focus on.

The two mentioned processing strategies to determine 'usefulness' are often called *perceptual* segmentation and *goal-oriented* segmentation, respectively (Gobet et al. 2001).[4] Perceptual segmentation is thought to be as an unconscious, fast-reaction process, assumed to be the result of lower level processing in the brain, hence mostly 'automated'. This implies a serial processing flow, i.e. first segments are bounded, then recognised as such, and lastly associated to other segments. Goal-oriented segmentation supports the idea that there is a deliberate, semi-conscious control of the segmentation process. This implies parallel (or feedback) processing, e.g. segment structure information may be used to define the identity of segments or detect their boundaries. While in the literature researchers often tend to side with only one of these processing strategies, in this dissertation we assume both goal-oriented and perceptual segmentation are possible.

2.3 Taxonomy of Segmentation Cues

In this section we introduce and describe our cue taxonomy. We focus on cues related to the 'music content', i.e. that relate to pitch, timbre, loudness, and so on.[5] In §2.3.1 we first collect a list of cues that have been observed in music psychology experiments, and then in §2.3.2 present the taxonomy.

2.3.1 Observed Cues

In the field of Music Psychology, researchers have investigated segmentation cues via listening experiments – refer to Appendix C for a review. The main idea behind the experiments is to have a group of participants listen to a number of music pieces or fragments, and ask them to indicate the location of boundaries while listening. To discern which cues the listeners might have used to perform the segmentation, these experiments follow one of three approaches:

1. *Test cues defined in segmentation theories.* That is, either generate or collect music where cues as defined in a known theory of segmentation can be observed, e.g. that contain a passage where *"in a sequence of four notes n1 n2 n3 n4, the transition n2-n3 marks a segment boundary if it has a greater intervallic distance than both n1 to n2 and n3 to*

[4] Also called bottom-up and top-down processes. We prefer the terms of Gobet et al. since we believe are more easily interpretable by non-specialist readers.

[5] For simplicity we do not consider non-musical factors. We ignore, for instance, the influence that linguistic factors might have on the segmentation of vocal melodies (e.g. word-level coarticulation, or the phrase and syntactic structure of text).

n4" - see Appendix Table B.1. Following this approach lists of tested cues have been provided by Deliège (1987) 1987; Ahlbäck (2004, pp. 409–449).

2. *Test segmentation principles from music theory.* That is, collect a sample of music pieces and have an expert analyst conduct a segmentation analysis of the pieces, indicating both boundaries and cues. Then have experiment subjects listen to the pieces and mark boundaries. Check the correlation between boundaries indicated by the expert and those indicated by the participants. Following this approach lists of cues have been provided by Spiro (2007, p. 356, 372).

3. *Let participants give a description of their strategies for segmenting.* That is, collect a sample of music pieces or fragments. Ask participants to mark boundaries. After the marking process is done, ask them to describe the cues that they used for the segmentation. Following this approach lists of cues have been provided by Bruderer (2008, pp. 121-132; Clarke and Krumhansl (1990, p. 243, 227).

Table 2.1 lists cues identified in these experiments. Due to our focus in melody, we list only cues observable in monophonic music. Moreover, we follow Spiro (ibid.) and use standard terminology from music theory to describe the cues.

Cue list
-Long note or rest, pitch jumps
-Change in dynamics, timbre, register, rhythm, motive, contour, meter, key, or tempo
-Consistency in dynamics, timbre, register, rhythm, motive, contour, meter, key, or tempo
-Exact or inexact repetitions
-Complete tonal motion, cadence preparation and completion, implicit harmonic progression
-Metrical accent: beat, bar, hypermetric
-Template form structure, recognition of stylistic motive or quotation

Table 2.1: List of segmentation cues tested/observed in monophony.

2.3.2 Taxonomy

Our taxonomy groups the cues listed in Table 2.1 into eight cue classes: *repetition, contrast, gap, alignment, closure, homogeneity, continuity,* and *template.* We illustrate our taxonomy in Figure 2.3, where cue classes are seen as specific instances of general cognitive processes, such as similarity processing and predictive processing, which in turn serve as input information to segmentation specific processes, such as detecting boundaries or segments. In the following subsections we describe each cue class in turn.

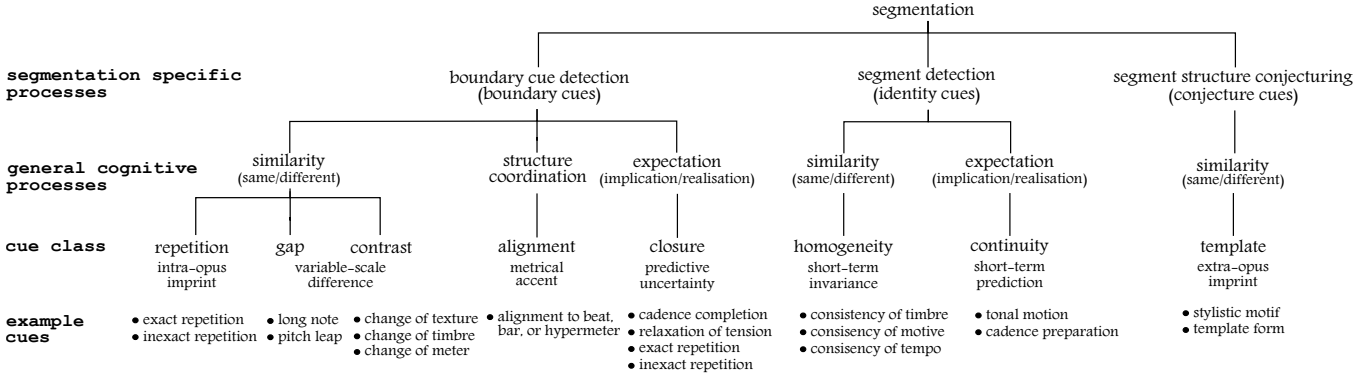

Figure 2.3: Segmentation cue taxonomy.

Note: Before moving onto the description of each class, it is important to stress that the processes outlined in Figure 2.3 operate in tandem, and that the output of one process is likely to be influenced by (or dependent on) the output of the other processes. The tree diagram employed to illustrate the taxonomy is used only for visualisation convenience, the independence relations it suggests should not be taken as meaningful.

Cues Related to Similarity Processing

The ability to make judgements about degrees of similarity is fundamental to cognition. Similarity processing in music, both within and across pieces, is an important and widely studied concept in the fields of Music Cognition and MIR (Toiviainen 2007, 2009; Margulis 2014; Hewlett and Selfridge-Field 1998; Volk et al. 2015). Many authors have acknowledged the importance of similarity in segmentation, for instance (Lerdahl and Jackendoff 1983; Cambouropoulos 2006; Deliège 2007; Ahlbäck 2007; Lartillot 2010). Among them, Deliège (2007) divides segmentation-guiding similarity judgements into *same* or *different*. Judgements of same are thought to enable listeners to 'cement' neighbouring segments and establish 'links' between temporally distant segments. Conversely, judgements of different are thought to enable listeners to 'demarcate' neighbouring segments. We have used similarity as a general class for the following cue classes: gap and contrast (different), repetition (same), homogeneity (same), and template (same).

Gap and Contrast. The gap and contrast classes contain cues such as change of timbre, pitch jumps, and so on. These cues are thought to suggest a transition between adjacent segments, and hence can be used by listeners to determine segment boundaries. We refer more generally to gaps and contrasts as *variable-scale differences*, given that changes seem to be perceived at variable (and on occasion multiple) time scales. Gaps are differences which can presumably be identified at a short time scale, not longer in length than that of a figure, where the listener's attention is assumed to be focused on note-to-note transitions. Gaps are generally associated to pitch jumps and long notes or rests. However, sudden changes in dynamics, timbre, or any other attribute that could be used to describe a note-like melodic event can be in principle considered a gap. Gaps are thought to suggest segment ends. Contrasts, on the other hand, are differences which presumably would require a longer time scale to be identified, where attention is focused on changes in the attributes describing sequences of notes. We associate contrasts to changes in, for instance, mode (e.g. major to minor or vise versa), melodic pitch contour, or motive. Contrasts are thought to suggest segment starts.

Homogeneity. The homogeneity class contains cues such as consistency of timbre,

tempo, or motive (caused by ostinato). We refer more generally to homogeneity as *temporal invariance* to stress that homogeneity, in a musical context, refers to time intervals where one or more musical attributes (such as the above mentioned tempo, timbre, and so on) change slowly or remain constant. Homogeneity seems to serve a dual purpose. On the one hand it seems to be a necessary condition so that temporally close musical events are perceived as part of the same whole, enabling listeners to form a nested segment structure (contiguous segments that share common characteristics are merged into larger segments). On the other hand it seems to enable the extraction of attributes that can be used to characterise segments, enabling listeners to create an identity for them. Homogeneity cues are hence thought to facilitate, and in cases enable, segment comparison and recognition.

Repetition. The repetition class contains cues such as exact or inexact repetitions of figures, phrases, or whole sections. These cues are thought to facilitate linking segments and thus are used by listeners to determine boundaries. We refer more generally to repetitions as *intra-opus imprint*. We do so to stress that (a) repetitions are taken to be identified from within the piece being listened to, and (b) their identification by humans requires concurrent formation of memory imprints and recognition of those imprints. Repetitions are most often thought to suggest segment starts.

Template. The template class contains cues such as stylistic motif recognition, quotation recognition, or segment structure recognition. These cues are thought to influence the organisation of segments into a segment structure, so that, from different segment structure hypotheses, listeners prefer those containing the recognised instances of motives or quotations, or prefer that segment structure which most resembles structures characteristic of the style or genre of the piece being heard (e.g. antecedent-consequent, strophic form, or 12-bar blues form). We refer more generally to the template class as *extra-opus imprint*, to stress that templates constitute a mapping of memory imprints from previously listened music onto the piece being listened to.

Cues Related to Predictive Processing

Huron (2006, p. 41) defines expectation as *"a form of mental or corporeal belief that some event or class of events is likely to happen in the future"*. Music expectation is an important and widely studied concept in the fields of Music Cognition and MIR (Meyer 1956; Narmour 1990; Huron 2006; Pearce and Wiggins 2006a; Abdallah and Plumbley 2009; Dubnov 2011). Many authors have acknowledged the importance of expectation in segmentation (Narmour 1990; Pearce and Wiggins 2006b; Wiggins and Forth 2015). Among them, Narmour (1990) suggests that *implicative* and *terminative* musical situations are specially important for segmentation. (Narmour uses

the term 'realisation' to refer to terminative situations.) Implicative situations enable listeners to make short term predictions of what might happen next. Conversely, terminative situations fail to stimulate predictions of continuation. The recognition of these situations seem to suggest to listeners whether the current listening point has a starting, middle, or ending quality. We have used expectation as a general class for two cue classes: closure (realisation) and continuity (implication).

Closure. The closure class contains cues such as cadence termination or recognition of figure/phrase/section ending. These cues seem to give the listener a sense of completion and finality, thus making it hard to predict what might follow. Closure cues are hence thought to suggest segment endings. We refer more generally to closure as *predictive uncertainty*, to stress that we do not refer only to *tonal* closure, but rather more generally to any disruption in an expectation process that might cause predictive uncertainty.

Continuation. The continuation class contains cues such as tonal motion and cadence preparation. These cues are thought to create in the listener an expectation of continuation, allowing him/her to formulate hypotheses about the music at the current listening point (e.g. whether it has a introductory or conclusive character), and make short-term predictions (e.g. estimate whether the end of a segment is approaching or not). Just like homogeneity, continuity seems to be a necessary condition so that neighbouring melodic events are perceived as part of the same whole. We refer more generally to closure as *short-term prediction*, to stress that predictions influencing segmentation seem to be limited to the immediate or short-term future.

Cues Related to Structure Coordination

As described in §2.2.1, theories of music perception posit that multiple mental structures are formed when listening, e.g. metric, harmonic, segmental, and so on. Even though these structures are normally studied separately (for simplicity), theories often stress that these structures are formed simultaneously. Hence during listening it is expected that the structures interact, altering one another's formation. We concentrate on the interaction between segment and metric structures.[6] We have used structure coordination as a general class for one cue class: alignment.

Alignment. The alignment class contains cues such as alignment to beats, bars, or hy-

[6] Metric structure refers to a multi-layered pattern of strong and weak accents in time. Metrical accents are organised hierarchically as either a subdivisions or a multiples of a central pulse called 'beat'. Metric structure is commonly notated using 'measures' or 'bars', containing a whole number of beats. The term 'hypermeter' is used to refer to metrical levels above the notated measure, containing some whole number of measures, usually between two and four (Love 2011).

permeter. In principle, segment and metrical structures are independent; segments do not necessarily begin or end on strong metrical positions – see (Lerdahl and Jackendoff 1983, p. 25–26) for a discussion. Nevertheless, it seems that segment and metrical structure tend to be roughly aligned. It has been shown that segment boundaries are likely to occur at beat positions (Palmer and Krumhansl 1987b,a; Stoffer 1985; Ahlbäck 2004, pp. 409–423), and also that strong metrical accents (on-beat and hypermeter) often occur near the beginning of phrases (Temperley 2003; Love 2011).

2.4 Melody and Melodic Segments

In §1.4 we restricted the scope of this dissertation to the segmentation of melody. In this section we provide working definitions for melody, define our assumptions as to how melody is mentally represented, and provide working definitions for the terms that we use to refer to melodic segments.

2.4.1 A Working Definition of Melody

In Western musicology, melody is considered one of the four basic 'materials' (Copland 1959) or 'ingredients' (Macpherson 1915) that composers use to create musical pieces (the other three being harmony, rhythm, and tone colour). Despite being a fundamental and ubiquitous concept in music, used by professional musicians and casual music listeners alike, melody is known for being notoriously difficult to formalise for scientific research – see for instance the discussion in (Salamon 2013, pp. 3–7). In this dissertation we use the definition provided by (Snyder 2000, p. 135):

Melody. *A melody is a temporal "sequence of acoustical events that contains recognisable patterns of contour ('highness' and 'lowness') ... with perceptible pitchlike intervals between successive events".*

We make three assumptions to complement and constraint this definition. Each assumption is described below.

1. The aforementioned acoustic events *can be approximately described by the music theoretic note*. Events are then primarily defined in respect to their pitch content, and have a duration best measured in seconds. These events are also assumed to be the primitives of cognitive representation, or, put another way, the elementary constituents used by human listeners in the conception of melodies. This assumption is not free of controversy – refer to Appendix B.2 for a discussion. That said, we believe that the melodies investigated in this dissertation are amenable to a note based representation, and thus overlook the controversy.

2. An event sequence generated by a single instrument, capable of producing only one pitched sound at a time, *leads to the perception of a single melody.* That is, we equate the notion of melody with that of *monophony.* This is not always the case, as human listeners can interpret a monophony as being comprised of multiple parallel melodies, or one melody plus an accompaniment. However, in this dissertation we take this simplifying assumption as true to make a clear distinction between segmentation and streaming.[7]

3. Event sequences classified by humans as melodies *give the perceptual impression of a connected and organised series.* Not any sequence of note-like events is perceived as a melody. If out previous assumption established that all melodies are monophonies, this assumption establishes that not all monophonies are melodies.

2.4.2 Working Definitions for Melodic Segments

From our third assumption above, we can expect that melodies have some form of syntax, which is reflected in (and at the same time motivates the need for) segment structure formation. Below we provide working definitions for the type of segments we expect a melody to have. We base our definitions in terminology from music theory.

Units of Musical Form

Form refers to *"how the various parts of a composition are arranged and ordered, ... how different sections of a work are organized into themes, and how the themes themselves break down into smaller phrases and motives"* (Caplin 1998, p. 9).[8] These sections, themes, phrases, and so on are often referred to as 'formal' (Copland 1959) or 'structural' (Stein 1979) units. Its commonly assumed that these units share some resemblance to segments. In Table 2.2 we collect definitions of various formal units relevant to this dissertation.

For relatively short melodies (30 notes or less), it is sensible to expect that melodic segments are conceived at least at two time spans (Ahlbäck 2007). We refer to these times spans as *figures* and *phrases*, due to their expected resemblance to these formal units. In longer melodies it would stand to reason to expect segments of a third, longer time span, which we refer to as *sections.*

[7] Segmentation focuses on the study of perceptually segregating sequential structural elements, such as phrases, sections, and so on. Streaming focuses on the perceptual segregation of structural elements that occur simultaneously in time, such as different voices of a polyphony, separating melody from accompaniment, and so on.

[8] The term form is also used as a class to categorise pieces of music by the way in which the main sections of the piece are arranged, such as 'ternary', 'cannon', or 'sonata'. In this dissertation we use the connotation of form as musical organisation and not as a categorical scheme.

Term	Definition
note	Basic unit of notation in Western music. Most traditionally it specifies one sound-producing action or gesture. It is minimally described in terms of duration and pitch. Ranges from a fraction of a second to several seconds (Roads 2001; Smalley 1997).
figure	Smallest musical unit with individual expressive meaning. Roughly 2-12 consecutive notes (Stein 1979, p. 3).
motive	Occasionally used as synonym of figure. Normally there is a distinction: the motive is a thematic particle (representative of the music) (Stein 1979, p. 3).
sub-phrase	Any unit smaller than a phrase (Rothstein 1989), similar in length to a figure.
phrase	Aggregation of consecutive notes encompassing a "substantial musical thought" (Benward and Saker 2008, p. 95). Roughly 2-4 to 8 measures in length (Temperley 2003).
section	Largest units of form. When use to classify whole pieces it is often labelled according to style into e.g. introduction, exposition, verse, chorus, refrain, conclusion, and so on.

Table 2.2: Definition of formal units relevant to this dissertation.

We selected these terms to classify melodic segments due to their alleged analytical value in form analysis. In most texts determining phrase units stands as particularly important. For instance, the phrase is often taken *"as the unit of measurement, a standard from which to base our consideration of other periods of varying length"* (Macpherson 1915, p. 13). Moreover, Stein (1979, p. 22) notes that the phrase constitutes *"the structural basis of compositions with homophonic forms ... most compositions with a predominant top-line melody may be divided into phrases"*.

Units such as figures and sections also play an important role in the analysis of form. For instance, Stein (1979, p. 25) notes that *"most vocal polyphonic forms and practically all imitative forms, both instrumental and vocal, are divided into sections rather than phrases"*. Also, in respect to figure level units he remarks *"the motive ... represent[s] the structural basis of the contrapuntal imitative forms such as the invention, fugue, or motet"* (ibid. 1979, p. 37).

The definitions given to formal units are either overly vague or overly reductive. (With the definition of phrase as appearing specially ambiguous.)[9] However, our interest in these units is not as a reference for specific definitions, but rather as a basis to broadly classify segment time spans (and so our focus is on approximate

[9] Lack of definition consensus in form analysis is a topic that could easily make for a whole journal article. However, for the sake of brevity, we do not elaborate on this issue and refer to Spiro (2003) for phrase term usage in classical music, and Attas (2011) for phrase term usage in popular music.

lengths, which in Table 2.2 are specified in notes). That said, we add some remarks to avoid terminological ambiguity in the following chapters.

1. In this dissertation the terms figures, phrases, and sections are used to denote cognitively relevant intervals of music. Therefore, to denote arbitrary time intervals, we use the terms *fragment* and *passage*.
2. We assume that in melodies for which a clear metric structure can be inferred, melodic segments are arranged in a hierarchical segment structure. So that briefer segments are exactly contained within longer ones. On the other hand, if for the listener it is not possible or difficult to determine a metric structure, the organisation of segments of different time scales will still convey the sensation of 'nestedness', but this sensation will be more free, subject to interpretation. In other words, the segment structure will be holarchical, so that segments across time scales are not necessarily aligned.
3. We assume phrases have a regulatory effect in segment structure, i.e. the length of the phrase regulates the length of longer and shorter segments.
4. We assume phrases represent a sort of perceptual present, so that their maximum length is restricted by the storage limitations of working memory. This assumption is made in many theories of music segmentation, see for instance (Snyder 2000, pp. 59-60). Note that we do not mean to imply that listeners can not perceive phrases longer than the average working memory span, rather just state that listeners would have a harder time processing these phrases, so that working memory limitations can serve as a regulator of expected phrase length.

2.5 A Review of Machine Melody Segmentation

In this section we review approaches to machine melody segmentation. In §2.5.1 the prototypical stages of a melody segmenter are first described. In §2.5.2 approaches are presented and existing melody segmenters are classified.

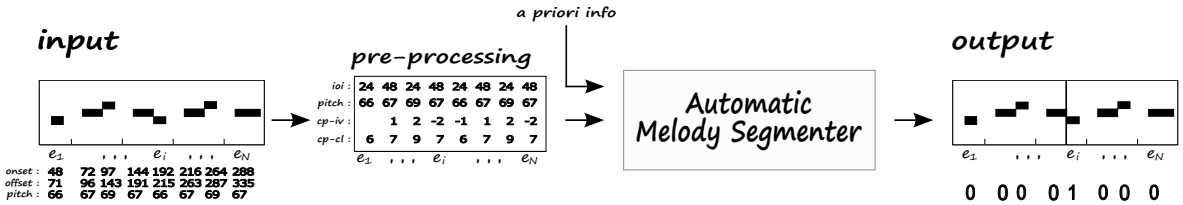

Figure 2.4: Input-output diagram of a prototypical computational model of melody segmentation. **From left to right**: melody in piano-roll representation, attribute profiles, segmenter, boundary list where a 1 represents a segment border. For abbreviations refer to Table 2.3.

2.5.1 I|O Description of a Prototypical Melody Segmenter

Figure 2.4 illustrates the pipeline of a prototypical automatic melody segmenter. Below we first describe the formalisation the melody segmentation task, review common architectures, and then describe each component of a prototypical segmenter in detail.

Task

Melody segmentation is formalised as a sequence segmentation problem: given a sequence of non-overlapping ordered events $e = e_1 e_2 \dots e_n$ a segmentation s of e is defined by $k+1$ boundary locations $1 = b_1 < b_2 < \dots < b_k < b_{k+1} = n+1$, yielding the sequence of segments $s = s_1 s_2 \dots s_k$ where $s_j = [e_{b_j} \dots e_{b_{j+1}}), \forall j = 1, \dots, k$. That is, the segmentation s partitions the sequence e into contiguous segments so that each event belongs to exactly one segment.

For a segmentation s we assume $|s_j| > 0, \forall j = 1, \dots, k$. We consider boundaries b_1 and b_{k+1} to be *trivial case boundaries*. Likewise, we consider a segmentation consisting of only one segment $|s_1| = n$ or of n segments each of $|s_j| = 1$ to be *trivial case segmentations*.

Common Architectures

Machine melody segmenters have, in most cases, single input – multiple output architectures. That is, they are meant to process only one melody at a time, and can be used to produce a single list of boundary locations, a ranked set of lists (each reflecting a different segmentation), or a set of lists reflecting different possible interpretations of segment structure. A few multiple input cases exist. For instance, in (Rafael and Oertl 2010) the segmenter introduced takes multiple instrumental parts as input, automatically reduces polyphonic parts to monophonies, and then integrates the segmentations of different parts into a single segmentation.

In this dissertation we evaluate segmenters of unaccompanied melodies, by comparing machine segmentations of to manual segmentations of such melodies. In most cases only a single reference manual segmentation is available, which moreover contains boundaries for one type of segments (often phrases). Thus, we focus on single input – single output architectures.

Melody representation

Two digital format categories exist for melody storage: *audio* and *symbolic*. Audio formats (e.g. a WAV file) store recordings of sound generated during performance. If a melody segmenter takes as input audio, then the sequence of events e is obtained by sampling the waveform file. That is, the recording is divided into a number of equal-length fragments, often 10 msec. This temporal resolution ensures that boundary locations can be specified with enough precision. Each fragment is then analysed to extract quantifiable descriptors (e.g. RMS energy, mel-frequency cepstral coefficients, chroma) that have been shown to correlate to musically-relevant perceptual attributes of sound (e.g. loudness, or timbre, harmony).

Existing melody segmenters most frequently process melodies in symbolic format (e.g. `MIDI` or `**kern`) Symbolic formats (e.g. MIDI) encode music as a sequence of events, similar to the music theoretic note. Melodies in symbolic format have been either encoded manually using a score editor, or recorded by performing using a digital interface (e.g. a MIDI keyboard). Input to a melody segmenter most often consists of a sequence of events e, where each event corresponds to a {`pitch`, `onset-time`} pair, or a {`pitch`, `onset-time`, `offset-time`} triplet. This reduced representation of notes is commonly referred to as 'piano roll' – see Figure 2.4 (far-left) for an illustration. Pitch is often represented as a number, usually a MIDI number, and an onset/offset is given by a real number representing a point in time. (A fair amount of segmenters require onset/offset to be 'quantised', i.e. aligned to metric units.)

Pre-processing

Input melodies are often transformed into a set of attribute sequences, one for each attribute considered relevant for computing the segmentation – see Figure 2.4 (centre-left) for an illustration. In this dissertation our segmenters use (subsets of) the attributes listed in Table 2.3. In the table attributes readily available from the melody encoding are referred to as 'basic'. Conversely, attributes computed from basic attributes are referred to as 'derived'. Formulas for the computation of derived attributes are provided in Appendix A. We limit ourselves to the use of attributes in Table 2.3 for pragmatic reasons. That is, other attributes that might be relevant are left out because they are either not included in the encoding format (e.g. timbre), or can not (at present) be automatically estimated with enough reliability (e.g. key or metric bars). The use of a specific subset of the attributes in Table 2.3 is motivated separately for each segmenter.

Class		Abbreviation	Description
basic		`on`	note onset time
		`of`	note offset time
		`cp`	note chromatic pitch
derived	rhythm	`rest`	duration of a musical rest: offset-to-onset-interval
		`ooi`	note duration$_1$: onset-to-offset interval
		`ioi`	note duration$_2$: inter-onset-interval
		`ioi-r`	`ioi` ratio of e_i relative to e_{i-1}
		`ioi-rc`	contour of `ioi-r`$_i$ relative to `dur-rat`$_{i-1}$ (longer, same, shorter)
		`ooi-r`	duration ratio of e_i relative to e_{i-1}
		`ooi-rc`	contour of `dur-r`$_i$ relative to `dur-rat`$_{i-1}$ (longer, same, shorter)
	pitch	`cp-cl`	pitch class: chromatic pitch interval under octave equivalence
		`cp-iv`	chromatic pitch interval of e_i relative to e_{i-1}
		`sl-iv`	"step-leap" (step ($\pm s$), a leap ($\pm l$), or a unison ($\pm u$)) classification of `cp-iv`
		`ct-iv`	pitch contour of `cp-iv`$_i$ relative to `cp-iv`$_{i-1}$ (up, down, same)
		`cp-rc`	chromatic pitch register class

Table 2.3: List of attributes used to describe melodic events in this dissertation. Basic attributes describe melodic events consisting of a single note. Derived attributes describe melodic events consisting of one or multiple notes. The horizontal line separating the attribute list of the rhythm class discriminates *absolute* from *relative* attributes. All attributes in the derived pitch class are *relative*.

2.5.2 Approaches To Machine Melody Segmentation

Existing melody segmenters define segmentation criteria by proposing a model of one or multiple segmentation cues, and then use these criteria to segment the input melody. Segmenters operate using predominantly one of two approaches: event classification or segmentation scoring. Below we describe each approach in turn.

Event Classification. The event classification approach aims to classify each melodic event as boundary or not-boundary. Event classification can be said to model more closely the process of boundary cue detection. Segmenters following this approach compute a score for each event location in the melody, resulting in what can be described as a 'boundary strength profile' of the melody. The interpretation of the boundary strength profile depends on the segmentation criteria used and it is thus segmenter specific. As an example, in Camboroupoulus' LBDM (2001), score values are proportional to the estimated perceptual salience of gap cues, while in the case of Pearce's IDyOM (2008) score values are proportional to the estimated perceptual salience of closure cues. Segmenters using the event classification approach require a post processing algorithm to select boundaries from the boundary strength profile. This is commonly implemented using heuristic peak selection methods.

Segmentation Scoring. The segmentation scoring approach aims to produce many pos-

sible segmentations of the input, and evaluate or score each of them to find the most suitable one. Segmentation scoring can be said to model more closely the process of segment structure conjecturing. As an example, Temperley's GROUPER (2001, Ch. 3) scores segmentations by considering alignment and template cues. More specifically, GROUPER select the best segmentation as that which is more congruent with metric structure (alignment cue), and also where each segment shows the minimum deviation from and ideal segment length (template cue). The advantage of the segmentation scoring approach over the event labelling approach is that inserting a segment boundary has an impact on the score of all other boundaries. A major disadvantage is that there are 2^{n-1} possible segmentations of a melody of length n, i.e. the space of possible segmentations is exponential in the number of events of the melody, which makes computation time and space requirements restrictive for all but very short melodies. For this reason, segmenters following the scoring approach often attempt to compute a subset of the space of all segmentations which is assumed to contain the 'correct' segmentation. For instance, GROUPER restricts possible segmentations to those whose boundaries are cued by temporal gaps.

Less Common Approaches

Another approach to segmentation is *event clustering*. This approach can be said to more closely model homogeneity cues. Several event clustering segmentes have been proposed in for the segmentation of music recordings – see (Paulus et al. 2010) for a review. They predominantly used some form of HMM based technique to model homogeneity. The event clustering approach was popular in the mid 2000s. However, it has since then fallen out of favour. The main reason is their intrinsic neglect for temporal information (i.e. the techniques used see music as a 'bag of features'). It was observed that this neglect made the approach very prone to over-segmentation. Many efforts were made to incorporate the temporality of music in the processing chain, but results did not improve significantly. In the field of melody segmentation very few approaches focus on modelling homogeneity cues. In fact, we are aware of only two (Thornton 2011; Lartillot and Ayari 2014). The former is a model of melody compression (conceptually similar to the time span reductions of the GTTM). It models homogeneity cues indirectly. Despite perceiving melodic temporality, its reported performance also suggest pronounced over-segmentation problems. The latter operates over a duration-only representation of the melody. It processes melodies sequentially (hence preserving temporality), and uses heuristics to determine whether two neigbouring events should be clustered or not. The segmenter has, however, not been systematically assessed.

2.5.3 Classification of Machine Segmenters

Figure 2.4 shows our classification of machine segmenters. All segmenters classified were developed to take as input music in symbolic input.[10] The large majority of them accepts monophony as input – exceptions are (Rowe 1992; Chew 2006; Zanette 2007). Segmenters are classified in respect to the cue(s) being modelled (from our taxonomy in Figure 2.3) and the approach used (from those discussed in §2.5.2). We also include information of the main technique used. However, at this stage of the dissertation, we avoid a description of the techniques employed, as we believe these make otherwise closely related segmenters appear as incompatible. We review subsets of these segmenters in the following chapters, where their technical aspects are discussed in more detail.

Multi-cue segmenter development seems to have been popular during the early 1990s, but research efforts since have moved to developing segmenters modelling single cues. Moreover, the vast majority of these segmenters have focused on gap cues, using an event classification scheme. There are a fair amount of repetition-based melody segmenters, however, most of them have not been systematically tested, and reported case study results suggest low performances can be expected. Closure cues have become a popular area of research since approximately the mid-2000s. Yet, just as with repetition cues, most closure-based segmenters have not been systematically tested. Contrast cues and template cues have not received much research attention.

In this dissertation we set out to provide a computational implementation of our conceptual model. To that end we first focus on modelling three cues which have either received little research attention, or have not been tested systematically, namely *repetition*, *contrasts*, and *template* cues. In Chapters 4, 6, and 6 we develop/expand and test single-cue segmenters using these cues. Then, in Chapter 7, we introduce a multi-cue framework that combines our proposed single-cue segmenters.

2.6 Conclusions

In this chapter we introduce a cognitively-conceptual framework to guide the development of machine segmenters. The framework is composed of a conceptual model and a taxonomy. Moreover, we provide working definitions of segments and segment types, and define computational modelling tasks. The conceptual framework is used to classify existing segmenters. From the classification we can observe that most re-

[10] For more extensive surveys refer to (Rodríguez-López and Volk 2012) for segmenters operating over symbolic formats, and to (Paulus et al. 2010) for segmenters opiating over audio formats

Author	Cues	Approach	Dominating Technique
(Tenney and Polansky 1980)	{G}	EC	distance measure
(Baker 1989b)	{G,R}	SS	context free grammar
(Baker 1989a)	{G,R,T}	SS	frames
(Camilleri et al. 1990)	{G,R}	SS	expert system
(Rowe 1992)	{G,-}	EC	expert system
(Large et al. 1995)	{-}	–	recursive auto-associative memory
(Cambouropoulos 1997b, 2001)	{G}	EC	distance, similarity measure
(Friberg et al. 1998)	{-}	SS	context free grammar, neural networks
(Takasu et al. 1999)	{G,R}	EC	similarity measure, grammar
(Lefkowitz and Taavola 2000)	{G}	EC	distance measure
(Temperley 2001)	{G,T,A}	SS	distance measure, dynamic programming
(Bod 2001, 2002)	{T}	SS	probabilistic grammars
(Weyde 2001, 2002)	{G,H}	SS	fuzzy neural networks
(Ferrand et al. 2003a)	{G,C_{st}}	EC	distance measure
(Ferrand et al. 2003b)	{C_{re}}	EC	information theory
(Harford 2003, 2006)	{-}	–	self organising maps
(Juhász 2004)	{T}	SS	information theory, optimisation
(Frankland et al. 2004)	{G}	EC	distance measures
(Ahlbäck 2004, 2007)	{G,R}	–	–
(Cambouropoulos 2004; 2006)	{R}	EC	string search
(Hamanaka et al. 2004; 2005; 2006)	{G,R,T}	EC, SS	distance measures
(Chew 2006)	{C_{st}}	EC	distance measure
(Pearce et al. 2006b; 2007; 2008)	{C_{re}}	EC	markov models, information theory
(Dubnov 2006)	{C_{re}}	EC	information theory
(Zanette 2007)	{C_{st}}	EC	probabilistic distance measure
(Wilder 2008)	{G}	EC	distance measure
(Rafael et al. 2009)	{R}	SS	genetic algorithms
(Abdallah and Plumbley 2009)	{C_{re}}	EC	information theory
(Cox 2010)	{C_{re}}	EC	recurrent neural networks
(Rafael and Oertl 2010)	{R}	EC	string search
(Thornton 2011)	{H}	ET	Bayes transform
(Wołkowicz 2013)	{R}	EC	similarity matrix
(Velarde et al. 2013)	{G,C_{st}}	EC	Haar wavelet
(Bozkurt et al. 2014)	{G}	EC	supervised learning
(Lartillot and Ayari 2014)	{H}	ET	rule-based clustering
(Lattner et al. 2015a; 2015b)	{C_{re}}	EC	Boltzmann machines

Table 2.4: Machine segmenters for music in symbolic format. Cue: R - repetition; G - gap; C_{st} - contrast; A - alignment; C_{re} - closure; H - homogeneity; C_{ty} - continuity; T - template. Approach: EC- event classification; SS- segmentation scoring; ET- event clustering.

search work is comprised of single-cue segmenters, and that most of them have focused in modelling gap cues. We hence find it necessary to develop/extend approaches to repetition, contrast, template, and multi-cue segmentation.

Chapter 3

Evaluation of Machine Melody Segmenters

In this chapter we discuss methodologies, corpora, and measures used in MIR to evaluate melody segmenters.

Chapter Contributions We critique and propose ways to improve the evaluation methodology currently at MIR. We motivate and study new quantitative evaluation measures. We introduce a new test dataset, consisting of 125 Jazz melodies.

This chapter is based on work presented in (Rodríguez-López and Volk 2013, 2015b; Rodríguez-López et al. 2015).

3.1 Introduction

Evaluating machine segmenters is highly non-trivial. At present current evaluation approaches do not allow generalisation, nor a clear interpretation of the evaluation results, making it complex to establish a meaningful and reliable comparison between different machine segmenters.

In this chapter we first discuss difficulties with the evaluation of machine segmenters, and suggest guidelines to overcome these difficulties. We then focus on a specific problem: the evaluation of boundary detection. We discuss current evaluation measures and describe their short comings. We then suggest a new evaluation measure, proposed in the field of text segmentation, which can deal with segment perception ambiguity better than traditional measures. Finally, we describe and analyse an novel benchmark corpus consisting of 125 jazz melodies.

3.2 Machine Segmenter Evaluation in MIR and CMMC

In this section we critique the way machine segmenters are evaluated in MIR. Our critique focuses on the approaches used to set-up evaluation experiments. Hence, we first discuss desirable properties of an evaluation, then describe currently used approaches, discuss their limitations, and finish with a list of suggestions to tackle these limitations.

3.2.1 Desirable Properties of an Evaluation of Machine Segmenters

From the perspective of MIR, a satisfactory evaluation should enable the characterisation of machine segmenters, i.e. reveal what the strengths and weaknesses of the evaluated segmenters are, what their best context of use is, and so on.[11] Hence, three desirable properties of an evaluation are:

- *Generality*: Evaluation results should be generalisable to music of different styles, as well as to different tasks or contexts in which the segmeter is used. Evaluation results should also provide insights as to how robust segmenters are to deformations of the input.

[11] In MIR machine segmenter evaluation is not concerned with the full psychological validation of machine segmenters, but it is rather taken as the starting point for a more general verification – see (Honing 2006) for a discussion on the requirements for computational model validation in the field of Music Cognition.

- *Interpretability*: Evaluation results should allow to unambiguously compare and rank different segmenters (or different configurations of a single segmenter).
- *Reproducibility*: Evaluation results should be reproducible.

3.2.2 Current Approaches to Machine Segmenter Evaluation

In MIR the goal of an evaluation is normally to benchmark a novel segmenter (or part of it) against existing segmenters and/or baseline segmenters. This requires defining a way to assess the 'quality' of a segmenter. The ISO/IEC standard (2001; 2003) defines three ways to assess software quality: *internal*, *external*, and *in-use*. Internal quality refers to aspects of the software that can be assessed without running it, such as time/space complexity, maintainability, and so on. External quality refers to aspects of the software that can be assessed by running it as a 'black box', using quantitative measures to estimate its precision, parameter sensitivity, and so on. In-use quality refers to aspects of the software that can be assessed via interaction with a human user, such as learnability, satisfaction, and so on.

Machine segmenter evaluations have mostly focused on assessing external quality. This has been done via one of two approaches: *reference-based* or *task-based*. Below we briefly describe each approach in turn.

In reference-based evaluation automatically produced segmentations are compared to a sample of acceptable, manually produced segmentations. Quantitative measures, often defined in terms of (quasi) distances, are used to measure the similarity between automatic and manual segmentations. Segmenter quality is then assessed in terms of the cognitive plausibility of its output. That is, the more the automatic segmentation deviates from the manual one, the less cognitively plausible it is assumed to be, and hence the lower the quality of the segmenter which produced it is.

In task-based evaluation automatically produced segmentations are used within a system performing other music processing tasks, such as classification or retrieval. Quantitative measures are used to assess the system's performance. Segmenter quality is then assessed in terms of task relevance. That is, high quality segmenters are those which produce segmentations that lead to improvements in the system's performance.

For both reference and task based approaches, performance scores obtained for each segmented piece are averaged, and statistical tests are used to check if the differences between performance means are significant.

Evaluations of melody segmenters are in most cases reference based – see (Thom et al.

2002; Wiering et al. 2009; Pearce et al. 2010a,b). This is also the case for evaluations of polyphonic audio segmenters – see (Paulus et al. 2010; Ehmann et al. 2011; Smith and Chew 2013a).

Audio segmenters have most often been evaluated in the subtasks of segment boundaries and segment labelling – see Figure 1.1. Recently some attempts to evaluate nested structure have been proposed (McFee et al. 2015). Melody segmenters have focused on the task of boundary detection.

3.2.3 Limitations of the Evaluation of Melody Segmenters

At present neither reference-based nor task-based approaches can, on their own, ensure an evaluation of melody segmenters with high generality and interpretability. Below we describe, in turn, the reasons why this is so.

Reference-based evaluation is limited by data sparsity. Ideally, test corpora should be comprised of a large number of melodies belonging to multiple styles and traditions. An also large number of manual segmentations per melody should be available. In actual fact, large and publicly available test corpora have (a) *low stylistic diversity* and (b) most often *a single manual segmentation*. In respect to point (a), corpora used to evaluate melody segmenters are most often comprised of vocal folk music. Both the physiology of vocal sound production and the stylistic traits of folk music are likely to influence segmentation, compromising the generality of the evaluation results. In respect to point (b), having only a single (or few) reference segmentation(s) makes it unfesible to estimate with enough reliability whether test melodies have one or more cognitively plausible segmentations. Hence, low scoring segmentations might not necessarily be cognitively implausible. This uncertainty compromises the interpretability of the evaluation results, making it impossible to reliably compare different machine segmenters.

Task-based evaluation is limited by its inherent lack of experimental control. Estimating the quality of a segmenter by its role in a larger music processing system makes the evaluation very sensitive to (a) *biases related to the system's architecture*, and (b) *artefacts that lead to determine spurious causal relations between segmentation and system performance* (often called 'confounding factors'). In respect to point (a), the system might inherently favour some segmentations over others. Thus, ranking results are heavily dependent on the system's components and architecture, decreasing the generality of the evaluation. This limitation motivates testing various architectures, components, and parametric settings, to have greater generalisation power. However, and in respect to point (b), the more components the system has (or the more systems

are tested), the harder it becomes to control for confounding factors, and hence the lower the interpretability of the evaluation.

Suggestions to Improve the Evaluation of Melody Segmenters

In scientific research normally evaluation interpretability is preferered over generality. Hence, reference-based evaluation is often preferred to task-based evaluation. In this section we focus on providing suggestions to tackle the main limitation of reference-based evaluation: data sparsity.

An obvious solution to the data sparsity problem is to develop new manually segmented corpora. However, manual segmentation is a time consuming and error-prone process. For one thing, it requires attentive, repeated listening, which can be tiring for annotators and implies that the annotation process is always longer than the duration of the melody. For another, segmentation is a task human listeners often perform in an unaware, unconscious fashion. Hence, it is non-trivial to communicate to annotators what to do, which often results in unwanted errors (e.g. the annotator did not understand the task) or biases (e.g. the experimenter over explained the task).

The complexity of segment annotation makes unrealistic to expect a substantial increase in the number and quality of new annotated corpora, at least in the short term. Thus, alternative ways to deal with data sparsity issues are needed. We suggest two strategies. First, to focus on the refinement of existing manually segmented corpora. Second, to develop evaluation frameworks that combine reference-based and task-based approaches. Below we describe each strategy in turn.

One way to refine existing corpora is to manually add extra information to its segment annotations. For instance, segment boundary annotations could be refined by adding human judgements of boundary 'confidence'. That is, have human annotators rate how strongly they agree with the location of segment boundaries present in available corpora. Annotating boundary confidence has the advantage of being easy to explain and fast to produce. Boundary confidence information can be used develop quantitative measures of performance that are more interpretable. For example, it can be used to down-weight penalties for 'missing' low confidence boundaries.

Another way to refine existing corpora is to develop methods to automatically characterise annotated segments. For instance, Smith and Chew (2013b) analysed manual segmentation in a large segment-annotated corpus of polyphonic music recordings. (This corpus contains segment boundaries and equivalent-class labels annotated by three human listeners.) Smith and Chew used an optimisation technique to estim-

ate which of five musical attributes the annotators where attending when segmenting. This analysis technique can be used to characterise annotations and inform segmenter evaluations. For example, estimated cue relevance can be used to identify pieces that should be excluded from the evaluation (take a case where a given melody was annotated attending primarily to cue 'A', and a given automatic segmenter aims to detect only cues of type 'B').

Lastly, there is a need for evaluation frameworks that combine task-based with reference-based approaches. For instance, lets take a case where three machine segmenters s1, s2, and s3 score 0.6, 0.58, and 0.48, respectively (using some measure where 1 indicates perfect match to the reference segmentation). The scoring suggests that s3 performs much worse than s1 and s2. However, as discussed previously, the low score of s3 does not necessarily indicate that its segmentation is cognitively implausible. A task-based evaluation could be used to support or refute the ranking and apparent segmentation quality difference between segmenters, compromising generality in favour of interpretability.

3.3 Performance Measures in Segment Boundary Detection

In this section we focus on reference-based evaluation of segment boundary detection. We first outline two issues that arise from the subjectivity of boundary perception when evaluating machine segmenters. We then review the traditional measures used to evaluate boundary detection performance, and summarise their short-comings in dealing with issues related to boundary perception subjectivity. Finally, we propose the use of a new measure that ameliorates some of the short-comings of traditional measures.

3.3.1 Segment Boundary Perception in Melodies

Listening studies point to two aspects of segment boundary perception that should be taken into consideration when evaluating machine segmenters: the possibility of *fuzzy boundaries* and of *multiple segmentations*. Below we address each aspect in turn.

Fuzzy Segment Boundaries

Segment boundary annotation studies (with human subjects) have shown instances when the locations of segment boundaries are not clear cut but rather flexible. We refer to these segments as having 'fuzzy' boundaries. Fuzzy boundaries might be the

result of ornamentation (e.g. appoggiatura or mordents), or the relationship between metric structure and segment structure (e.g. anacrusis). In this situations listeners fail to reach consensus as to whether a given mordent or anacrusis is part of one segment or of the subsequent one. Fuzzy boundaries might also be the result of segmentation cues that differ at a local level (e.g. a repetition that occurs near a gap).

Multiple Cognitively Plausible Segmentations

Listener studies have provided evidence that human listeners might also assign different segmentations to melodies. (That is, instances where not only local disagreements between boundaries are observed, but also the total number of boundaries between annotators differs.) We hypothesise that these differences are mainly do to two reasons. First, the annotators have an inherently different concept of the length of segments. Second, the annotators might be attending to different segmentation cues.

Full and Near Misses

As discussed previously in this chapter, the problem of the uncertainty of whether a melody has one or more cognitively plausible segmentations is largely ill-defined. We proposed evaluation frameworks that could be used to mitigate the problem. In the rest of this section we focus on the problem of fuzzy segment boundaries.

3.3.2 Traditional Measures of Performance

Segment boundary detection is generally evaluated in a reference-based scenario, where automatically identified boundaries are compared to manually identified boundaries.[12] Automatically and manually identified boundaries need to be made comparable, and so both are encoded, respectively, as binary vectors $\mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{m} = (m_1, \ldots, m_n)$, where $a_i, m_i \in \{0, 1\}, \forall i = 1, ..., n$. Vector element positions represent potential-boundary-locations,[13] a 1 encodes boundary presence, and a 0

[12] The process of manual boundary identification requires human annotators to listen to a piece or fragment of music, and 'mark' the time points where they believe segments have finished/begun. The marking process can be actual or notional. Actual marking refers to the case when boundaries are identified by marking a visual (waveform, score, other) depiction of the music. Notional marking refers to the case when no visual aid is provided. Manually identified boundaries are stored as time stamps.

[13] Potential boundary locations can be absolute time windows, note positions or beats. When sengmenting music recordings often time windows ($\sim$100 msec) or beats are used. When segmenting symbolically represented music often note positions or beats are used. Melody segmenter evaluations have used note positions.

encodes boundary absence.

Once the binary encoding procedure is carried out, the most common evaluation strategy is to first check for boundary misplacement, and then use misplacement information to compute the similarity between **a** and **m**. A value of 0 should reflect that all boundaries in **a** are misplaced by comparison to **m**, and a value of 1 should reflect that all boundaries in **a** perfectly coincide with those of **m**.

Boundary misplacement is viewed as a classification problem. That is, taking **a** and **m**, each pair of corresponding vector elements is classified as either a true positive tp ($a_i = 1 \wedge m_i = 1$), true negative tn ($a_i = 0 \wedge m_i = 0$), false positive fp ($a_i = 1 \wedge m_i = 0$), or false negative fn ($a_i = 0 \wedge m_i = 1$). Then, the similarity between **a** and **m** is most often computed using the F_β measure (with $\beta = 1$)

$$F_\beta = \frac{(1+\beta^2) \cdot P \cdot R}{(\beta^2 \cdot P) + R} \in [0, 1], \tag{3.1}$$

where Precision P and Recall R are defined as

$$P = \frac{TP}{TP + FP}, \tag{3.2}$$

$$R = \frac{TP}{TP + FN}, \tag{3.3}$$

and TP, FP, and FN correspond, respectively, to the total number of tp, fp, and fn.

Benefits of the F_1, *Precision*, and *Recall* measures: Quantifying binary vector similarity using the F_1, P, and R measures has the benefit of not considering information on true negatives, which due to the strongly unequal proportions of boundary presence/absence values in music segmentation data would result in biased performance estimates.[14] Moreover, the P and R measures allow two interpretations of boundary misplacing: 'over-segmentation', i.e. introducing too many spurious boundaries (high R, low P), and 'under-segmentation', i.e. missing too many annotated

[14] Segment boundaries are sparse. For example,in (Pearce et al. 2010a) is indicated that in a melodic dataset adding up to ∼79000 notes, only about 12% of the note locations correspond to phrase level segment boundaries. Thus, standard evaluation measures in information retrieval using TN information, such as $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, would result in a biased assessment value. For instance, if a manual segmentation for a piece marks 20% of possible-boundary-locations with boundary presence, a naïve automatic segmentation predicting only boundary absences (an all-zero vector) would still receive an accuracy score of 80%.

boundaries (high P, low R).

3.3.3 Issues When Penalising Near Misses

The most common strategy to handle the near miss problem is to allow for a small tolerance δ when determining boundary matches.

In an ideal situation, a significant number of human listeners would have annotated the pieces, and this would allow to compute distributions of possible boundary locations. These distributions could then be used to estimate how large δ should be, and what score should be awarded to near misses. Some measures that formalise these ideas have been proposed – see (Melucci and Orio 2002; Spevak et al. 2002). However, as mentioned earlier at present large benchmark datasets for segmentation have been annotated by at most three human listeners, which impedes a reliable estimation of boundary location distributions.

At present, δ is most often set according to intuition. In the MIREX Structural Segmentation track (audio input) two tolerance settings have been used: narrow $\delta = \pm 0.5$ seconds and broad $\delta = \pm 3$ seconds. In comparative studies of melody segmenters (symbolic input) three tolerance settings have been used: no tolerance $\delta = 0$, narrow $\delta = \pm 1$ note events, and broad $\delta = \pm 2$ note events. No partial score is awarded to near misses, i.e. if the automatically determined boundary falls within the interval set for δ then it is classified as a true positive, otherwise it is classified as a false positive. Not awarding partial scores to near misses implies that narrow tolerance intervals might result in overly pessimistic performance estimates, while broad tolerance intervals might result in overly optimistic estimates. These inaccurate estimates complicate the interpretation of the 'true' performance of a machine segmenter, directly affecting the ranking of the segmenters participating in the evaluation. Additionally, inaccurate estimates might also affect subsequent analyses of performance, such as correlation analyses or outlier analyses.

Alternative performance measures and near misses: In MIREX the $mt2g$ measure has been used as an alternative to evaluate boundary detection performance. The $mt2g$ computes the median distance from each annotated boundary to the nearest predicted boundary. The $mt2g$ can be interpreted in terms of Recall (a high score corresponds to low Recall), and can also be seen to provide a rough account of near misses (a low score indicates a dominance of close near misses). However, assessing the influence of near misses on boundary detection performance can only be achieved indirectly, i.e. by cross-analysing F_1 and $mt2g$ scores, which makes the analysis complex and ultimately unreliable.

Other measures have been tested to complement/replace the F_1, Precision, and Recall measures, such as the *kappa statistic* and the *sensitivity index* d' (Pearce et al. 2010a), and also the $1-f$, $1-m$, $mg2t$, and $mt2g$ measures (Smith and Chew 2013a). However, aside from the previously discussed $mt2g$, none of the measures takes into account near misses.

3.3.4 New Measures of Performance that Account for Near Misses

In this dissertation we complement the use of the F1 measure with the *Boundary Edit Distance based boundary Similarity* (BED-S) proposed by Fournier (2013b) in the field of text segmentation. BED-S is an improvement upon the state of the art in boundary detection evaluation of text segmenters.

BED-S models the problem of identifying misplaced boundaries as an alignment problem. To that end Fournier introduces a new edit distance called *boundary edit distance* (BED), which differentiates between full and near misses between **a** and **m**. BED uses two main edit operations to model boundary misplacements: additions/deletions (A) for full misses, and n-wise transpositions (T) for near misses.

BED is based on the Darmeau-Levenshtein edit distance, which formalises A and T operations. An A type operation is a single-unit edit, which as seen in Figure 3.1 can correspond to either a false positive or a false negative. A T type operation is an adjacent-unit edit, i.e. the act of swapping one unit in a sequence with adjacent units (e.g. the sequence of characters 'ab' becomes 'ba'). Figure 3.1 depicts a transposition spanning one unit. Since in text segmentation (and also music segmentation) near misses can span more than one possible-boundary-location unit, BED extends the Darmeau-Levenshtein edit distance, which is limited to single-unit transpositions, to accommodate for multiple-unit transpositions. Lastly, if $a_i = m_i$ for $i \in \{1, \ldots, n\}$ (M in Figure 3.1), BED stores it as a full match (true positive).

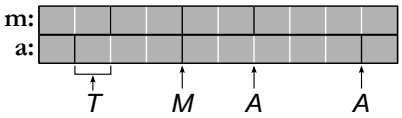

Figure 3.1: Boundary edit operations, adapted from (Fournier 2013b).

The counts of edit operations are then used to model boundary misplacement penalties as specified in Table 3.1.

Using the counts $|A_e|$, $|T_e|$, and $|B_M|$, BED-S can be defined as:

Operation	*Codomain*	*Range*	*Penalty-per-Edit*	*Description*
A_e	$\mathbb{N}_0^2$			set of A edits
T_e	$\mathbb{N}_0^2$			set of T edits
B_M	$\mathbb{N}_0^2$			set of matching boundaries
$\lvert A_e \rvert$	$\mathbb{N}_0$	$[0, n-1]$	1	number of A edits
$\lvert T_e \rvert$	$\mathbb{N}_0$	$[0, \frac{1}{2}\lfloor n-1 \rfloor]$	1	number of T edits
$\lvert B_M \rvert$	$\mathbb{N}_0$	$[0, \frac{1}{2}\lfloor n-1 \rfloor]$	0	number of B_M
$W_T(T_e, n_t)$	$\mathbb{Q}_+$	$[0, \frac{1}{2}\lfloor n-1 \rfloor]$	$[0, 1]$	Weighted T_e operations

Table 3.1: Details for the edits determined using BED, adapted from (Fournier 2013a).

$$BED\text{-}S(m, a) = 1 - \frac{|A_e| + W_T(T_e, n_t)}{|A_e| + |T_e| + |B_M|}, \tag{3.4}$$

where

$$W_T(T_e, n_t) = \sum_{j=1}^{|T_e|} \left(bc_t + \frac{abs(T_e[j][1] - T_e[j][2])}{max(n_t) - 1} \right)$$

Moreover, n_t is a user defined parameter that controls the maximum transposition distance (in potential-boundary-location units), and bc_t is a user defined bias constant.

The intuition for using $W_T(T_e, n_t)$ is simple. It is assumed that penalties for near misses should be proportional to the distance between the reference and predicted boundaries. $W_T(T_e, n_t)$ then corresponds to a distance function whose purpose is to scale transposition errors.

The output value of $BED\text{-}S$ serves as a summary measure of the similarity between **a** and **m**, just like the F_1 score. However, during evaluation one might also want to have higher interpretative power, e.g. in terms of over-segmentation and under-segmentation. To that end Fournier defines a confusion matrix so that TP, TN, FP, and FP are computed using counts of $|A_e|$, $|T_e|$, and $|B_M|$. The confusion matrix can then be used to compute BED-based Precision, Recall, and F_1-measures, which would have the advantage that near misses are accounted for (i.e. TP= $|B_M| + W_T(T_e, n_t)$).

3.4 A New Benchmark Database: The Jazz Tune Corpus (JTC)

In this section we describe and analyse a new benchmark database for machine melody segmenters. We first describe how the database was assembled, its annotation procedure, and its main characteristics. We then analyse the main characteristics of its segments, inter annotation agreement, and cues that might have been involved in the annotation.

3.4.1 The JTC in Brief

The JTC is a dataset of Jazz theme melodies constructed to evaluate computational models of melody segmentation. A list of global statistics describing the dataset is presented in Table 3.2.

Total number of melodies	125
Total number of notes	19419
Total time (in hours)	3.103
Approximate range of dataset (in years)	1880-1986
Total number of composers	81
Total number of styles	10

Table 3.2: Global statistics of the JTC

All melodies are available in MIDI. Each melody in the JTC is annotated with phrase boundaries (by three human listeners) and boundary salience (by two human listeners).[15] In Table 3.3 we present the total number of phrases and mean phrase lengths (with standard deviation values in parenthesis) per annotation.

	Number of	Mean Phrase Length	
Annotation	Phrases	Notes	Seconds
1	1881	10.32 (4.85)	5.94 (3.16)
2	1701	11.42 (6.55)	6.57 (3.93)
3	1682	11.55 (5.78)	6.64 (4.01)

Table 3.3: Summary statistics of annotated phrases. (Standard deviation in parenthesis.)

All segment boundaries and salience annotations were produced using MoSSA – an interface for segment boundary annotation described in (Rodríguez-López et al. 2015).

[15] We use the term 'boundary salience' to refer to a binary score that reflects the relative importance of a given boundary as estimated by a human annotator.

Annotations are provided in Audacity's label file format (Li et al. 2006). The JTC also provides metadata for each melody. The metadata includes information of tune title, composer, Jazz sub-genre, and year of the tune's composition/release.

3.4.2 JTC assembly

To assemble the JTC, we consulted online sources that provide rankings of jazz tunes, albums, and composers.[16] We employed a web-crawler to automatically collect MIDI and MusicXML files from a number of sources in the internet. (The majority were crawled from the now defunct *Wikifonia Foundation*.[17]) We cross referenced the rankings and the collected files, and selected 125 files trying to find a balance between tune ranking, composer ranking, sample coverage, and encoding quality. We describe the JTC's sample coverage (in terms of time periods and sub-genres) below, and discuss the encoding quality of the files in §3.4.4.

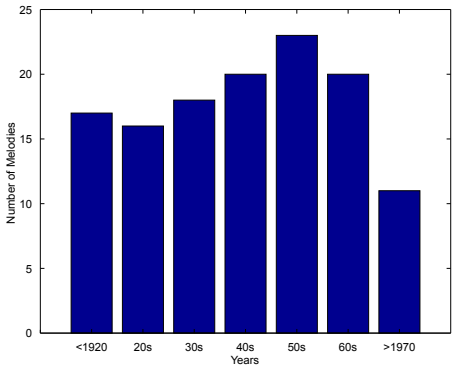

Figure 3.2: JTC: number of melodies per time period

The JTC can be divided in seven time periods (see Figure 3.2). Each time period contains between 11 and 23 tunes from representative sub-genres (see Figure 3.3) and influential composers/performers of the period. The year of release/composition, Jazz sub-genre, and composer metadata was obtained by consulting online sources.[18]

[16] The main sources consulted were: `www.allmusic.com`, `www.jazzstandards.com`, `en.wikipedia.org`

[17] `www.wikifonia.org`

[18] in most cases `en.wikipedia.org` and `www.allmusic.com`

Class Label	Sub-Genre
C1	Bebop
C2	Big Band, Swing, Charleston
C3	Bossa Nova, Latin Jazz
C4	Cool Jazz, Modal Jazz
C5	Dixieland
C6	Early, Rag time, Folk Song
C7	Electric Jazz, Fusion, Modern
C8	Other
C9	Musical, Film, Broadway
C10	Post Bop, Hard Bop

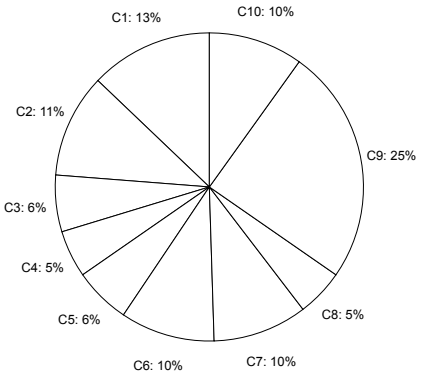

Figure 3.3: Distribution of sub-genres in the JTC

3.4.3 Melody encoding quality and corrections

From the 125 melodies making up the JTC, 64 correspond to perfomed MIDI files, 4 to manually encoded MIDI files, and 57 to manually encoded lead sheets in MusicXML format. In most cases the performed MIDI files encoded polyphonic music, so the melody was extracted automatically by locating the MIDI track labelled as 'melody'.[19]

All melodies were exported as MIDI files, using a resolution of 480 ticks-per-quarter-note, which successfully encoded the lowest temporal resolution of the melodies. All melodies were inspected manually, and, if needed, corrected. Correction of the melodies consisted in adjusting note onsets, as well as removing ornamentation. Notated leadsheets from the Real Book series were used as reference for the correction process.[20] It is important to notice that not all ornamentation was removed, only that

[19] If no such track was found the file was automatically filtered from the selection process.
[20] The Real Book editions used as reference for editing are published by `www.halleonard.com`.

which was considered to severely compromise the intelligibility of segment structure. Also, while JTC melody encodings might contain information of meter, key, and dynamics, this information was not checked nor corrected, and thus its use as 'a priori' information by machine segmenters is discouraged.

3.4.4 Segment Structure Annotations

For each melody, segment boundaries and salience were annotated by one amateur musician and one degree-level musician. These are referred to, respectively, as 'annotation 1' and 'annotation 2' in the tables and figures of this section. For each melody there is also a third annotation of segment boundaries, produced by one of a group of extra annotators. This annotation is referred to as 'annotation 3' throughout the section.

The group of extra annotators consisted of 27 human listeners (18 male and 9 female), ranging from 20 to 50 years of age. In respect to the level of musical education of the extra annotators, 6 reported to be self taught singer/instrumentalist, 10 reported to having some degree of formal musical training, and 11 reported to having obtained a superior education degree in either musicology or music performance. Moreover, the extra annotators were asked to rate their degree of familiarity with Jazz (on a scale from 1 to 3, with 1 being the lowest, and 3 the highest), 12 annotators rated their familiarity as '1', 7 rated their familiarity as '2', and 8 rated their familiarity as '3'. Lastly, none of the extra annotators reported to suffering from any form of hearing impairment, and 2 reported having perfect pitch.

3.4.5 Analysis of phrase annotations in the JTC

In this section we analyse phrase annotations in the JTC. We start with an analysis of two global properties of the annotated phrases: length and contours. Then, we analyse inter-annotator-agreement using two different measures that score agreement. Finally, we check the vicinity of annotated phrases for evidence of two factors commonly assumed to be of high importance to segment boundary perception: *gaps* (in duration and pitch related information) and phrase start *repetitions* (also in duration and pitch related information).

Phrase Lengths and Contours

Average phrase duration lengths presented in Table 3.3 are in line with durations reported in previous manual segmentation studies (Fraisse 1982; Ash 1997; Frieler

et al. 2014).

The box plots presented in Figure 3.4 show that the phrases of annotations 2 and 3 tend to be larger than those in annotation 1. This observation is supported by two key differences. First, both boxes and whiskers of annotations 2 and 3 tend to be larger than those of annotation 1. Second, the notch of box plot 1 does not overlap with those of box plots 2 and 3, which indicates, with 95% confidence, that the difference between their medians is significant.

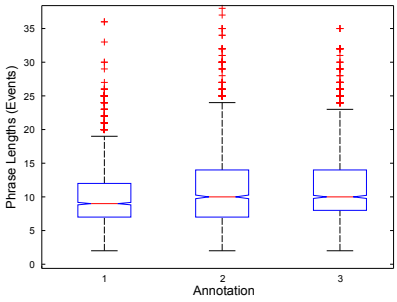

Figure 3.4: Annotated phrase lengths

To get further insights into these apparent preference for longer phrases, we consulted the degree-level musician of annotation 2 and some of the extra annotators for their choice of phrase lengths. The most common reply was that on occasion relatively long melodic passages suggested multiple segmentations, where phrases "seemed to merge into each other" rather than having clear boundaries. For these passages the consulted annotators reported choosing to annotate just one long phrase with 'clear' boundaries rather than attempting to segment the melodic passage into multiple segments.

We also manually checked the outliers identified in Figure 3.4 for the presence of potential annotation errors. In most cases outliers simply correspond to melodic passages with high tempo and high note density, and are not particularly large in terms of time in seconds. Two examples of these type of outliers (common to all annotations) are phrases in the melodies of Dexterity and Ornithology of Charlie Parker.

We classified the annotated phrases in respect to their type of gross melodic contour using the contour types of Huron (1996). Table 3.4 shows the classification results, expressed as a percentage of the total number of phrases per annotation. The results show that all annotators agree in the ranking given to the four dominant contour

Huron's Contour Classes	Annotation		
	1	2	3
convex	33.86	35.10	36.15
descending	23.71	24.99	24.14
ascending	19.30	20.16	19.62
concave	19.99	16.34	17.06
ascending-horizontal	1.33	1.00	1.13
horizontal-descending	0.58	0.88	0.54
horizontal-ascending	0.37	0.59	0.48
descending-horizontal	0.48	0.47	0.42
horizontal	0.37	0.47	0.48

Table 3.4: Contour class classification of annotated phrases

classes, namely convex, descending, ascending, and concave (these four contour classes describe ∼96 percent of the phrases in each annotation). The ranking of the four dominant classes is also in line with the ranking obtained by Huron (1996), who performed phrase contour classification on ∼36000 vocal melodic phrases.

Inter-annotator-agreement (IAA) analysis

We checked the inter-annotator-agreement for each melody annotation using Cohen's κ (1960). Table 3.5 shows the mean pairwise agreement $\overline{\kappa}$, with standard deviation σ_κ in parenthesis. According to the scale proposed by Klaus (1980) the mean agreement on phrase boundary locations between annotations can be considered 'tentative', and according to the scale of Green (1997) it can be considered 'fair'. However, if for each melody we consider only the two highest κ scores, then $\overline{\kappa} = 0.86$, which can be considered by both the Klaus and Green scales as 'good/high'. Moreover, this 'best two' mean agreement also shows a substantial reduction in σ_κ. This indicates that, for any melody in the JTC, it is likely that at least two segmentations have good agreement.

Annotation	$\overline{\kappa}$
1 vs 2	0.72 (0.22)
1 vs 3	0.71 (0.24)
2 vs 3	0.69 (0.26)
Best two	0.86 (0.15)

Table 3.5: Mean pairwise IAA (*kappa*)

Annotation	$\tilde{B}$ (tolerance = 1 note)	$\tilde{B}$ (tolerance = 4 notes)	WSRT
1 vs 2	0.67	0.70	h: 1, Z: 4.54, $p < 0.001$, r: 0.41
1 vs 3	0.62	0.67	h: 1, Z: 5.23, $p < 0.001$, r: 0.46
2 vs 3	0.60	0.65	h: 1, Z: 5.23, $p < 0.001$, r: 0.47

Table 3.6: WSRT of B scores, tilde is used to denote the median

Manual inspection of the boundary annotations showed that, even in cases when the annotators roughly agree on the total number of boundaries for a melody, constructing histograms of boundary markings results in clusters of closely located boundaries. We observed that these boundary clusters are in cases a side effect of dealing with ornamentation during segmentation (i.e. deciding whether grace notes, mordents, or fills should be part of one or another segment). We argue that boundary clusters are examples of 'soft' disagreement and should not be harshly penalised when estimating agreement.

The κ statistic does not take into account the possibility of, nor is able to provide partial scores for, points of 'soft' disagreement when estimating agreement. Hence, to investigate the effect of soft disagreement in the JTC we employed an alternative measure, namely the Boundary Edit Distance Similarity (B), described in §3.3.4. One of the parameters of the B measure is a tolerance window (in notes). Within this tolerance window boundaries are given a partial score proportional to their relative distance. We tested the effect of soft disagreement by computing the B for each melody in the JTC using two tolerance levels: one note (giving score only to points strong agreement) and four notes (giving score also to points of soft agreement). We then computed whether the differences between the medians of the two sets of scores is statistically significant using a paired Wilcoxon Signed Rank test (WSRT). The results of this analysis are presented in Table 3.6. The WSRT confirms that the difference in medians is significant ($p < 0.001$), with medium effect size ($r = 0.41 - 0.47$). These results suggest that the number of points of 'soft' disagreement is not negligible and it should be taken into consideration when benchmarking machine melody segmenters.

3.5 Test Corpus

To test the ability of the machine segmenters presented in this dissertation to locate melodic phrase boundaries, we used a set of 125 instrumental folk songs randomly sampled from the Meertens Tune Collection[21] (`MTC`), 125 vocal folk songs randomly

[21] `http://www.liederenbank.nl`

sampled from the German subset of the Essen Folk Song Collection[22] (EFSC), and the 125 comprising the Jazz Theme Collection (JTC). Throughout this dissertation we refer to this corpus as FJ375.

On the Choice of Test Melodies

Using vocal folk songs for scientific investigation of segmentation seems like natural and intuitive choice. There are two main reasons for this. First, it is unlikely that a vocal melody is perceived as anything but a monophony (which is not necessarily true of instrumental melodies, as they might trigger the perception of parallel melodies). Second, folk melodies are often considered to have a relatively simple segment structure, which allows a higher degree of experimental control.

Melody segmenter evaluation has been confined mainly to vocal melodies. This is likely to introduce a bias during evaluation (physical breath, segmentation might be driven by text, and so on). Thus, to generalise results and evaluate on more 'complex' melodies, we use instrumental folk melodies and jazz head theme melodies.

Short Description of the Collections

Note: refer to §3.4.1 for a description of the JTC.

The EFSC consists of ~6000 songs, mostly of German origin. The EFSC was compiled and encoded from notated sources. The songs are available in EsAC and **kern formats. The origin of phrase boundary markings in the EFSC has not been explicitly documented, yet it is commonly assumed markings coincide with breath marks or phrase boundaries in the lyrics of the songs. Thom et al. (2002, pp. 68–69) cites a comment of made by Ewa Dahlig (who at the time mantained the EFSC) on the phrase markings in the collection: *"When we encode a tune without knowing its text, we do it just intuitively, using musical experience and sense of the structure. The idea is to make the phrase not too short (then it is a motive rather than a phrase) and not too long (then it becomes a musical period with cadence etc.). But, of course, there are no rules, the division is and will be subjective."*

The instrumental (mainly fiddle) subset of the MTC consists of ~2500 songs. The songs were compiled and encoded from notated sources. The songs are available in MIDI and **kern formats. Segment boundary markings for this subset comprise two levels: 'hard' and 'soft'. Hard (section) boundary markings correspond with structural marks

[22] http://www.esac-data.org

found in the notated sources. Soft (phrase) boundary markings where annotated by two experts. Instructions to annotate boundaries were related to performance practice (e.g. "where would you change the movement of the bow"). The annotators agreed on a single segmentation, so no inter-annotator-agreement analysis is possible. For our experiments we use the soft boundary markings.

Corpus Cleaning and Sampling Considerations for the FJ375

Melodies collected from the EFSC and MTC are selected via random sampling. However, following the corpus cleaning procedures of Shanahan and Huron (2011), we filtered out melodies which contained rests at annotated phrase markings, and also excluded melodies with just one phrase. The reason to exclude melodies with rests at annotated phrase markings is that, according to transcription research, sometimes musicologists transcribing the folk melodies use rests at phrases as 'breath marks', regardless of whether performers would actually take breaths or not, making these rests an artefact of the transcription process.

It must be also be noted that in the JTC nearly half of the melodies were performed by a musician (most likely using a midi keyboard). Thus, even though melodies might sound monophonic, they might in fact not be strictly so. There is a large possibility of partial overlapping between consecutive notes due to legato articulation. A simple procedure to eliminate overlapping notes was applied: if the offset time of note n is larger than the onset of note $n+1$, then the offset time is truncated and takes the value of the onset of note $n+1$.

Also, due to human motor capacity, performed note duration (onset/offset) information is likely to vary significantly from duration classes used in score notation. Note onsets where 'corrected' manually, i.e. the onsets where aligned to the nearest beat. Offsets, however, where not corrected. For this reason, when computing note duration values our models do it based only on onset information, i.e. duration is measured by computing inter-onset-intervals – see Appendix A.

Corpus Formatting

The sampled melody collections used to test our segmenters are either in MIDI, **kern, or MusicXML formats. For segmentation analysis melodies are converted into 'note lists', i.e. a list of the notes in a melody, where each note is described as a {pitch, onset, offset} set as:

	onset	offset	pitch	boundary
Note	0	243	66	0
Note	250	475	64	0
Note	500	711	66	0
Note	750	999	69	1

This is the preferred format for the Melisma suite (Sleator and Temperley 2001) and is also equivalent to a 4-column note matrix of the Miditoolbox (Eerola and Toiviainen 2004).

Encoding Boundaries

For the JTC boundary data was encoded as a list of time points (with msec precision). This time points are then transformed to binary vector representation, i.e. a vector where each element corresponds to a note event in the melody, and a 1 indicates the starting note of a segment – see the note list example above. The same procedure is applied to boundaries of the EFSC and MTC.

3.6 Guidelines

In this dissertation we use the following guidelines when evaluating machine segmenters:

1. *Prefer reference-based evaluation to task-based evaluation.* Whenever possible reference-based evaluation is preferred to task-based evaluation. The reason for this choice is that the former strategy gives, at present, the least biased evaluation scenario. For task-based evaluation we (as a research community) are still lacking user data and user feedback, which complicates the interpretation of evaluation results – see §3.2.3.
2. *Compare single-cue segmenters only to other segmenters modelling the same cue.* Since it is not possible to penalise false positives in an unbiased way – see §3.3.1 – the least biased evaluation is to compare performances of a segmenter to other segmenters modelling the same cue.
3. *Design baseline segmenters specifically for each evaluation.* The modelling of segmentation cues are known to exhibit different issues. For instance, repetition and gap segmenters are known to be prone to over-segmentation. Conversely, contrast segmenters are known to be prone to under-segmentation. We hence design baselines that present a 'worse case' scenario for each segmenter, so as to better interpret the scale of the evaluation measure used.

4. *Use both $F1$ and B evaluation measures.* We use both $F1$ (with P and R) described in §3.3.2 to evaluate our segmenters. We set an strict tolerance level: if the predicted boundary coincides with *either the last note or the first note* of a manually annotated segment boundary the prediction is considered a true positive.[23] Otherwise it is considered a false positive. To investigate the possibility of fuzzy boundaries, we use the B measure, described in §3.3.4. We set the tolerance of B to 4 notes.

3.7 Conclusions

In this chapter we discuss methodologies, corpora, and measures used in MIR to evaluate melody segmenters. We critique and propose ways to improve the evaluation methodology currently at MIR. We motivate and study new quantitative evaluation measures. We introduce a new test dataset, consisting of 125 Jazz melodies.

The database and evaluation measures proposed here are used throughout the rest of this thesis. The directions and suggestions given to improve the segmenter evaluation are left as future work.

[23] In MIREX absolute time windows have been used to allow for a degree of tolerance in the presence of tempo variation. The idea is that the amount of tolerance should be proportional to tempo. So that higher tempos should have higher tolerance, and lower tempos lower tolerance. To put the time window in better perspective, a ± 0.5 second window has a precision of 1 or 2 notes at lower tempos (say 60 BPM) and 2 to 4 notes at higher tempos. Conversely, a ± 3 second window has a precision of 6-12 notes at lower tempos, and 12-24 notes at higher tempos. As stated above for music stored in symbolic format tolerance is most often specified directly in notes. This is due because of the common object of study: vocal melodies. Vocal melodies are assumed to be sparse (low note density in time) and remain recognisable despite large variation in tempo. Most databases in symbolic format do not even specify a fixed tempo. It is then expected that tolerance in notes won't have large effect on performance. In the benchmark dataset used in this dissertation has vocal and instrumental melodies of Folk music, and melodies of Jazz music. The most melodies comply with the assumptions of sparsity and tempo stability/range. We then also measure tolerance directly in note events.

Chapter 4

Repetition Based Segmentation

In this chapter we tackle the problem of automatically segmenting melodies into phrases via repetition identification. We focus on investigating the role of location related information on the automatic identification of repetitions.

Chapter Contributions

We introduce three complementary scoring functions based on location information to identify (segmentation determinative) repetitions.

To test the ability of our functions to identify segmentation determinative repetitions, we incorporate them in a state-of-the-art repetition based segmenter proposed by Müller and Grosche (2012). The original and altered versions of MUL are used to segment melodies from the `FJ375` corpus. Results show that by using our scoring functions the segmenter achieves a statistically significant 14% average improvement over its original version.

This chapter extends work presented in (Rodríguez-López et al. 2014b; Rodríguez-López and Volk 2015a).

4.1 Introduction

In this chapter we tackle the problem of segmenting complete melodies into phrases by modelling repetition cues.

Main Concepts. We use the term *repetition* to refer to a human-identified reoccurrence of a fragment of music within a piece. One fragment is a repetition of another fragment if it is judged to be same (in some respect) by a human listener. *Repetition cues* refer to the identification of repetitions during a segmentation process, and *repetition based segmentation* to machine segmentation approaches where it is assumed that repetition cues play a central role in the detection of segment boundaries. More specifically, that the start or ending points of identified repetitions are likely to be points of segmentation (segment boundaries).[24]

Modelling Tasks. Modelling repetition cues requires (I) a search process to locate candidate repetitions within a melody, (II) a method to estimate if candidate repetitions are perceived as such or not (and are moreover likely to influence segmentation), (III) a method to use identified repetitions to locate segment boundaries.

Focus. While each of the modelling tasks listed above have their own challenges, in this chapter we tackle those related to subtask II. Our motivation to do so is based on the findings of a recent comparative study (Rodríguez-López et al. 2014b). In the study it was observed that repetition based segmenters tend to detect too many repetitions, leading to over-segmentation. The tendency for machine repetition detectors to identify much more repetitions that those actually recognised by human listeners is also a known issue in automatic motif identification – see discussions in (Meredith et al. 2001, pp. 5–6; Lartillot 2007, pp. 244–245). For segmentation modelling this issue is all the more acute, given that the number of repetitions relevant for segment boundary perception is likely to be much smaller than the total number of recognised repetitions. (One could argue that in many cases human repetition identification would require having interiorised the segment structure of a piece.) Improving automatic identification of perceived, segmentation-determinative repetitions is thus a crucial step to the development of more accurate repetition based segmenters.

Contributions. The problem posed by subtask II is essentially to model the conditions that facilitate repetition recognition. We refer to these conditions as *cognitive constraints*. Existing repetition based segmenters most often model cognitive constraints

[24] While for some readers it might seem obvious, it is relevant to stress that repetition-based segmenters do not assume all segment boundaries are linked to repetition identification. It is presupposed (though most often left implicit) that repetition based segmenters will not discover all possible boundaries.

based on information of the frequency, length, and temporal overlap of/between detected candidate repetitions. In this chapter we investigate the role of location-related cognitive constraints on repetition recognition. For brevity we refer to them as *location constraints*. We focus on information of the location of repetitions relative to (a) each other, (b) the whole melody, and (c) temporal gaps. We introduce three scoring functions that make use of this information to rank candidate repetitions. (Where a high rank indicates that the repeated fragment is more likely to be perceived.)

We incorporate our scoring functions in an optimisation framework for repetition based segmentation proposed by Müller and Grosche (2012). For brevity, through this chapter we refer to this framework as MUL. The original and constraint-extended versions of MUL are used to segment the `FJ375` corpus. Results show the constraint-extended version of MUL achieves a statistically significant 14% average improvement over MUL's original version.

Chapter Structure. This chapter is organised as follows. In §4.2 we discuss repetition cues in more detail to limit the scope of our study. In §4.4 we describe the MUL segmentation model. In §4.5 we introduce our location constraints and describe how they are integrated into MUL. In §4.6 we describe the experimental setting, present results, and discuss how location constraints affect the performance of MUL. Finally, in §4.7 we present our conclusions and outline future work.

4.2 Discussion on Repetition Cues

In §2.3.2, page 20, we introduced 'repetition' and 'homogeneity' as classes of segmentation cues related to similarity processing. We used the class homogeneity to refer to cues that contribute to the perception of unity within a segment, and the class repetition to cues that help establish links between segments. For the sake of taxonomisation clarity we treated repetition and homogeneity cues as independent. Assuming independence is, however, an oversimplification. There are many situations in which identifying repetitions influences the perception of unity and cohesiveness rather than that of boundaries. A clear separation of the situations in which repetition assumes one or the other role (as a segmentation cue) is not straightforward. The reason being that this role heavily depends on the time span of the fragments being repeated (whether similar in size to figures, phrases, etc.) and the time scale of the segmentation (phrases into subphrases, whole melody into phrases, etc.). Take for instance the repetitions of fragments b and c in Figure 4.1, bars 21–30. In this example repetitions are likely to influence the perception of boundaries in *figure-size* segments, but at the same are likely to influence the perception of cohesion in *phrase-*

size segments.

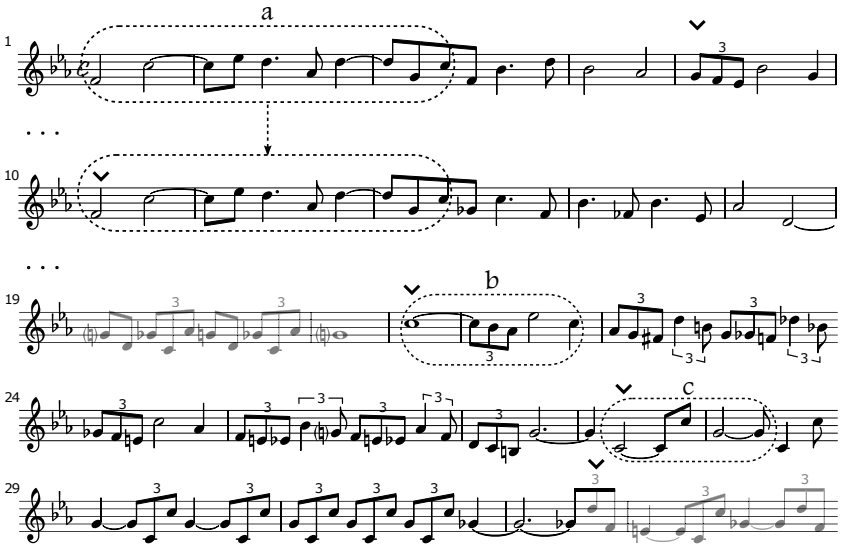

Figure 4.1: Fragment of the English horn solo from *Tristan und Isolde* by R. Wagner. Dotted regions enclose fragments that repeat. Arrow heads mark phrase boundaries identified by human listeners – refer to (Deliège 2007, pp. 15–18) for details.

In this chapter we concentrate on phrase level segmentation and repetition of fragments whose length ranges between figures and phrases. We use these limits to propose a rough typification of the role of repetitions in segmentation, by separating between *temporally close* and *temporally distant* repetitions. Below we discuss each class in turn.

Close Repetitions. If the repeated fragments are temporally close, it is thought that their identification contributes to the sensation of unity. The unifying role of temporally close repetition in a segmentation process seems to always be for 'the next level up', i.e. close note repetition assists/enables figure cohesion, close figure repetition assists/enables phrase cohesion, and so on. To illustrate this we can refer to the same example as before, i.e. the repetitions of fragments b and c in Figure 4.1. We argue that, at least in this particular case, the immediate and frequent repetition of b and c not only contributes to the sensation of within-phrase cohesion, but rather that it is determinant of it.

(Note: It is necessary at this point to make a clarification. What makes a repetition pair temporally 'close' is relative to the time span of the analysis. Two consecutive

repetitions are close if their temporal distance is smaller than the average time span of the segmentation. Or, in other words, if the two are likely to be located in the same segment. Conversely, temporally distant repetitions are those who are likely to be located in different segments.)

Distant Repetitions. If identified repetitions are temporally distant their starting and/or ending points are thought to indicate likely boundary locations. Take for instance the repetition of fragment a in Figure 4.1, which suggests the beginning of bar 10 to be the starting point of a new phrase. There is experimental evidence which suggests that identification of temporally distant repetitions, like that of fragment a, has a considerable influence on the perception of phrase boundaries (Spiro 2007, pp. 356–357), and form section boundaries (Clarke and Krumhansl 1990; Bruderer et al. 2006).

In this chapter we are interested in modelling *distant* repetition identification as a cue to phrase boundary perception. However, doing so posses a Catch-22 problem: identifying temporally distant repetitions is itself likely to depend on segment structure. Music psychology experiments have provided evidence of situations where this is the case. To name one, in (Margulis 2012) a short piano piece was manually segmented into phrases. Exact (note-by-note) matching fragments where identified (again manually) on the score of the piece. Evidence was found that human listeners had more difficulties to aurally identify exact matching fragments if they occurred across phrases rather than within phrases.

Moreover, repetition identification has also been shown to, in cases, depend on tonal or metric structure information. For instance, a common observation is that repetitions are more likely to be identified by listeners if the fragments start in metrically strong positions (Ahlbäck 2007). Also, sameness and difference judgements between melodic fragments may depend on diatonic pitch interval perception, which requires the conception of a tonal centre (Cambouropoulos 1996, 1997a). And yet, tonal/metric structures are themselves thought to require segment structure information for their conception – see for instance (Cuddy 1993, pp. 35–36) for a discussion on the influence of segmentation on the formation of tonal structure and (Ahlbäck, 2004, pp. 101–106; Temperley, 2001, pp. 49–51) for discussions on the influence of segmentation on the formation of metric structure.

We use these issues to further motivate our research, and also establish more limits on its scope. In respect to motivation, the aforementioned issues stress the importance of accurate simulation of repetition identification for repetition-based segmenters. They make it sensible to assume that the number of repetitions relevant for phrase boundary detection is much smaller than the total number of perceived repetitions. This in

turn suggests that the conditions and information required for successful repetition identification goes beyond the commonly investigated frequency and length. In respect to scope, to avoid conflicts with tonal/metric structure formation, we focus on what is often called 'surface' similarity when estimating repetition (Cambouropoulos and Tsougras 2004; Lalitte et al. 2004; McAdams et al. 2004; Lamont and Dibben 2001). That is, we do not distinguish between tonally or metrically important/unimportant parts of the fragments being compared when estimating if one is a repetition of the other.

4.3 Related Work

The processing chain of repetition-based melody segmenters generally has four stages: (1) input a melodic sequence, (2) automatically identify exact or approximate repetitions in the sequence, (3) select only those repetitions that might be relevant for segmentation, and (4) output the start and/or ending points of the selected repetitions as segment boundaries. Based on these four stages, Table 4.1 gives a summary of the main characteristics of the segmenters reviewed in this section.[25]

The listed segmenters have mostly focused on the segmentation melodies into either stanzas, phrases, or subphrases.[26] Their approaches to computing the segmentations are, however, quite diverse, showing differences in almost every aspect of the processing chain. The way repetitions are automatically identified tends to be the most similar, with a dominance of exact or approximate string search techniques. Conversely, the largest differences are observed in the way the input melody is represented, and the way repetitions are selected. In this section we focus on discussing aspects related to repetition selection. For thorough discussions on melody representation and automatic identification we refer to (Cambouropoulos et al. 1997a; 2001; 2009).

Repetition selection simulates the cognitive constraints influencing human repetition

[25] Our review concentrates on approaches that fulfil at least two of the following criteria: (a) the approach focuses on processing symbolic encodings of music, (b) the approach focuses on identifying repetitions for phrase level segmentation, (c) the approach has been designed-for or tested-on melodies. For more complete reviews of computational modelling of music similarity (within and across pieces) we refer to (Cambouropoulos et al. 2001; Meredith et al. 2002; Lartillot 2007; Janssen et al. 2013).

[26] Two exceptions are the segmenters proposed by Wołkowicz (2013) and Rafael and Oertl (2010). The former attempts to automatically locate form-level segment boundaries in polyphonic music. During preprocessing polyphony is automatically reduced to monophony. The latter takes as input multi-part polyphonic music and aims to infer segment boundaries at various segment granularities. In the publication it is not clear if the parts, which in principle can be polyphonic, are reduced to monophony or not.

Segmenter	Attribute Sequence	Identify Repetitions		Select Repetitions	
Author(s)		Search/Store	Similarity	Information	Technique
Ahlbäck (2007)	`cp-iv*,beat`	–	–	–	preference rules
Cambouropoulos (2006)	`sl-iv*` and `ioi-r`	string search	E	`L`, `F`, `TO`	scoring function
Rafael et al. (2009)	–,`beat`	string search	A (dtw)	`L`, –	optimisation
Rafael and Oertl (2010)	–,`beat`	string search	E,A (dtw)	`L`, –	optimisation
Takasu et al. (1999)	`cp`	string search	A (lcs)	`P`, `TO`	preference rules
Wołkowicz (2013)	`cp-iv` and `ioi-r`	similarity matrix	A (cos)	`L`, `TO`	scoring function
MUL	chroma vector	similarity matrix	A (cos)	`L`, `F`, `TO`	optimisation

Table 4.1: Reviewed repetition-based segmenters. *Attribute*: attribute sequence used to describe melodies in the publication, the first attribute indicates pitch specification, the second attribute (if present) indicates duration specification, 'and' indicates both attributes are used to describe a single melodic event, 'comma' indicates attributes are processed in parallel (as independent sequences), 'asterisk' indicates the specification of the attribute is non-standard. *Search/Store*: data structure construction method used to search and store repetitions. *Similarity*: E - exact matching, A - approximate matching, in parenthesis there is an abbreviation for the similarity measured employed in the publication (cos - cosine, lcs - longest common subsequence, dtw - dynamic time warping). *Information*: `L` - length, `F` - frequency, `TO` - temporal overlap, `P` - position. For all columns a hyphen indicates 'unclear/unspecified'.

identification. The most commonly used information to do so is: *degree of similarity*, *frequency*, *length*, and *temporal overlap*. The motivation to use the first two should appear intuitive to most readers. The more similar two fragments are the more likely it is that a listener will judge them to be repetitions. Likewise, the more times a fragment is repeated through a melody the more chances are that a listener will notice it. Length, on the other hand, seems to have an indirect effect on repetition identification. Short fragments (say one or two intervals) might appear too often through a melody, and, because of their commonality (of 'being everywhere'), loose their power as segmentation cue. Longer fragments that repeat are thus assumed to give more specific information to listeners, and are hence preferred. Lastly, it is known that listeners have more difficulty recognising repetitions if these temporally overlap. For this reason most of the approches reviewed completely reject overlaps.

The above mentioned information is used to estimate whether the automatically identified repetitions of a fragment are perceived or not. This has been done using either *preference rules*, (user-controlled) *scoring functions*, or (machine-controlled) scoring functions – notice we refer to the latter as *optimisation* in Table 4.1.

An example of a segmenter using preference rules for selection is that proposed by Takasu et al. (1999). It pre-segments by locating temporal gaps, then computes the similarity between all resulting fragment pairs, and uses an automatically determ-

inable threshold technique to locate all fragments that constitute a set of candidate repetitions. It then selects repetitions which have an instance located (a) after a long temporal gap, or (b) at the beginning of the melody. In preliminary experiments we found that the segmenter performs with high precision but with very low recall. It seems that many repetitions that humans would be able to identify may have been discarded during pre-segmentation.

Cambouropoulos (2006), conversely, uses an efficient exact-match string search algorithm to compute all candidate repetitions (up to a user defined maximum length). He then defines a user-controlled scoring function to rank repetitions. The highest scoring repetitions are taken as boundary candidates. In preliminary experiments we found that the optimal parameter setting of the scoring function varies greatly from one melody to the next (and from one melody representation to another). This high context dependency had a pronounced adverse effect on performance when a single setting was chosen to evaluate the segmenter over a large corpus of melodies. We observed that the same negative effects of having a user-determined scoring function affected the performance of the segmenter proposed by Wołkowicz (2013, Ch. 6).

The segmenter used for experimentation in this chapter, MUL, uses an optimisation approach which allows the segmenter to automatically determine the parameter settings of the repetition selector. In (Rodríguez-López et al. 2014b) we conducted an evaluation study where MUL performed best, suggesting that the proposed optimisation framework helps ameliorating the context dependence issues that lower the performance of the other segmenters of Cambouropoulos and Wołkowicz. Moreover, by not starting from a predefined segmentation (as the segmenter of Takasu), it is able to achieve higher recall.[27]

4.4 Description of the MUL Segmentation Model

MUL searches for the 'most representative' melody fragment and uses its repetitions to segment the melody. As shown in Figure 4.2, MUL first computes a similarity matrix representation of the input melody, where repetitions can be visualised as diagonal

[27] We have left three segmenters in Table 4.1 out of our critical commentary: Rafael et al. (2009); Rafael and Oertl (2010); Ahlbäck (2007). The segmenters proposed by Rafael et al. (2009); Rafael and Oertl (2010) propose alternative and interesting approaches to the problem at hand. However, it is difficult to estimate their contribution since they have not been systematically evaluated. We were not able to test the proposed segmenters because they are not described in sufficient detail for implementation, and no implementation has been made available to the community. The segmenter proposed by Ahlbäck (2007), conversely, has been tested extensively – see (Ahlbäck 2004) – but the results are difficult to summarise (the evaluation is most often qualitative, on a case-by-case basis). We were not able to test the segmenter for the same reasons as with those proposed by Rafael et al.

or quasi diagonal stripes. Then, it uses an exhaustive stripe search technique to identify repetitions, and scores each repetition set according to the degree of similarity, frequency, and length of the repetitions contained in the set. Finally, it takes the highest scoring set of repetitions, and uses the start/end points of these repetitions as segment boundaries. Below we briefly describe the different processing stages of MUL – for a more detailed description we refer to (Müller et al. 2013).

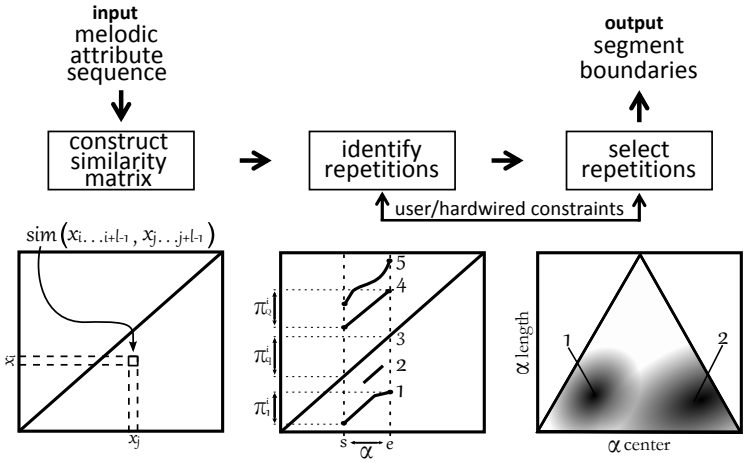

Figure 4.2: Processing chain of the repetition-based segmentation model. Left: SM construction process. Middle: simplified SM depiction of a fragment α and five stripes, stripes {1,3,4} constitute an 'accepted' set of repetitions $\mathcal{P}$. Right: scape plot representation of the space of fragments $\mathcal{A}^*$, shading depicts ϕ score (fitness), points 1 and 2 mark two fragments with high fitness.

4.4.1 Similarity Matrix Construction for Symbolic Data

MUL was originally developed to take audio data as input. In this chapter, conversely, MUL takes symbolic data as input. The melody is thus represented as a sequence of symbols. Each symbol represents an attribute describing either a note (e.g. its pitch or duration) or a short sequence of notes (e.g. a chromatic pitch interval or inter-onset-interval ratio). Let then $x = x_1 \dots x_N$ be a sequence of melodic symbols of length N, and let $x_{i\dots j} = x_i \dots x_j$ be a sub-sequence of x, with $i, j \in [1 : N]$. A similarity matrix SM of x corresponds to the matrix $\mathcal{S} = [s_{ij}]_{N\times N}$ of pairwise similarities between subsequences $s_{ij} = sim(x_{i\dots i+l-1}, x_{j\dots j+l-1})$, where l indicates the length of the subsequence and sim is a similarity measure. Figure 4.2 (left) depicts the construction process of an SM. In this chapter we employ SMs that fulfil the normalisation properties $0 \leq \mathcal{S}(i,j) \leq 1$ for $i, j \in [1 : N]$, and $\mathcal{S}(i,i) = 1$ for $i \in [1 : N]$. In a SM temporally distant repetitions are visualised as diagonal or quasi

diagonal stripes. Figure 4.2 (middle) depicts a SM schetch, where five diagonal stripes mark potential repetitions of fragment α. The stripe structure of SMs computed from music data are often noisier than that shown in our simplified SM example. Thus, denoising and smoothing methods are commonly used to post-process SMs, aiming to enhance desired structural properties of the SM (stripes in our case) while suppressing unwanted ones – for details on enhancement techniques refer to (Müller and Kurth 2006; Müller and Clausen 2007).

The parameter settings to construct the SMs used in this chapter (i.e. melodic representation, fragment length, similarity measure, denoising, and smoothing) are listed in Table 4.2, §4.6.3.

4.4.2 Constraint-based Identification and Selection of Repetitions

MUL uses information on the number or *frequency* of repetitions, their *length*, and the amount of *temporal overlap* between repetitions as well as their *degree of similarity* to model cognitive constraints of human repetition identification. Below we describe how this information is used to extract and score stripes from SMs, and then to select sets of stripes.

Repetition identification (stripe extraction)

The goal is to identify and store repetitions for all fragments ranging in length from one event to all the events in the melody. To that end MUL defines the space of fragments $\mathcal{A}^*$ as a superset containing all sets $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots \alpha_K\}$ of pairwise disjoint fragments $\alpha_h \cap \alpha_k = \varnothing$ for $h, k \in [1:K]$ and $h \neq k$, where $\alpha = [b:e] \subseteq [1:N]$ is a fragment of the melody. Repetitions of each melodic fragment are identified by extracting quasi diagonal stripes from $\mathcal{S}$ in the region encompassed by the fragment, e.g. Fig. 4.2 (middle SM) shows a fragment α and five stripes marking potential repetitions. If we take the tuple $(i_l, j_l) \in [1:N]^2$, $l \in [1:L]$ to denote a cell of $\mathcal{S}$, then a stripe of length L can be defined as any sequence $\pi = (i_l, j_l), \ldots, (i_L, j_L)$ forming a path within the region encompassed by fragment α. A path π has two projections $\pi^i = [i_l : i_L]$ and $\pi^j = [j_l : j_L]$. The constraints for a set of stripes $\mathcal{P} = \{\pi_1, \pi_2, \ldots \pi_Q\}$ to be a set of repetitions are:

1· stripe projections π^j must be of the same length as α (i.e. $j_1 = b$ and $j_L = e$),

2· stripes must be diagonal or quasi diagonal, for which user defined diagonal distortions are allowed (the default setting requires the slope of a stripe to lay within the bounds 1/2 and 2), and

3· the set of stripe projections π^i must not temporally overlap.

In Fig. 4.2 (middle SM) we exemplify how MUL enforces these constraints. From the set of stripes {1,2,3,4,5}, the set of stripes complying with the criteria is {1,3,4}, since stripe 2 is unacceptably short, and stripe 5 is both unacceptably distorted and its π^i projection overlaps with that of stripe 4. Since a fragment can have more than one acceptable set of repetitions, MUL uses an optimisation procedure to search for the best possible set of repetitions. MUL defines the optimal set of repetitions $\mathcal{P}^o$ as that containing the most frequent and similar repetitions (using Eq. 4.1 below). The identification of repetitions and the search for the optimal set of repetitions is computed simultaneously, using a modification of the classic dynamic time warping algorithm.

Repetition selection (fitness function)

To select which fragment α to use for segmentation, MUL enforces constraints on the degree of similarity, length, and frequency of its associated set of repetitions $\mathcal{P}^o$. The main idea is to search for the 'most representative' fragment. MUL defines the most representative fragment as that which contains *the highest repeating and most similar set of repetitions*, which moreover *covers the largest portion of the melody*. To formalise this idea MUL employs two heuristic functions. The first is a *repetition* score function

$$\rho(\mathcal{P}) = \sum_{q=1}^{Q} \rho(\pi_q), \tag{4.1}$$

with $\rho(\pi) = \sum_{l=1}^{L} \mathcal{S}(i_l, j_l)$. The function $\rho(\mathcal{P})$ awards a high score to sets with highly similar and frequent repetitions. The second is a *coverage* score function

$$\kappa(\mathcal{P}) = \sum_{q=1}^{Q} |\pi_q|, \tag{4.2}$$

with $|\cdot|$ used to denote the length of π. The function $\kappa(\mathcal{P})$ awards a higher score to repetition sets that cover a large part of the melody. MUL uses normalised versions of $\rho(\cdot)$, $\kappa(\cdot)$. For brevity we omit a description of the normalisation procedures and refer to (Müller et al. 2013). The normalised scoring functions (denoted by $\tilde{\rho}(\cdot)$, $\tilde{\kappa}(\cdot)$)

are combined using a harmonic mean, i.e.

$$\phi(\alpha) = 2 \cdot \frac{\tilde{\rho}(\mathcal{P}^o) \cdot \tilde{\kappa}(\mathcal{P}^o)}{\tilde{\rho}(\mathcal{P}^o) + \tilde{\kappa}(\mathcal{P}^o)}, \tag{4.3}$$

MUL uses $\phi(\cdot)$ as a 'fitness' measure whose score represents a balance between having highly frequent/similar repetitions and covering large portions of the melody. The most representative fragment is that containing the repetition set of maximal fitness:

$$\alpha^m = \underset{\alpha}{\operatorname{argmax}}\, \phi(\alpha). \tag{4.4}$$

4.5 Location Constraints for Repetition Selection

Repetition recognition is a recall process, i.e. the realisation that what is currently been heard occurred earlier. Recognising repetitions is then likely to be heavily influenced by location related information, such primacy, order, recency, and so on (Murdock 1962; Greene 1986). Thus, we introduce scoring functions based on the location of repetitions relative to (a) each other, (b) the whole melody, and (c) the location of temporal gaps. Below we describe and motivate our scoring functions in turn.

Scoring Repetition Dispersion

In respect to (a), we hypothesise that repetition sets in which instances are roughly evenly spaced (within the melody) are more salient than those that are not. We do so based on the observation that phrases tend to have a narrow distribution of possible phrase lengths (Temperley 2001, Ch. 3). Hence, if we assume that salient repetitions mark mainly the starting points of phrases (Rodríguez-López et al. 2014b), then the distribution of *inter-repetition-onset-intervals* (IROI) of salient repetitions should also be dominated by relative few and similar IROIs. We propose λ_1 (Eq. 4.5) as a scoring function that gives higher score repetition sets with low IROI dispersion.

$$\lambda_1(\mathcal{P}) = \frac{1}{\sigma_{iRoi} + 1} \tag{4.5}$$

where $\text{IROI}(\pi_q) = \pi_{q+1} - \pi_q$, $\forall q = 1, \ldots, |\mathcal{P}| - 1$ and σ is the standard deviation.[28]

[28] Since $\sigma \in \mathbb{R}_{\geq 0}$, normalisation of the λ_1 values is required.

Scoring Repetition Priming

In respect to (b), we hypothesise that repetition sets in which the first instances occur earlier in the melody are more salient than those containing first instances appearing later in the melody. We do so based on the notion that melodic 'vocabulary' is mostly emergent, and so the earlier a 'vocabulary term' is introduced the higher its relevance (Cambouropoulos 2006; Takasu et al. 1999). To quantify this notion, we use λ_2 as a scoring function that prefers sets of repetitions with instances located both at the 'beginning' (I_b) and the 'rest' (I_r) of a melody.

$$\lambda_2(\mathcal{P}) = \sqrt{I_b \cdot I_r} \tag{4.6}$$

where $I_b = \frac{\mathcal{O} \cap \mathcal{B}}{|\mathcal{B}|}$ and $I_r = \frac{\mathcal{O} \cap \neg\mathcal{B}}{|\mathcal{O}|}$, $\mathcal{O}$ is the set of repetition onsets from $\mathcal{P}$, and $\mathcal{B}$ is the set of possible note locations at the beginning of the melody. We take $x_{1\ldots\lfloor N/n \rfloor}$ to be the melody 'beginning', with n defined by the user (see settings in Table 4.2).[29]

Scoring Repetition Alignment to Temporal Gaps

In respect to (c), we hypothesise that repetition sets that better align to temporal gaps are more salient than those which do not. (In melodies temporal gaps can be overly long note durations, musical rests, or a combination of the two.) The motivation for this hypothesis is based on the observation that temporal gaps often precede phrase starts (Temperley 2001; Takasu et al. 1999), and repetitions often mark the starting points of phrases (Margulis 2012; Huron 2006). To quantify this notion, we use λ_3 as a scoring function that prefers sets of repetitions containing one or more instances starting right after temporal gaps.

$$\lambda_3(\mathcal{P}) = 2 \cdot \frac{T_p \cdot T_r}{T_p + T_r} \tag{4.7}$$

where $T_p = \frac{\mathcal{T} \cap \mathcal{O}}{|\mathcal{O}|}$ and $T_r = \frac{\mathcal{T} \cap \mathcal{O}}{|\mathcal{T}|}$, $\mathcal{O}$ is the set of repetition onsets from $\mathcal{P}$, and $\mathcal{T}$ is the set of temporal gap locations. To automatically obtain temporal gap locations, we use the temporal gap detection component of the LBDM segmenter (Cambouropoulos 2001) – settings specified in Table 4.2. Each estimated temporal gap location in $\mathcal{T}$ has been incremented on one note event to align with repetition onsets.

[29] While theoretically $\lambda_2 \in [0, 1]$, considering $\lim_{\mathcal{O} \to N} \lambda_2(\mathcal{O}) = 1$, in practice the values of λ_2 will never reach the maximum of the function's range, and so re-scaling is required.

Combining Scores

In our experiments we incorporate the arithmetic mean $\overline{\lambda}$ of the scores $\lambda_{1,2,3}$, in the fitness measure (Eq. 4.3) which results in

$$\phi(\alpha) = 3 \cdot \frac{\tilde{\rho}(\mathcal{P}^o) \cdot \tilde{\kappa}(\mathcal{P}^o) \cdot \overline{\lambda}(\mathcal{P}^o)}{\tilde{\rho}(\mathcal{P}^o) + \tilde{\kappa}(\mathcal{P}^o) + \overline{\lambda}(\mathcal{P}^o)} \tag{4.8}$$

To select a meaningful set of repetitions from the ϕ-space the same criterion used in Eq. 6.3 is employed, namely the most representative fragment is that containing the repetition set of maximal fitness.

4.6 Evaluation

In this section we describe the test database and evaluation metrics, list experimental parameter settings, and present the results obtained in our experiments. For our experiments we use the implementation of MUL provided in the `SM toolbox` (Müller et al. 2014). We coded additional functions that compute SMs from symbolic data and implement the location constraints described in §4.5.

4.6.1 Experimental Setting: Test Dataset

To test the ability of MUL and its extended versions to locate melodic phrase boundaries, we use the `FJ375` corpus – refer to §3.5 for a description.

4.6.2 Experimental Setting: Evaluation Measures

We use the well known $F1$, precision P, and recall R measures, defined in Equations 3.1, 3.2, and 3.3, respectively. To take into account the possibility of fuzzy boundaries during evaluation (see discussion in §3.3.1), we also use Boundary Edit Distance Similarity B, defined in Equation 3.4. One of the parameters of the B measure is a tolerance window (in notes). Within this tolerance window boundaries are given a partial score proportional to their relative distance. We tested the effect of soft disagreement by computing the B using a tolerance of four notes (giving score also to points of soft agreement).

4.6.3 Experimental Setting: Parameters

In Table 4.2 we specify SMs construction parameters and repetition identification/selection parameters used for experimentation. The choice of parameters is the result of previous experimentation with MUL reported in (Rodríguez-López et al. 2014b).

Parameters		Setting used for experimentation
SM construction		
melodic fragment length	fl	fragment length of 4 notes
similarity measures	sm	cosine similarity
melody representation	mr	`cp-iv`, `ioi-r`
matrix blending	smb	geometric mean, $w_{p,d} = 0.5$
Repetition identification/selection		
allowed stripe distortion (step size)	sts	default={(1,2), (2,1), (1,1)}
minimum repetition length	mnl	minimum = 5 notes
predict boundary	sb	starting points of selected repetitions
setting for λ_3	$p\lambda_3$	temporal gap component of the LBDM segmenter (Cambouropoulos 2001); $k = 0.4$
setting for λ_2	$p\lambda_2$	$n = 4$

Table 4.2: MUL parameter settings.

Previous experimentation showed that using either thresholding or smoothing methods is detrimental to the performance of MUL.[30] Hence, in our experiments we use clean, non post-processed SMs. Moreover, we also tested eight melody representation schemes and eight similarity measures, yet no combination of representation scheme and similarity meaure resulted in statistically significant improvements over other combinations. Hence, we opt for a commonly used representation scheme: chromatic pitch interval (`cp-iv`) and inter-onset-interval ratio (`ioi-r`). Similarity is measured using the widely employed cosine similarity. We combine the pitch and duration representation using a geometric mean. That is, if we take $\mathcal{S}_p$ as an SM constructed using pitch information, $\mathcal{S}_d$ as an SM constructed using duration information, the geometric mean is computed as $(\mathcal{S}_p w_p \circ \mathcal{S}_d w_p)^{\circ \frac{1}{2}}$, with $\circ$ denoting the Hadamard or element-wise product.

[30] We tested both standard thresholding and the thresholding method provided in the SM toolbox (with the default parameters). We also tested Gaussian smoothing with window sizes $\in \{2, 3, 6\}$ notes.

Database	Folk Vocal				Folk Instrumental				Jazz			
Segmenter	$\overline{R}$	$\overline{P}$	$\overline{\mathbf{F1}}$	$\overline{\mathbf{B}}$	$\overline{R}$	$\overline{P}$	$\overline{\mathbf{F1}}$	$\overline{\mathbf{B}}$	$\overline{R}$	$\overline{P}$	$\overline{\mathbf{F1}}$	$\overline{\mathbf{B}}$
ORG	0.36	0.34*	0.33*	0.37*	0.32	0.25*	0.26*	0.30*	0.29	0.29*	0.25*	0.30*
ORG$\overline{\lambda}_{123}$	**0.48**	°0.53	°**0.47**	°**0.50**	0.39	°0.48	°**0.39**	°**0.45**	0.37	°0.45	°**0.37**	°**0.43**
ORG$\overline{\lambda}_{12}$	0.43	0.49	0.42	0.47	0.27*	0.37*	0.28*	0.33	0.28*	0.34*	0.27*	0.33
ORG$\overline{\lambda}_{13}$	0.44	°0.54	0.45	0.46	0.33	°0.51	0.37	0.39	0.30	°0.49	0.36	0.39
ORG$\overline{\lambda}_{23}$	0.43	0.38	0.38	0.42	°**0.43**	0.33	°0.33	°0.39	**0.41**	0.30	0.32	0.35
FIT$\overline{\lambda}_{123}$	0.39	°**0.59**	0.43	0.48	0.27*	°**0.57**	0.34	0.43	0.27*	**0.55**	0.33	0.40
FITλ_1	0.29	0.40	0.31	0.36	0.16	0.31	0.20	0.25	0.27	0.28	0.21	0.26
FITλ_2	0.37	0.25	0.29	0.34	0.42	0.18	0.25	0.31	0.40	0.20	0.26	0.29
FITλ_3	0.38	0.44	0.35	0.40	0.27	0.52	0.31	0.38	0.22	0.49	0.29	0.34
RND10%	0.17	0.25	0.20	0.22	0.17	0.19	0.17	0.15	0.17	0.25	0.20	0.21
ALWAYS	1.00	0.10	0.17	0.16	1.00	0.02	0.04	0.03	1.00	0.10	0.17	0.16
NEVER	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.3: Performance of MUL variants and baselines (abbreviations are defined in the text). From left to right: mean recall $\overline{R}$, precision $\overline{P}$, $\overline{F1}$, and $\overline{B}$ Highest performances are marked in bold. $*$ indicates performances that *are* significantly different ($\alpha = 0.05$) to the *highest* performance. $\circ$ indicates performances that *are* significantly different ($\alpha = 0.05$) to the ORG performance.

4.6.4 Results, Baselines, and Significance Testing

In Table 4.3 we present mean recall $\overline{R}$, precision $\overline{P}$, and $\overline{F1}$ results obtained by the different variants of MUL. Phrase boundaries are considered as predicted correctly (a *tp*) if the prediction identifies either the last event of an annotated phrase or the first event of the following phrase.

The variants of MUL in Table 4.3 are abbreviated as follows: ORG corresponds to the original version of MUL, and hence computes the fitness ϕ using Eq. 4.3; ORG$\overline{\lambda}_{123}$ computes ϕ using Eq. 4.8; ORG$\overline{\lambda}_{ij}$ also computes ϕ using Eq. 4.8, but this time the mean $\overline{\lambda}$ is computed over the pairs $ij \in \{12, 13, 23\}$; FIT$\overline{\lambda}_{123}$ uses the mean $\overline{\lambda}$ of scores $\lambda_{1,2,3}$ instead of ϕ as a fitness function; finally, FITλ_i, for $i \in \{1, 2, 3\}$, uses λ_i instead of ϕ as a fitness function.

To define a lower bound of performance we tested three naïve baselines: RND10%, which predicts a segment boundary at random in 10% of the melody (10% approximates the mean number of estimated boundaries produced by the tested variants of MUL); ALWAYS, which predicts a segment boundary at every melodic event position; NEVER, which does not make predictions (for completeness). We also tested the statistical significance of the paired $F1$, B, P, and R differences between the compared configurations of MUL and the baselines. For the statistical testing we used a non-parametric Friedman test ($\alpha = 0.05$). Furthermore, to determine which pairs of measurements significantly differ, we conducted a post-hoc Tukey HSD test.

4.6.5 Discussion

In this section we analyse the results shown in Table 4.3. We first discuss aspects related to the general performance of MUL. Then we discuss more specific aspects of performance: the possible benefits of using our location information, and the relative importance of each location constraint scoring function $\lambda_{1,2,3}$.

Note: The performance estimates for $\overline{B}$ are in all cases higher than the $\overline{F1}$ estimates. We believe the B estimates are more likely to match segmentation performance as judged by humans than those of the $F1$. However, to make the analysis easier to follow, in the following sections we focus on discussing $F1$, P and R performances. Moreover, due to annotation issues (discussed in depth in §3.3.1), determining if a machine estimated boundary is a false positive or not is not possible. Therefore, our discussion tends to favour precision over recall. We focus on comparing only very related approaches to be able to have some grounds in assuming a segmenter with high precision is in fact 'better' than one with high recall.

General performance observations

First, the performances obtained for vocal melodies are in general higher than those obtained for instrumental melodies. However, the $F1$ performance differences between each MUL variant for vocal and instrumental melodies are not statistically significant. This suggests that MUL generalises to these two sets. The performance drops a bit more with Jazz melodies. In this case the difference is statistically significant, suggesting that the model is more appropriate to folk music than other styles. Second, for both folk and jazz melodies $\overline{P}$ is tends to be higher than $\overline{R}$ ($\sim$9% higher for vocal melodies, $\sim$11% for instrumental melodies, $\sim$12% for jazz melodies). This can be explained by recalling that MUL models only repetition-based segmentation cues, while the annotated boundaries might have been perceived taking into account other cues. This is even more pronounced for Jazz melodies, due to the higher likelihood of having forms other than strophic. Third, all pairwise $F1$ performance differences between MUL variants and baselines showed to be significant at the 5% level.

Benefits of location constraints

For both folk and jazz melodies ORG$\overline{\lambda}_{123}$ obtains the highest performance. The $\overline{F1}$ improvements of ORG$\overline{\lambda}_{123}$ over ORG are of 14% in the vocal set, 13% in the instrumental set, and again 14% in the jazz set. For both sets their $F1$ performance differences are statistically significant. These significant improvements support our hypothesis, suggesting that location constraints are an important addition when attempting to discern which repetitions human listeners might recognise and use for segmentation. Furthermore, the fact that the differences in $F1$ performances between FIT$\overline{\lambda}_{123}$ and ORG (for both sets) are *not* significant stresses the level of importance of location constraints. To be more precise, while for all (vocal/instrumental/jazz) sets the $\overline{R}$ of FIT$\overline{\lambda}_{123}$ is comparable to that of ORG (R differences are not significant), the $\overline{P}$ of FIT$\overline{\lambda}_{123}$ shows large and statistically significant improvements over ORG, suggesting that the human annotators of the melodic datasets might be recognising repetitions by using location constraints in a greater degree than constraints on repetition frequency or length.

Role of location constraints 1, 2, and 3

In both folk and jazz sets the $\overline{F1}$ performances of each variant FIT$\overline{\lambda}_{1,2,3}$ is similar, with the best one for both sets being FIT$\overline{\lambda}_3$. For all sets the difference between the $F1$ performances of FIT$\overline{\lambda}_3$ and FIT$\overline{\lambda}_1$ is significant, and the one between FIT$\overline{\lambda}_3$ and FIT$\overline{\lambda}_2$ is not. Moreover, when $\lambda_{1,2,3}$ are used in combination in FIT$\overline{\lambda}_{123}$, the

$F1$ performances of FIT$\overline{\lambda}_{123}$ are not significantly different to those of FITλ_2 and FITλ_3. This suggests that the impact of repetitions aligned to temporal gaps λ_3 and repetitions with instances at the beginning of the melody λ_2 is higher than that of having evenly distributed repetitions λ_1. That said, it is only when all location constraints are used (ORG$\overline{\lambda}_{123}$) that a significant performance increase over ORG is obtained. This suggests that, even though in isolation $\lambda_{2,3}$ seem to have higher importance than λ_1, when associated to other constraints, such as repetition frequency and length, all location constraints become essential.

4.7 Conclusions

In this chapter we have proposed a set of location constraints for repetition based modelling of melody segmentation. Our proposed constraints aim to enhance repetition selection of repetition based segmenters. To test our constraints, we quantified and incorporated them in a state-of-the-art repetition based segmenter (Müller et al. 2011, 2013). The original and constraint-extended versions of MUL are used to segment melodies from the `FJ375` corpus. Results show the constraint-extended version of the segmenter achieves a statistically significant 14% average improvement over the model's original version. This suggests the influence of location information on human repetition recognition to be much more important than previously thought.

Future Work

Influence of Tonal and Metrical Structure. In future work the role of metrical structure has to be taken into consideration. As shown in (Ahlbäck 2007), even exact repetitions of melodic material might not be recognised by humans if these are not congruent with the metric structure of the melody. We also plan to extend our analysis to audio data, given that the constraints proposed in this chapter are independent of the representation scheme (although the robustness of temporal gap detection in automatically extracted onset information would need to be assessed).

Chapter 5

Contrast Based Segmentation

In this chapter we tackle the problem of automatically segmenting melodies into phrases via contrast identification. We focus on investigating the role of attention and multi-scale perception when determining contrasts.

Chapter Contributions

We introduce a novel approach to model automatic contrast identification based on statistical hypothesis testing. We tackle attention modelling using methods from information theory and statistical model selection. We tackle multi-scale modelling using methods from classifier combination.

To test the ability of our segmenter to identify segmentation determinative contrasts, we use it to segment melodies from the `FJ375` corpus. Results show that our segmenter achieves a statistically significant 10-12% improvement in precision in respect to the reference segmenters. If our segmenter is combined with a gap segmenter the combination achieves a statistically significant 10-20% average F1 improvement over the reference segmenters.

This chapter extends work presented in (Rodríguez-López and Volk 2012).

5.1 Introduction

In this chapter we tackle the problem of segmenting complete melodies into phrases by modelling contrast cues.

Main Concepts. A music fragment is in *contrast* to an immediately preceding fragment if a listener considers that the latter significantly alters a trend established by (or within) the former.[31] *Contrast cues* refer to the identification of contrasts during a segmentation process. *Contrast based segmentation* refers to machine segmentation approaches where it is assumed that contrast cues play a central role on the detection of segment boundaries. More specifically, that the starting point of the contrasting fragment is likely to be a segment boundary.

Modelling Tasks. A contrast-based segmenter requires (I) the formalisation of a sequential processing approach such that, for each time step, some model of the immediate past is compared to a model of the present (or immediate future if it is assumed the listener has heard the piece before), (II) the formalisation of the likely mental description(s) of the music to be compared, (III) the formalisation of a selection mechanism to model attention, i.e. that chooses the description(s) of the music to which the listener is likely to attend, and (IV) the formalisation of a comparison function to measure the amount of perceived contrast between the descriptions of past and present.

Focus. Research in contrast-based segmentation has focused mainly on tasks I, II, IV. Modelling of attention (task III) is most often not addressed. Hence, existing contrast segmenters are generally non-adaptive. That is, their parameters are set manually, at initialisation, and remain constant through the analysis. This non-adaptivity feature has been noted to have a negative effect on performance – see for instance discussions by Kaiser and Peeters (2013); Lartillot et al. (2013).

Contributions. We model contrast-based segmentation using a multi-resolution analysis based on statistical hypothesis testing techniques. We tackle attention modelling using methods from information theory and statistical model selection. Our segmenter is hence able to estimate parameter settings automatically, at run time.

We evaluate our segmenter on the `FJ375` corpus. To have a comparison point we also

[31] An example of the usage of the term contrast for the musicological analysis of antecedent-consequent phrase structures is as follows: *"If the direction of the melodic line in the consequent phrase differs from the direction of the melodic line in the antecedent phrase, the period is said to be in contrasting construction. The rhythm in both phrases may be similar or even identical, but if the melodic direction is different in each phrase, the period is nevertheless identified as being contrasting"* (Stein 1979, p. 42).

evaluate three existing contrast-based segmenters and two naïve baseline segmenters on the same corpus. Results show that our segmenter achieves a statistically significant 10-12% improvement in precision in respect to the reference segmenters. If our segmenter is combined with a gap segmenter the combination achieves a statistically significant 10-20% average F1 improvement over the reference segmenters.

Chapter Structure. This chapter is organised as follows. In §5.2 we discuss contrast cues in more detail to further motivate our approach. In §5.3 we review previous work on contrast-based segmentation. In §5.4 we describe our proposed approach to model contrast-based segmentation. In §5.5 we describe the experiments conducted to test our segmenter and discuss results. Finally, in §5.6 we present our conclusions and outline future work.

5.2 Discussion on Contrast Cues

In §2.3.2, page 20 we discussed contrasts as a class of segmentation cues related to similarity processing. More specifically, to instances where human listeners judge neighbouring fragments in a piece as being dissimilar/different. In this section we motivate two factors that are often no considered when modelling differences: multi-scale perception and attention. We discuss each factor in turn below.

Multi-scale perception. We posit that segmentation-influencing differences are perceived at variable (and on occasion multiple) time scales. We use the terms gap and contrast to make a rough distinction of the scales commonly used when studying difference perception (for segmentation) in MIR and CMMC. Gaps refer to differences for which short temporal contexts (roughly 2-4 notes long) are though to be enough for their perception. Contrasts refer to differences for which larger temporal contexts are necessary. (While we don't discard other scale distinctions, in this chapter we focus on the two mentioned ones.)

There is empirical evidence of both gaps and contrasts to occur at (or near) phrase boundaries – see (Deliège 1987) for the former, and (Spiro 2007) for the latter.[32] However, machine gap detectors are known to locate more gaps than there are phrase boundaries (oversegment) – see for instance (Temperley 2001, pp. 69-68) for a discussion on the oversegmentation issues of temporal gap segmenters. Conversely, contrast segmenters have shown to be more selective (Ferrand et al. 2003a; Rodríguez-López

[32] Spiro tested 'changes', some of which correspond to our notion of contrasts. It must noted that, differently from the study of Deliège, Spiro's is by no means exclusive to gap/contrast cues. However, to the best of our knowledge there are no perceptual studies focusing exclusively on contrast cues for phrase segmentation.

and Volk 2012). This duality provides some support to the idea that multiple scales are at play when segmenting phrases by difference. Spiro (2007, pp. 375–377) shares our position. She argues that in phrases gap perception is (seemingly) immediate, while contrasts are most often perceived retrospectively. Spiro then posits that the latter provide confirmatory information. That is, when a gap is perceived, it triggers a hypothesis of boundary location, then if a contrast also occurs the hypothesis is confirmed.

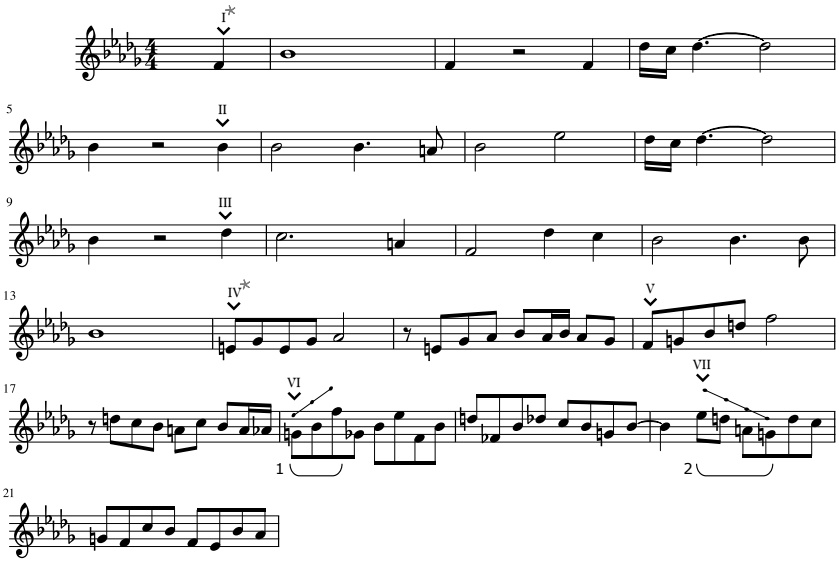

Figure 5.1: Fragment of Black and Tan Fantasy (1927) by Duke Ellington and Bubber Miley. Head (bars 1-13) and interlude (bars 14-21). Part of the JTC corpus (see §3.4.1). Arrows mark phrase starts, asterisks mark section starts. The segmentation was produced by an amateur musician.

It must be noted that we are not simply talking about points where boundaries of segments of different time spans coincide (say phrases and form sections). For instance, in bars 19–20 or Figure 5.1 one could argue that the temporal gap separating phrases *VI* and *VII* might be too 'weak' for it to be determinant of boundary perception. We can then argue that the change between the predominant presence of figure 1 (three notes in ascending motion) to that of figure 2 (four notes in descending motion), produces a sensation of contrast, which in turn confirms to the listener the beginning of figure 2 as a phrase start.

Attention. When determining contrast it is reasonable to expect attention shifts. For instance, in Figure 5.1 we could argue that perception of IV might be cued by the

change in the temporal density of notes, while some moments later boundaries VI and VII might be cued by the introduction of repeating figures 1 and 2. These possible sources of attention drift have been a recent focus in music psychology studies using boundary annotated databases (Smith et al. 2013b; 2014; 2015). It is then necessary for machine contrast detectors to have access to different representations of the input, and be able to decide (during processing) which representation might be more likely to influence segmentation.

5.3 Related Work

Existing contrast-based segmenters have been mostly designed to segment music (audio) recordings into form sections.[33] Automatic form-level segmentation of music recordings has been an active and popular topic of research for the better part of a decade. Consequently many approaches (including many contrast-based ones) have been proposed. For brevity we refer to (Paulus et al. 2010) for individual segmenter descriptions,[34] and focus on first describing popular approaches, and then providing a general discussion of their parameter automation characteristics.

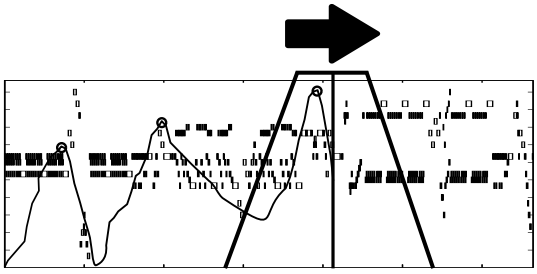

Figure 5.2: Sliding window technique applied to the piano roll of *12th Street Rag* (1914) by E. Bowman (part of the `JTC` database).

Approaches. Contrast-based segmenters often use of (some variation of) a 'sliding window' technique to detect boundaries – see Figure 5.2 for an illustration. This technique consists of first defining a time window within which evidence for a contrast

[33] To the best of our knowledge using symbolic formats of music for for contrast-based form section segmentation are Chew (2006); Zanette (2007). For a description and analysis of these approaches refer to (Rodríguez-López and Volk 2012).

[34] Paulus uses the term 'novelty-based' to refer to segmenters modelling difference cues. In this dissertation we purposely chose a different term. The novelty-based class, as used in the publication, does not seem to distinguish between short term and long term difference. Segmenters in this class are (implicitly) treated as multi-purpose machines. We, conversely, choose to treat segmenters modelling long and short term difference perception as belonging to different classes.

is to be searched. Then 'sliding' the window across the input piece, from beginning to end, in equal (and normally small) step-sizes. For each time step contrast within the window is estimated. To that end window is split into two sections, for each section the music contained therein is described in some way, and the descriptions are compared using some similarity/distance measure. Dissimilar splits are given a high score. This results in a profile of the piece in which peaks are assumed to mark points of significant contrast. A peak picking algorithm is used to select peaks, and their locations are taken as boundary estimates.

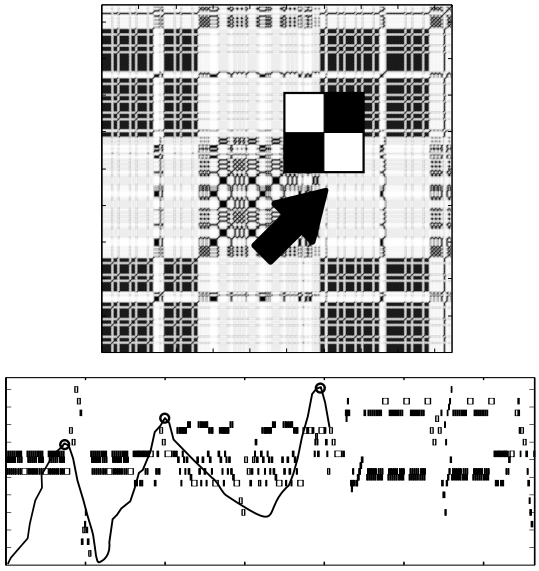

Figure 5.3: Contrast-based segmentation of the theme of *12th Street Rag* based on similarity matrix processing. The matrix was computed from a piano roll rendition (bottom) using a ioi-ratio representation of the melody. Dark indicates high similarity. As in Figure 5.2 a sketch of the resulting contrast profile is superimposed over the piano roll input, and selected peaks have been circled.

A popular variation of the sliding window technique is based on similarity matrix (SM) processing – see Figure 5.3 for an illustration. In SM depictions of music the edges of block-like shapes have been shown to coincide (or be near) to segment boundaries (Smith et al. 2010; 2013). The problem of finding contrasts then is posed as that of detecting block edges. To that end a subsection of the matrix around the diagonal is compared with an ideal representation of a block edge, modelled as a 2x2 checkerboard. The checkerboard is slide across the diagonal, and a score at each point is computed. Sections that resemble the checkerboard structure get a high score. Just

as with the standard sliding window approach, step-by-step scoring results in a profile where peaks indicate points of high contrast. Peaks are then automatically selected and their locations are outputted as boundary estimates.

Parameter Handling in Form-level Contrast Segmenters. Form section segmenters (whether SM based or not) are most often completely handcrafted. The set of music attribute representations is normally small and predefined.[35] Window/checkerboard settings (such as length, split position, and shape) are defined at initialisation, and remain constant through the analysis. The same is true for the parameters of smoothing and thresholding methods that are often applied during pre/post processing. For the reasons outlined in §5.2, the single-scale, static parameter feature of form-level contrast segmenters is likely to negatively affect their performance when segmenting at phrase granularity.

Parameter Handling in Phrase-level Contrast Segmenters. Not much research has focused on proposing contrast segmenters for phrase segmentation. Table 5.1 provides a summary of the only investigations that, to the best of our knowledge, do so.[36] All approaches listed tackled the segmentation of melodies. And all focus on modelling contrasts at a single scale. Rodríguez et al. (2012) is the only one that used a technique to select between different viewpoints of the melody. However, viewpoint selection is preformed at initialisation, and it is not recomputed during processing.

Segmenter Author(s)	Window Length used	Search Process	Input Representation
Ferrand et al. (2003a)	4-8 notes	sliding window	{`on`, `pi-iv`}
Rodríguez-López and Volk (2012)	all past, all future	top-down	{`cp`, `ioi`, `ioi-rc`, `cp-cl`}
Velarde et al. (2013)	1-2 notes to 1 bar	sliding window	{`cp` }

Table 5.1: Reviewed contrast-based segmenters. For abbreviations see Table 2.3.

[35] Two exceptions are (Turnbull et al. 2007; Ullrich et al. 2014), who used supervised learning to automate some aspects of the traditional (non-SM based) sliding window approach. For both cases the idea is to take a sample of temporal contexts around annotated boundaries, and then train a classifier to distinguish the degree of difference necessary for a boundary to be perceived. While both approaches have show to perform better than their unsupervised counterparts, they are still limited by one big factor: annotated data sparsity. Existing annotated music collections are limited in size, and are often annotated by at most three experts. This is specially true for collections annotated with phrase level boundaries. Producing human segmented data is a very time consuming and error prone task. It is unlikely that in the short term the amount of training data will scale to the sizes necessary to ensure generality and flexibility for supervised segmenters.

[36] It also must be noted that, while the approaches of Ferrand et al. and Velarde et al. can in principle be used to model contrasts, the focus of their research only implicitly addresses it.

5.4 Approach

We take a hypothesis testing approach to modelling contrast. We assume regions of a piece/melody perceived as homogeneous can be modelled as regions with a quasi-stationary probability distribution. That entails that boundary-cueing contrasts can be determined by testing if contiguous regions have distributions that significantly differ. We use this idea to propose a multi-scale approach based on the sliding window technique. In the following subsections we first give an overall view of our approach, then present a description of the process for a single scale, and, lastly, describe the process for multiple scales.

5.4.1 Overall Description

Figure 5.4 shows a diagram of our approach. The user chooses a number of time scales (window sizes) for analysis, and a number of viewpoints of the melody to be analysed. The segmenter then proceeds by computing, for each scale, contrast profiles for each viewpoint. The viewpoints are then merged into a single scale profile. The merging process gives more weight to viewpoints that are likely to stimulate listener's attention. Weights are computed for each time point, thereby modelling online, dynamic attention. Each resulting scale profile is then merged into the output profile, using techniques from model selection and data fusion.

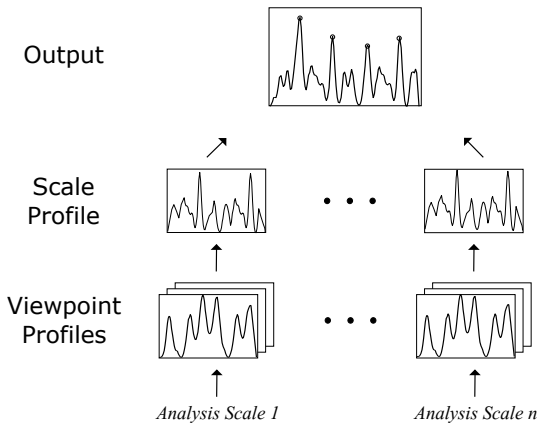

Figure 5.4: Overall view of our approach to contrast-based segmentation.

5.4.2 Analysis of a Single Scale

The analysis of a single scale is directly based on the sliding window technique. This method requires specifying (a) the length of the window, (b) the length of the unit for which statistics are going to be collected (single notes, note sequences), (c) the musical attribute representing the unit (pitch class, pitch interval, duration ratio, and so on). Analysis parameter (a) is specified by the user. Parameters for (b) and (c) are inferred by the machine.

Automatically estimating parameters (b) and (c) require tackling of both cognitive modelling issues and technical issues. The former relate to choosing an appropriate music representation and modelling attention drift. The latter to issues related to estimating probability distributions and significance from small-sample statistics. Below we first formally describe the sliding window technique for melodies, and then describe in turn the techniques used to tackle automatic parameter estimation.

Sliding Window Technique for Melodies in Symbolic Format

In this dissertation melodies are represented as a sequence of symbols. Each symbol represents an attribute describing either a note (e.g. its pitch or duration) or a short sequence of notes (e.g. a chromatic pitch interval or inter-onset-interval ratio). Let then $\mathbf{x} = x_1 \dots x_N$ be a sequence of melodic symbols of length N. Each symbol $x_i \in \mathcal{A}$, where $i \in [1 : N]$ and $\mathcal{A}$ is an alphabet representing the space of possible discrete categories use to mentally represent the attribute.

The task is to locate which positions i in the sequence are likely points of perceptually salient contrast. We define a window $\mathbf{w}$ of length $2M$. We slide the window along the sequence, in steps of one symbol. At each step we need to refute the hypothesis that $\mathbf{w}$ is *statistically stationary*. In other words, the likelihood that $\mathbf{w}$ was generated by a single probability distribution. We then bisect $\mathbf{w}$ into two subsequences $\mathbf{w}_L = [i-M, i)$ and $\mathbf{w}_R = [i, i+M)$. Statistics for $\mathbf{w}$ and $\mathbf{w}_{L,R}$ are collected and distributions estimates $\hat{p}$ and $\hat{p}_{L,R}$ are computed. A statistical test is used to quantitatively measure the degree to which $\hat{P}_1$ (from $\hat{p}_{L,R}$) provides a better probabilistic description of the windowed fragment of music than $\hat{P}_2$ (from $\hat{p}$). A high value indicates the fragment was more likely generated by two distinct distributions, separated at the split point. As in the standard sliding window technique, computing this test for every position in the attribute sequence results in a profile were peaks are assumed to indicate contrast-cued segment boundaries.

Estimating and Comparing Probability Distributions

We need to define a process to estimate probability distributions $\hat{P}$ and $\hat{P}_{L,R}$. When doing so, we want to have the flexibility to collect statistics for both single symbols (one or two notes) and symbol sequences (two notes or larger). To do so we model the melodic attribute sequence $\mathbf{w}$ and $\mathbf{w}_{L,R}$ as Markov chains of order K (with $K = 0$ single symbol, $K = 1$ two symbols, and so on).

We then count order-K symbol sequences, taking into account that

$$c_{\mathbf{t}a} = c^L_{\mathbf{t}a} + c^R_{\mathbf{t}a}, \quad \text{and} \sum_{\mathbf{t}\in\mathcal{A}^K}\sum_{a=1}^{A} c_{\mathbf{t}a} = N \tag{5.1}$$

Here A is the size of the alphabet $\mathcal{A}$, and $\mathbf{t}$ is shorthand notation for the K-tuple of symbols. The counts $c_{\mathbf{t}a}$, $c^L_{\mathbf{t}a}$, and $c^R_{\mathbf{t}a}$ are then the number of times figures of length $K+1$ appear in sequences $\mathbf{w}$, $\mathbf{w}_L$, and $\mathbf{w}_R$, respectively. With the counts we compute maximum-likelihood probabilities as

$$\hat{p}_{\mathbf{t}a} = \frac{c_{\mathbf{t}a}}{\sum_{a'=1}^{A} c_{\mathbf{t}a'}}, \quad \hat{p}^L_{\mathbf{t}a} = \frac{c^L_{\mathbf{t}a}}{\sum_{a'=1}^{A} c^L_{\mathbf{t}a'}}, \quad \hat{p}^R_{\mathbf{t}a} = \frac{c^R_{\mathbf{t}a}}{\sum_{a'=1}^{A} c^R_{\mathbf{t}a'}}, \tag{5.2}$$

and estimate probabilities for the complete and bisected fragment as

$$\hat{P}_1 = \prod_{\mathbf{t}\in\mathcal{A}^K}\prod_{a=1}^{A}(\hat{p}_{\mathbf{t}a})^{c_{\mathbf{t}a}}, \quad \hat{P}_2 = \prod_{\mathbf{t}\in\mathcal{A}^K}\prod_{a=1}^{A}(\hat{p}^L_{\mathbf{t}a})^{c^L_{\mathbf{t}a}}(\hat{p}^R_{\mathbf{t}a})^{c^R_{\mathbf{t}a}} \tag{5.3}$$

To measure how much better $\hat{P}_2$ models $\mathbf{w}$ than $\hat{P}_1$, we use the general *Jensen-Shannon divergence* (JSD), which is defined as:

$$\Delta = log(\frac{\hat{P}_2}{\hat{P}_1}) = \sum_{\mathbf{t}\in\mathcal{A}^K}\sum_{a=1}^{A}[c_{\mathbf{t}a}\ log\ \hat{p}_{\mathbf{t}a} + c^L_{\mathbf{t}a}\ log\ \hat{p}^L_{\mathbf{t}a} + c^R_{\mathbf{t}a}\ log\ \hat{p}^R_{\mathbf{t}a}] \tag{5.4}$$

The JSD was introduced by Lin (1991) as a symmetric variant of the Kullback-Leibler divergence. Lin's JSD formulation was only defined for zero-order distributions. Equation 5.4, introduced by Thakur et al. (2007), presents a generalised version

of the JSD, which can handle higher orders. We choose to use the JSD over other divergence measures proposed in the literature due to its properties of symmetry, non-negativity, finiteness and boundedness, which are desirable in our context – see (Lin 1991; Grosse et al. 2002) for detailed discussions of these properties.

Representation Issues: Selecting an Appropriate K

When processing a windowed fragment, Markov models up to a (user defined) maximum K are computed. We want our segmenter to estimate the optimal order K of the Markov model. In other words, we want it to estimate the length of the unit (note, note sequences) which is more appropriate for statistically describing the music within the window. This problem is, however, fraught with uncertainty. It is not clear a priori if the window actually contains statistically distinct areas. Moreover, if it does, it is also not known whether these areas differ in respect to statistics collected from low-order models (notes, pairs of notes) or high-order models (figures). In the field of mathematical statistics the problem of determining model parameters in the presence of this kind of uncertainty is called *model selection* (Claeskens and Hjort 2008). Several techniques to determine the suitability of a given model (order) in a given situation have been proposed. They are collectively known as *information criteria* (Konishi and Kitagawa 2008). Information criteria normally consist of computing a penalty score ϕ for the sequence likelihood estimate $\hat{P}$ estimated with a model of order K as

$$\phi(K) = -2P(K) + \theta, \tag{5.5}$$

where θ is a penalty function that depends on the specific criterion. In our segmenter we test three popular information criteria: the Akaike information criterion (AIC) (Akaike 1974), the Schwarz information criterion (SIC) (Schwarz 1978), and the Bayes information criterion (BIC) (Katz 1981). For a Markov chain of length N and order K, over an alphabet of S symbols, their penalty functions are

$$\theta(N,S,K)\begin{cases} S^K(S-1) & AIC; \\ \frac{1}{2}S^{K+1}logN, & SIC; \\ S^K(S-1)logN, & BIC; \end{cases} \tag{5.6}$$

We then select the $P(K)$ that minimises ϕ as

$$K^* = \underset{K}{\operatorname{argmax}}\, \phi(K). \tag{5.7}$$

Representation Issues: Attention Drifting

When processing at one scale, a profile is computed for each (user defined) melody representation – refer to Figure 5.4 for a depiction. We want our segmenter to take into account the possible drifts in attention, from one representation to the other, that human listeners might experience as they hear the input melody. We consider a simplified case of attention drift. We assume situations where listeners do not willingly/consciously switch their attentional focus. Rather, that attention is stimulated by music content factors, to which listeners simply react. To model attention drift profiles are merged using a weighted average scheme. Weights should ideally put attention-stimulating representations in the foreground.

Weights are calculated at each step of the computation of each profile (i.e. for each windowed fragment) as

$$\alpha = (H(d)/H(S))^b, \tag{5.8}$$

where d is a probability distribution, S is the size of the alphabet, $b \in Z_{\geq 0}$ is a bias term and $H(d) = -\sum_{i=1}^{S} d_i \log_2 d_i$ is the Shannon entropy. In the context of our segmenter, the input distribution d corresponds to the probability estimate $P(K^*)$. Entropy weights α give preference melodic representations which are less predictable, which researchers have posited might influence attentional drift between mental representations of music (Ferrand et al. 2002; Madsen et al. 2008; Culpepper 2010).[37]

We test three averaging techniques: geometric, arithmetic, and maximum (Tax et al. 2000; Kuncheva 2002). The geometric average emphasises points of agreement, and penalises points if disagreement (essentially a conservative type of voting). The arithmetic average emphasises aggregated agreement (essentially a type of majority voting). The maximum average selects only the maximum of all profiles at each point.

[37] Entropy-based measures have also been used with a fair degree of success to study attention related musical notions such as style and interestingness (Snyder 1990; Margulis and Beatty 2008; van Balen et al. 2015; Jensen and Hebert 2015).

Technical Issues: Dealing with Statistical Noise

In addition the automation of parameters K and α, which can be seen tackling cognitive modelling issues, we also need to tackle issues related to the statistical approach itself. Technical issues are related to the fact that we are estimating probabilities by collecting statistics from small samples. Using small samples is known to result in unrealiable probability distribution estimates. As a concequence we have to deal with 'noisy' profiles, where not all peaks can be trusted to be perceptually meaningful. We employ smoothing and bootstrapping techniques in an attempt to emphasise meaningful peaks and remove spurious ones. Below we describe each technique in turn.

Smoothing. We test two widely used smoothing techniques and one created specifically for statistical change detectors. The first two are a *moving-average* smoother and a (one-dimensional) *Gaussian kernel* smoother. The third is a *mean-field kernel* smoother, proposed by Cheong et al. (2009) for divergence profiles of biological sequences. Cheong's kernel is defined as

$$mf(i) = \begin{cases} (1 + \frac{z}{n})^2, & z < 0; \\ (1 - \frac{z}{n})^2, & z \geq 0; \end{cases} \tag{5.9}$$

where n is the length of the kernel, $i \in [1, n]$, $z = i - c$, and c is the centre of the kernel. To specify the kernel's shape, Cheong's studied divergence profiles derived from artificial sequences (where change points were known in advance). The kernel is an approximation of the observed curve shapes around change points.

Bootstrapping. In the profiles obtained by computing local, windowed JS divergences, it is assumed that large values indicate points of local contrast. However, the statistical significance of the estimates of divergence is not guaranteed. Unfortunately, there is no analytical form to estimate the significance of a JSD estimate.[38] We hence use a simple bootstrapping methodology to assess significance. The main idea is to compute JSD values for randomly ordered sequences. And assume the JSD estimate is significant if it is larger than the sum of the mean divergence (plus one standard deviation) computed for the random sequences.

To this end, random sequences are generated by resampling the original sequence (with replacement). The generated sequences are of identical length, and can be

[38] Grosse et al. (2002); Thakur et al. (2007) proposed Monte Carlo approximations, but both parameters and recommended values are specific to biological sequences.

expected to have similar symbol statistics as the original sequence. Then the JSD is computed for the generated sequence. The mean and standard deviation of the JSD values obtained are then calculated and stored. This results in a profile with divergence values that are likely to be obtained by chance.

We use the 'chance' profile to estimate the confidence of our divergence estimates. To do so we compute a residual profile as

$$\rho(i) = \begin{cases} \Delta(i) - \Delta^\star(i), & \Delta(i) - \Delta^\star(i) > 0; \\ 0, & \Delta(i) - \Delta^\star(i) \leq 0; \end{cases} \tag{5.10}$$

where Δ is the estimated profile, and $\Delta^\star$ is the chance profile. Values $\rho(i)$ can be seen as confidence estimates for the divergence estimates. We can hence obtain more reliable estimate of the divergence profile by weighting each value by its corresponding residual counterpart.

5.4.3 Merging Scale Profiles

The last stage of processing in our segmenter deals with merging the profiles created for each (user defined) temporal scale – refer to Figure 5.4 for a depiction. To merge scale profiles we test the same techniques used to merge representation profiles, namely, geometric, arithmetic, and maximum averaging.

5.5 Evaluation

We evaluate our segmenter in a traditional scenario, i.e. by comparing machine-estimated segment boundaries to human-annotated segment boundaries. The experiments described in this section have three main goals. First, benchmark our contrast-based segmenter against existing ones, to compare and validate our hypothesis testing approach to contrast cue detection. Second, analyse the extent to which our approach might be modelling contrast cues, to validate/reject the hypothesis that human contrast perception might depend on statistical processing mechanisms. Third, test if our contrast segmenter complements exiting gap segmenters, to validate/reject the hypothesis that human contrast perception might use different processing strategies at different time scales.

5.5.1 Preliminaries

In the following subsections we motivate our choice of reference and baseline segmenters, and describe our test corpus and evaluation metrics.

Selecting Reference Segmenters. As discussed in §5.3, not much research has targeted contrast detection for phrase level segmentation. From the few that do, summarised in Table 5.1, we implemented and tested (Ferrand et al. 2003a; Rodríguez-López and Volk 2012). We did not consider the segmenter of Velarde et al. (2013). Published results indicate that it appears to perform best at narrow time scales, being then closer to a gap detector than a contrast one. Thus it would not constitute an informative reference point.

Due to their popularity, we implemented and tested two similarity matrix based segmenters, designed to estimate form boundaries in music recordings by detecting change (gaps or contrasts depending on the kernel size setting). We selected the segmenter presented in (Cooper and Foote 2003), which is considered a landmark in the area. We also tested (Serrà et al. 2012) as an example of a recent and successful variation on the similarity matrix approach. It must be noted however, that Serra's segmenter was intended for the detection of both change and repetitions. Care is then taken when comparing and discussing performances between Serrà's segmenter and our own.

Baselines. To define a lower bound of performance we tested three naïve baselines: RND10%, which predicts a segment boundary at random in 10% of the melody (10% approximates the mean number of estimated boundaries produced by the tested segmenters); ALWAYS, which predicts a segment boundary at every melodic event position; NEVER, which does not make predictions (for completeness).

Test Dataset. All machine segmenters studied in this chapter are evaluated in the `FJ375` corpus – refer to §3.5 for a description.

Evaluation Measures. In this thesis we use the well known $F1$, precision P, and recall R measures, defined in Equations 3.1, 3.2, and 3.3, respectively. Phrase boundaries are considered as predicted correctly (a 'true positive') if the prediction identifies either the last event of an annotated phrase or the first event of the following phrase.

To take into account the possibility of fuzzy boundaries during evaluation (see discussion in §3.3.1), we also use the Boundary Edit Distance Similarity B, defined in Equation 3.4. One of the parameters of the B measure is a tolerance window (in notes). Within this tolerance window boundaries are given a partial score proportional

to their relative distance. We tested the effect of soft disagreement by computing the B using a tolerance of four notes (giving score also to points of soft agreement).

5.5.2 Parametric Settings

All tested segmenters produce as an output a contrast likelihood profile, from which boundaries need to be selected via automatic peak picking. We experimented with several peak selection algorithms, settling for the algorithm proposed in (Pearce et al. 2010b). This peak selection algorithm has only one parameter k. The optimal values of k are specified for each segmenter in Tables 5.2 and 5.3.

All tested segmenters consist of multiple modules, each with multiple parameters. In the rest of this section we list and discuss parameter settings for each segmenter.

Parameters		Setting used for experimentation
SER (Serrà et al. 2012)		
melodic fragment length	*fl*	fragment length of 3 notes
similarity measure	*sm*	Cosine similarity
melody representation	*mr*	{`cp-iv`, `ioi-r`}
matrix blending	*smb*	geometric mean, $w_{p,d} = 0.5$
matrix smoothing	*smt*	–
matrix thresholding	*thr*	$t = 1.2$
peak picking	*pp*	$k = 0.8$
COO (Cooper and Foote 2003)		
melodic fragment length	*fl*	fragment length of 3 notes
similarity measure	*sm*	Jaccard similarity
melody representation	*mr*	{`cp-iv`, `ioi-r`}
matrix blending	*smb*	geometric mean, $w_{p,d} = 0.5$
Kernel wifth	*krl*	{4,**6**,8,16,24} notes
peak picking	*pp*	$k = 0.8$
FER (Ferrand et al. 2003a)		
melody representation	*mr*	{`cp-iv`, `onsets`}
window size	*pp*	{4,6,**8**,16,24} notes
peak picking	*pp*	$k = 1$
ROD (Rodríguez-López et al. 2012)		
melody representation	*mr*	{`cp`, `cp-cl`, `cp-iv`}

Table 5.2: Parameter settings of reference segmenters.

Parametric Setting for Reference Segmenters

In Table 5.2 we present the parametric settings of the reference segmenters. (For conciseness, in this section we refer to them by the abbreviations given on the table.) All parameters were optimised for performance on the `FJ375` corpus.

Segmenters SER and COO were originally developed to take as input an `SM` computed from music recordings in audio format. Here we use `SM`s computed from melodies in symbolic format – refer to §4.4.1 for details on their construction. For both segmenters best results are obtained using a geometric blending of a pitch interval (step-leap) matrix with a `ioi`-ratio matrix. This representation setting agrees with the one used for repetition based segmenters in §4.6.3. SER uses various forms of smoothing and thresholding. In our experiments smoothing proved detrimental to performance, and so it was bypassed. Conversely, thresholding proved indispensable. The recommended setting in the original publication is 0.02. During manual optimisation this setting proved too strict, a more suitable setting was found for values around 0.4.

Parameters for the CON segmenter		Setting used for experimentation
Markov order K	K	0-1
information criterion	ic	BIC
divergence measure	dm	Jensen-Shannon Divergence
melody viewpoints	mr	{`cp-cl`, `cp-iv`, `cp-rc`, ... `iv-sl`, `ioi`, `ioi-r`}
time scales for analysis	ts	{8,16,24} notes
profile smoothing	sm	mean-field kernel smoothing
melodic viewpoint merging	vm	maximum averaging
time scale profile merging	tsm	geometric averaging
peak picking	pp	$k = 1$

Table 5.3: Best configuration and parameter settings for our segmenter. For melody representation abbreviations see Table 2.3

Configuration and Settings of our Segmenter

Our segmenter, henceforward CON for short, consists of multiple modules. For some of these modules we proposed multiple alternative techniques for experimentation. In Table 5.3 we list the modules, techniques, and parametric settings that maximised its performance on the `FJ375` corpus.

Choice of Melodic Viewpoints. We create zero-order models from four representations: pitch classes, pitch intervals (chromatic), pitch range classes, and `ioi`s. Divergences

in pitch class distributions may indicate changes in local key, mode, or chroma. Divergences in pitch intervals distributions may indicate changes in local interval usage. Divergence in pitch range classes may indicate changes in local melodic register or tessitura. Lastly, divergence in `ioi`s may indicate changes in local temporal note density and duration.

We also create higher-order models (up to 3 symbol sequences) for: pitch intervals (step-leap), and `ioi`-ratios. Divergences in these models may indicate local changes in short melodic figures and in rhythmic cells, respectively. For this attributes the choice of model order (figure length) is determined automatically using the BIC information criterion.

Combination of Contrast and Gap Segmenters

We test whether CON produces boundaries that complement those produced by a gap segmenter. For our experiments, we chose the LBDM gap segmenter, proposed by Cambouropoulos (2001). We use LBDM with its recommended settings. The combination between the CON and LBDM segmenters is performed by combining their output profiles with simple arithmetic weighting. The same peak picking algorithm as used for CON is used. A setting of $k = 1$ was found to be appropriate.

5.5.3 Results and Discussion

In Table 5.4 we present mean recall $\overline{R}$, precision $\overline{P}$, and $\overline{F1}$ results. We tested the statistical significance of the paired $F1$, B, P, and R differences between the compared segmenters. For the statistical testing we used a non-parametric Friedman test ($\alpha = 0.05$). Furthermore, to determine which pairs of measurements significantly differ, we conducted a post-hoc Tukey HSD test.

Note 1: In our discussion, when we say a difference is significant, we mean it is significant according to our hypothesis testing settings above.

Note 2: The performance estimates for $\overline{B}$ are in all cases higher than the $\overline{F1}$ estimates. We believe the B estimates are more likely to match segmentation performance as judged by humans than those of the $F1$. However, to make the analysis easier to follow, in the following sections we focus on discussing $F1$, P and R performances. Moreover, due to annotation issues (discussed in §3.3), determining if a machine estimated boundary is a false positive or not is at present unfeasible from our test corpora. Therefore, our discussion tends to favour precision over recall. We focus on comparing only very related approaches to be able to have some grounds in assuming

Database	Folk Vocal				Folk Instrumental				Jazz			
Segmenter	$\overline{R}$	$\overline{P}$	$\overline{\mathbf{F1}}$	$\overline{\mathbf{B}}$	$\overline{R}$	$\overline{P}$	$\overline{\mathbf{F1}}$	$\overline{\mathbf{B}}$	$\overline{R}$	$\overline{P}$	$\overline{\mathbf{F1}}$	$\overline{\mathbf{B}}$
CON	0.25	0.40	0.29	0.45	0.29	0.43	0.31	0.47	0.30	0.41	0.30	0.42
CON + LBDM	0.42	0.66	**0.49**	0.51	0.40	0.73	**0.39**	0.49	0.39	0.70	0.39	0.43
LBDM	0.33	0.44	0.33	0.46	0.35	0.68	0.35	0.40	0.39	0.74	**0.40**	0.42
FER	0.27	0.28	0.26	0.41	0.33	0.32	0.31	0.44	0.31	0.31	0.29	0.41
ROD	0.25	0.29	0.27	0.43	0.32	0.29	0.29	0.40	0.29	0.33	0.31	0.42
COO	0.20	0.14	0.16	0.39	0.24	0.19	0.20	0.37	0.24	0.20	0.21	0.39
SER	0.25	0.18	0.20	0.38	0.20	0.15	0.17	0.31	0.20	0.14	0.16	0.36
RND10%	0.17	0.25	0.20	0.22	0.17	0.19	0.17	0.15	0.17	0.25	0.20	0.21
ALWAYS	1.00	0.10	0.17	0.16	1.00	0.02	0.04	0.03	1.00	0.10	0.17	0.16
NEVER	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 5.4: Performance of CON variants, reference segmenters, and baselines (abbreviations are defined in the text). From left to right: mean recall $\overline{R}$, precision $\overline{P}$, $\overline{F1}$, and $\overline{B}$ Highest performances are marked in bold.

a segmenter with high precision is in fact 'better' than one with high recall.

In what follows we analyse the results shown in Table 5.4. We first discuss how CON compares to the selected reference segmenters and baselines. Then we discuss more specific aspects of performance: the possible benefits of using statistical modelling for contrast detection. Finally we discuss the performance of the combined contrast and gap model.

Comparison to Reference Segmenters

CON shows similar $\overline{F1}$ performances for all subsets of the `FJ375` corpus (differences are not significant). The same can be observed for the reference segmenters. This suggests a generalisation of contrast cues to all the three different melodic traditions contained in the `FJ375`. The $\overline{F1}$ performance of CON is well above that of the `SM` based segmenters COO and SER, but similar to that of FER and ROD.[39] However, CON shows improvements on $\overline{P}$ of up to 12%, which proved statistically significant. This suggest our proposed statistical testing approach might be more suitable to model contrast cues than the popular `SM` based approaches. Moreover, the $\overline{F1}$, $\overline{P}$, and $\overline{R}$ performances of CON also show to be well above the performances of the baseline segmenters, suggesting a fair distance to chance performance.

Possible Benefits of our Approach to Modelling Contrast

The observed increase in precision of CON in respect to the reference segmenters is specially high for the instrumental folk and jazz subsets of the `FJ375`. Since melodies in these two subsets are generally longer than in the vocal set, the improvement in $\overline{P}$ suggests a benefit of the use of multiple time scales. We then tested CON for only one time scale (window of 8 notes) and the margin of improvement diminished. This suggest that the use of multiple scales contributes to getting more 'fine tuned' boundary locations, without compromising $\overline{R}$. A second, test we performed was removing those melodic representations for which higher order (figure level) statistics where computed, namely pitch interval (step-leap) and `ioi`-ratios. $\overline{P}$ performance again lowered, giving some support to the idea that using wider temporal contexts and

[39] The performance of the `SM` segmenters is the lowest of amongst the segmenters tested in this chapter, ranging between 0.15-0.25 on the $\overline{F1}$ scale. This is somewhat lower that reported performances with audio datasets. However, reported performances most often correspond to evaluations of section level boundaries. For instance, evaluating on the 'large-scale' annotations of the SALAMI dataset. Our segment boundaries considered for our evaluation are closer to the 'small-scale' annotations of SALAMI. To our knoweldge the tested `SM` segmenters have not been evaluated on these annotations. Moreover, our tolerance level when computing F1 performances is more strict than the one used for audio – refer to §3.5 for a more extended discussion on this issue.

statistics based on larger 'melodic' units is of importance when estimating boundary-cueing contrasts. Lastly, we observed that removing the profile smoothing stage also lowered performance. And that removing the bootstrapping stage improved recall at the expense of performance. Both of these findings suggest that the techniques for eliminating statistical noise are of main importance when modelling contrast with a statistical approach.

Combining Contrasts and Gaps

When the Con is combined with Lbdm we observe 10-20% $\overline{F1}$ improvements over the reference segmenters. The improvement are in most cases significant (performances are statistically equivalent to those of Rod in the Jazz subset, and of Fer in the intrumental folk subset). When we compare the combination performances to those of Lbdm alone, we see large increases in both recall and precision for the instrumental folk and Jazz subset. On the other hand, performance increases are rather small for the vocal set. This can be expected, Jazz and Intrumental folk melodies are probably more inclined to contain distinct parts, both at the level of sections and phrases. (For instance, *abab* or *abcabc* type section structures are common in Jazz, while antecedent-concequent can phrase pair structures can be expected in both Jazz and intrumental folk.) The fact that we are observing large increases in recall is indicative of a complementary role of contrast. As hypothesised earlier, the role of contrast in phrase boundary perception might not just be additive, but rather confirmatory. That is to say, it might not just be that gaps and contrasts tend to coincide at section boundaries (which would have resulted only in an increase in precision), but rather that contrast is also present at the phrase level. And that the simple arithmetic combination (majority voting) of the profiles generated by Con and Lbdm was able to emphasise peaks captured by Lbdm but that were not large enough to be selected by the peak picking algorithm.

5.6 Conclusions

In this chapter we introduced a novel contrast-based segmenter, targeted to segmenting melodies into phrases. We proposed an approach to modelling segmentation cues based on statistical hypothesis testing. The main contributions in respect to existing contrast segmenters is the use of multiple scales and attention modelling modules.

We evaluated our segmenter on the `FJ375` corpus. To have a comparison point we also evaluated three existing contrast-based segmenters and two naïve baseline segmenters on the same corpus. Results show that our segmenter achieves a statistically

significant 10-12% improvement in precision in respect to the reference segmenters. If our segmenter is combined with a gap segmenter the combination achieves a statistically significant 10-20% average F1 improvement over the reference segmenters. This suggests that the perception of local difference has two facets, one at narrow time-scale and another at a wider time scale. And that both of these facets lead to cues (gaps and contrasts, respectively) which a seemingly complementary role in phrase boundary perception

Chapter 6

Template Based Segmentation

In this chapter we tackle the problem of segmenting melodic phrases into subphrases based on template cues. More specifically, we focus on modelling segmentation influenced by the recognition of previously heard melodic figures.

Chapter Contributions We introduce a statistical learning approach to learn melodic figures. We introduce a maximum entropy approach to model template-based segmentation.

We test the usefulness of the subphrases produced by our template-based segmenter in a classification experiment. In the experiment automatically determined subphrases are used as features to estimate the cultural origin melodies. We compare to three reference segmenters and two naïve baselines. Results show that using subphrases produced by our segmenter results in a 5% to 8% increase in classification accuracy over the references and baselines.

This chapter is based on work presented in (Rodríguez-López and Volk 2015c).

6.1 Introduction

In this chapter we tackle the problem of automatically segmenting melodic phrases into subphrases based on template cues. More specifically, we focus on modelling segmentation driven by the recognition of previously heard melodic figures.

Main Concepts. We use the term *template* to refer to mentally-encoded knowledge about both segments and segmentation, acquired through exposure to music. *Template cues* hence refer to the recognition of templates during a segmentation process, and *template based segmentation* to machine segmentation approaches where it is assumed that template recognition is determinative of the outcome of a segmentation process.

Modelling Tasks. Template recognition involves matching auditory stimuli (in our case figures in short-term memory) to templates (in our case figures in long-term memory). Thus, template based segmentation requires (I) the formalisation and construction of a model of long-term memory (LTM) containing exemplar figures, (II) the formalisation of a matching strategy that identifies exemplar figures within a phrase (which we assume delimits short-term memory), and (III) the formalisation of how identified exemplars are used to segment phrases.

Scope. We focus on modelling template based segmentation as performed by listeners well acquainted with a given melodic tradition. Throughout this chapter we refer to them as *enculturated* listeners.[40] (The motivations to do so will be covered in the following two sections.)

Contributions. The research presented in this chapter has two main contributions. First, to tackle problem I, we propose to model LTM probabilistically, and introduce an optimisation approach to learn melodic figure distributions from melodic corpora. Existing machine segmenters that model LTM probabilistically often use learning approaches that store information indiscriminately into LTM, simulating listeners with perfect retention and recall. Our learner differs from these approaches in that it is able select what to store in LTM. (We introduce and test two selection techniques.) We argue that this way of learning, hereafter *selective acquisition* learning, is more suitable for modelling the LTM of human listeners, especially that of enculturated listeners.

[40] It is worth stressing that we take the meaning of enculturation in a very restricted sense. Musical enculturation is a broad and multifaceted process. It includes the development of specialised perceptual processing mechanisms for musical structures (e.g. segment, metric, tonal), the shaping of aesthetic and expressive norms, and linkage between music and social situations. In this chapter we take enculturation only in respect to the first of the alluded facets.

Our second contribution tackles problems II and III. We introduce an approach to template based segmentation. Our approach frames segmentation as a noisy message decoding problem.It proceeds by first computing all possible segmentations of a phrase, and then searching for the one that gives the most likely interpretation of the message.

We use our segmentation and learning approaches to design two 'enculturated' segmenters. That is, two template based segmenters, each using a probabilistic LTM model trained with a different selection technique of our learner. We use the two enculturated segmenters to segment the phrases of 3000 folk melodies, each belonging to one of three different cultural traditions. We test the usefulness of the segmentations produced using our enculturated segmenters in a melody classification experiment. In the experiment the automatically determined subphrases are used as features to estimate the cultural origin of each segmented melody. To have a comparison point we also segment the melodies using three (non-enculturated) segmenters and two naïve baseline segmenters. Our results show that using the subphrases produced by enculturated segmenters substantially improves classification accuracy when compared to those produced by the reference segmenters.

Chapter Structure. The remainder of this chapter is organised as follows. In §6.2 we discuss template cues in more detail. In §6.3 we motivate and formally introduce our template based segmentation approach. In §6.4 we motivate and formally introduces our selective acquisition learner. In §6.5 we describe the classification experiment used to test segmenters, presents results, and discuss them. Finally, in §6.6 we summarise our conclusions and outline possibilities of future work.

6.2 Discussion on Template Cues

Templates can be abstract or concrete. Abstract templates can be seen as *prototypes* (Roach 1975) or *schemata* (Rumelhart 1980). In the context of segmentation, prototypes correspond to mental structures summarising the most stereotypical characteristics of segments, segment organisation, or of the process of segmentation itself. For instance, a listener well acquainted with music from the Classical Period might have distilled norms of tonal motion that allow her/him to characterise (and hence recognise) antecedent-consequent phrase structures common to that era. Prototypes are abstract in that, despite being constructed from reference segments or segment structures, they do not correspond to any particular reference instance. Concrete templates, conversely, can be seen as *exemplars* (Medin and Schaffer 1978). Exemplars correspond to actual instances of segments or segment structures that listeners

have heard (and retained in memory) at one point or another during their lives. For instance, a listener well acquainted with the music of a given composer might have a mental library of figures or phrases that characterise the composer's works.

6.3 Approach to Template Based Segmentation

In this section we both motivate and formally introduce our approach to template based segmentation. We also review related work to situate the approach in context.

6.3.1 Motivation: Phrase Segmentation Driven by Exemplar Recognition

Melodic phrases are likely to have more than one cognitively plausible segmentation. For instance, Lerdahl and Jackendoff (1983, p. 63) discuss how the phrase in Figure 6.1 can have different interpretations depending on whether attention is focused on figure repetitions or on note durations. Ahlbäck (2004, p. 252) goes further and suggests that (through repeated exposure) it might be possible for some listeners to experience and conceive both interpretations, and be able to switch between them at will.

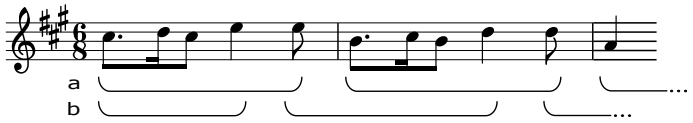

Figure 6.1: Soprano voice of a theme's phrase by Mozart (Piano Sonata K 331 in A major). The phrase "... *is supplied with two possible groupings. We favor grouping a, but grouping b has not been without its advocates; see Meyer 1973.*" (Lerdahl and Jackendoff 1983, p. 63).

In this chapter we posit that, at least during first time exposure, listeners might use template cues to disambiguate the different possible segmentations of a phrase. Figure 6.2 presents a simplified scenario depicting segmentation driven by the recognition of two melodic figure exemplars. This depiction embodies what is commonly referred to as 'top-down' processing (Narmour 1990, 1992), in that segment boundary perception is a by-product of the figure recognition process.

An important aspect of the depicted approach to segmentation is that we assume the recognised exemplars lead to the 'best possible interpretation' of the phrase. Since music, unlike vision or language, does not refer to objects in the world, defining what it means to 'interpret' or 'understand' music is certainly non-trivial – see discussions in (Bartel 2006, 2007). However, for the purpose of motivating our approach, we use an analogy to language understanding. Let us take a native Spanish speaker trying to make sense of the Italian utterance "*l'albero è secco*" [which translates to "the tree

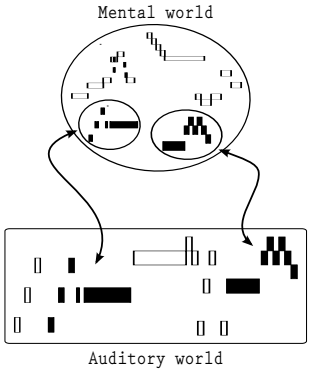

Figure 6.2: Segmentation of a melodic phrase (bottom) based on the recognition of two figures imprinted in a listener's long-term memory through exposure to music (top). Melodic events are depicted visually using a piano roll like representation. Rectangle length symbolises duration, the vertical dimension symbolises pitch, the horizontal dimension time.

is dry" in English]. Lets also assume that the (s)he does not speak Italian. In this scenario we can expect that the speaker, in an attempt to interpret the utterance, might segment it by matching closely sounding words from her/his native language, for instance *árbol*∼*albero* and *seco*∼*secco*. It follows from this example that the more matches to closely sounding Spanish words are found in the utterance, the better the interpretation the speaker will get. We argue that this analogy might hold true for how listeners interpret music, particularly for listeners well acquainted with a given melodic tradition or style. For these listeners phrase segmentation could be seen as a goal-directed search process, where the goal is to find the segmentation containing the most informative exemplar matches. Informative matches should allow the listener to establish connections between the piece being heard and previously heard music. And enable categorisation: the melody being listened to '... belongs to this or that tradition', or '... is in the style of so and so composer', or '... contains a musical quotation from so and so piece'.

6.3.2 Formalisation of Our Approach

Our approach to segmentation proceeds by processing a melody a phrase at a time. For each phrase it computes a number of possible segmentations. For each segmentation it searches for exact matches to exemplar figures, and scores them. Figure scores are used to give an overall score to the segmentation. The segmentation with the highest score is then given as output.

In the following subsections we describe each step in turn. We start by describing the input (choice of melody representation and phrase segmentation markings). Then we describe how the space of possible segmentations is computed. Lastly, we describe how figures matching exemplars are scored, and how these scores are used to rank segmentation candidates.

Melody Representation. Our segmenter takes as input melodies represented as a pitch interval sequences, constrained to a range of two octaves.[41] Formally, we take $p = p_1 \dots p_N$ to be a sequence of pitch intervals, where each interval $p_i \in \mathcal{A} = \{-12, \dots, 0, \dots, +12\}$. In $\mathcal{A}$ each numerical value encodes the distance in semitones between the pitches of contiguous notes, and the $\pm$ symbol encodes its orientation (ascending, descending).

Phrase Segmented Melodies. We assume input melodies are annotated with phrase boundaries, so that our segmenter can process melodies on a phrase by phrase basis, finding for each an optimal segmentation. We choose to process phrases based on cognitive constraints, as exhaustively evaluating several segmentations for a whole melody would break known limitations of human memory.

Computing Possible Segmentations. Ideally, our segmenter should evaluate all possible segmentations of a phrase. However, processing time is exponential on the number of notes in the phrase, so in practice evaluating all segmentations is unfeasible. Thus, we use the algorithm proposed by Opdyke (2010) to efficiently compute a constrained space of possible segmentations. The algorithm takes as input the minimum and maximum length of subphrases, as well as the minimum and maximum number of subphrases. We limit subphrases to be sequences of 1-5 intervals in length (which is a sensitive choice considering the discussion in §2.4.2). We also limit phrases to be composed of at most 6 subphrases (by doing so we are able to cope with phrases of a maximum length of 30 intervals).

[41] As through most of this thesis, input melodies are assumed to be in symbolic format. For our experiments the symbolic encodings processed correspond to computer readable representations of scores transcribed by experts (see §6.5.1 for more details). Symbolically encoded melodies can be represented in a variety of ways, e.g. chromatic pitch, step-leap pitch intevals, inter onset intervals, and so on. In statistical learning this multi-dimensional attribute representation of melodic events can be dealt with using *multiple viewpoint systems* (Pearce 2005; Conklin and Witten 1995). However, using multiple viewpoints comes at expense of a considerable increase in the complexity of the statistical model architecture, resulting in an increase in processing time and space requirements, as well as lower interpretability of the model. In this chapter we favour using a single melodic representation to simplify the evaluation our of segmenters, which is important considering that we evaluate our segmenters indirectly, by means of a classification experiment (we discuss this further in §6.5). Moreover, we focus on melodic pitch information motivated by findings in (folk) melody recall experiments – see (Hébert and Peretz 1997) – which suggest pitch information leaves stronger imprints in LTM than rhythmic information.

Scoring Matches to Exemplars

As motivated in the previous section, exemplar matches should be as 'informative' as possible. In this chapter we take this to mean that they should be 'characteristic' of a melodic culture, enabling the listener to perfom cognitive tasks such as melody categorisation. One way to measure how characteristic figures are is by searching for 'common' figures in a corpus representative of a melodic culture. However, common figures are mainly of short duration, and normally less specific and informative than figures of larger duration (Wołkowicz et al. 2008). There is hence a trade-off between how common a figure is and how specific to a given tradition it can be.[42] Thus, we need a way to automatically determine how long do the figures we are after need to be, so that we search for the longest possible common figures instead of only the most common ones. One way to do so is by attempting to determine if a given figure is somehow 'complete' on its own, or if its part of a larger figure. Our search then would be for figures that are common, yet large enough so as to be perceptually complete. According to melodic expectation theory (Meyer 1957; Pearce and Wiggins 2006b), the perceptual completeness of a melodic figure is inversely proportional to the degree by which it stimulates expectation. In other words, melodic figures for which it is hard to predict what comes next are perceived as more complete than those for which is easy to predict what comes next.

Using information theory we can attempt to jointly quantify the commonness and completeness of a figure. If from within a phrase of length T we take a figure $w = p_i \dots p_j$, with $i, j \in [1 : T]$, we can compute its conditional entropy h as

$$h(x|w) = P(w) \sum_{x \in \mathcal{A}} P(x|w) log\left(P(x|w)\right) \tag{6.1}$$

where x is used to symbolise melodic events that can follow w, and P denotes probability. In Eq. 6.1 the first term $P(\cdot)$ will be high for common figures in a corpus, and the second term $\sum P(\cdot) log\left(P(\cdot)\right)$ will be high if it is hard to predict what comes after w. Hence, h will be high for figures that are common and complete in an information theoretic sense.

The values of probabilities $P(\cdot)$ can be estimated from the counts of w and the concatenation wx that our learner has extracted from a given melodic corpus: $P(w) \sim N(w)/N_T$ and $P(x|w) \sim N(wx)/N(w)$, where $N(\cdot)$ denotes counts, and N_T denotes

[42] In natural language this is also a commonly found problem, 'content' or informative words (e.g. nouns) tend to be of greater length that 'non-content' words (e.g. determinants).

the total number of counts for figures of length equal to w in the corpus.

Selecting a Segmentation. If we have a space of possible segmentations $\mathcal{S}$, the average $\overline{h}$ of a candidate segmentation $s = w_1, \ldots, w_m$ is

$$\overline{h}(s) = \frac{h(w_1) + \cdots + h(w_m)}{m} \tag{6.2}$$

where $h(w)$ is computed using Eq. 6.1. We can then select the best segmentation by computing

$$\overline{h}^* = \operatorname*{argmax}_{s \in \mathcal{S}} \overline{h}(s) \tag{6.3}$$

6.4 Approach to Selective Acquisition Learning

In this section we both motivate and formally introduce our approach to selective acquisition learning. We also review related work to situate the approach in context.

6.4.1 Motivation for Selective Learning of Exemplars

For our purposes LTM can be modelled as a relational database of melodic figures, which gives some account of the salience or importance of each figure in the characterisation of a particular melodic tradition or style. This model of LTM could in principle be constructed from musicological resources (e.g. catalogues or dictionaries of melodic figures). However, to the best of our knowledge existing melodic catalogues are scarce, expensive, rarely found in machine readable formats, and most often document whole themes rather than figures. Thus, it is necessary to construct the LTM model automatically from melodic corpora by simulating human learning of melodic figures.

In existing machine segmenters that model LTM learning is approximated rather crudely, normally as n-gram counting. That is, all possible sequences of notes from a melody are identified, counted, and stored in a data structure. This method simulates a learning with perfect retention and recall – a listener with eidetic or photographic memory.

The n-gram approach to model learning has one big limitation. Learning is passive, yet goal oriented. This means that acquisition of exemplars seems to be selective. After listening to a melody, not all figures will be stored in LTM. Most will be activated

to aid cognitive processes during listening, but will not be stored in LTM. We argue that this is specially significant to model the LTM of an enculturated listener. We assume only those figures leading to the best interpretation of a phrase are retained in memory. Thus, the LTM of an enculturated listener initially is eager to learn just melodic figures, but after been exposed to certain number of melodies starts to pick up regularities in these melodies [in our case that would mean figures representative of the style]. This process is retroactive in that it influences segmentation. We hence introduce an approach to figure learning that embeds a statistical learning approach into an optimisation framework. We propose an unsupervised statistical learning approach in which figure learning and phrase segmentation are performed jointly.

6.4.2 Formalisation of the Approach

In this chapter the goal of selective acquisition learning is to construct an enculturated LTM model. We model enculturation as a refinement process. That is, our learner takes two inputs: (1) a LTM model, which is simply a collection of melodic figures acquired during prior listening experience, and (2) a corpus of melodies of a given culture to which the learner is to be exposed. The output is a LTM model in which, ideally, only melodic figures characteristic of the culture to which the learner has been exposed are preserved. Our learning approach is summarised as pseudo code in Algorithm 1.

Input: LTM model, Phrase-segmented Melodic Corpus
Output: LTM model

while *termination condition not met* **do**
    read melody from corpus;
    for *each phrase in melody* **do**
        Compute possible segmentations;
        Select the optimal segmentation;
        Store suphrases in LTM;
    Check termination condition;

Algorithm 1: Selective Acquisition Learning

As shown in Algorithm 1, our learner 'listens' to each melody one phrase at a time, and decides which figures to store in LTM by evaluating different segmentations. That is, the learner stores in LTM only the figures that allow it to segment the phrase in an optimal way. This process is continued until the learner has acquired the melodic vocabulary that allows it to perform optimal segmentations. In the following sections we describe each part of the approach in more detail.

Long-Term Memory Model. We model LTM probabilistically using a Markov modelling strategy. Essentially this boils down to constructing a data structure to hold the number of times melodic figures up to 5 intervals appear in a corpus, and then use those counts to estimate probabilities (as described in §6.4.2).[43]

Select the optimal segmentation

Below we present two techniques to select an optimal segmentation. One in which the learner selects subphrases that give it the 'clearest' possible understanding of a phrase, and another in which the learner uses subphrases it 'knows well' to increase its vocabulary.

Monitoring LTM. Using conditional entropy we can monitor the state of our LTM before and after a new melodic figure is listened to. So, first, the total entropy for figures w of the same size is

$$H^o = -\sum_{w \in \mathcal{A}^*} P(w) \sum_{x \in \mathcal{A}} P(x|w) log P(x|w) \tag{6.4}$$

where we use $\mathcal{A}^*$ to denote the space of all figures of size o with attribute space $\mathcal{A}$. In our LTM $o = \{1, \ldots, 5\}$ and hence its total entropy is

$$H = H^1 + \cdots + H^5 \tag{6.5}$$

and then we can define ΔH as

$$\Delta H = H_{\text{after listening to } w} - H_{\text{before listening to } w} \tag{6.6}$$

which allows us to monitor the evolution of our LTM.

Selection Technique 1. We have now the necessary information to formulate our first selection technique. Since common and complete figures are expected to have high entropy, a 'good' phrase segmentation among a group of possible segmentations is that

[43] The input LTM model can also be computed by sampling from known parametric distributions, e.g. in (Abdallah and Plumbley 2009) the LTM model is constructed sampling from a Dirichlet distribution. However, by using corpus statistics we can assess how different (and perhaps more suitable) are the segmentations produced by one of the learners in respect to the others when exposed to the same melodies, which is a better way to try to prove or disprove our hypothesis.

segmentation with the highest average ΔH. That is, if we have a space of possible segmentations $\mathcal{S}$, the average ΔH of a candidate segmentation $s = w_1, \ldots, w_m$ is

$$\phi(s) = \frac{\Delta H(w_1) + \cdots + \Delta H(w_m)}{m} \tag{6.7}$$

and hence our first selection technique is

$$s^* = \underset{s \in \mathcal{S}}{\operatorname{argmax}}\, \phi(s) \tag{6.8}$$

Where s^* denotes the segmentation with maximal score. Note that, to ensure convergence, the learner stores in LTM only the subphrases in s^* for which ΔH is positive.

One problem with our first technique is that it makes our learner very conservative. The melodic figures stored are characteristic of the corpus as a whole. Hence, the technique operates under the assumption that the corpus is stylistically homogeneous. For most cultural traditions the assumption of complete stylistic homogeneity is too strong (it is likely that certain figures are important but only characteristic of subsets of the corpus).

Selection Technique 2. Our second technique aims to relax the assumption of homogeneity and stimulate the learner to expand its vocabulary. More importantly, it aims to reveal segmentations in which one or more subphrases are common and complete, and others are representative of the melody, yet relatively rare in the corpus. For a figure w the latter idea can be quantified as

$$\rho(w) = -P_{melody}(w) * log\left(P_{corpus}(w)\right) \tag{6.9}$$

with $P_{melody}(w) \sim M(w)/M_T$ and $P_{corpus}(w) \sim N(w)/N_T$, where M denotes counts of w in the melody, M_T is used to indicate the total number of counts of figures of size equal to w in the melody/corpus, and N denotes counts of w in the corpus.

For a complete segmentation we take the average of ρ

$$\overline{\rho}(s) = \frac{\rho(w_1) + \cdots + \rho(w_m)}{m} \tag{6.10}$$

Finally, we combine the $\overline{\rho}$ and ϕ using a geometric mean:[44]

$$\lambda(s) = \sqrt{\phi(s) \cdot \overline{\rho}(s)} \tag{6.11}$$

and compute our second technique as

$$s^* = \underset{s \in \mathcal{S}}{\operatorname{argmax}}\, \lambda(s) \tag{6.12}$$

Where s^* denotes the segmentation with maximal score. Our leaner stores all subphrases of s^* in LTM.

Termination Condition

We keep track of the scores of s^* when processing the corpus, expecting that, as the learner reaches convergence, the score difference between subsequent instances s^* gets smaller and smaller. We hence assume convergence has been reached if $\Delta s^* < \epsilon$.

Since Eq. 6.12 encourages learning new vocabulary, convergence is slow and not guaranteed. Thus, in addition to $\Delta s^* < \epsilon$, we also set a maximum number of learning iterations as a second termination condition.

6.5 Evaluation

At present, freely available corpora annotated with subphrase boundaries do not exist. This implies we are unable to evaluate our segmenters in a traditional scenario (i.e. by comparing automatic segmentations to human-annotated segmentations). Hence, we opt for a 'task based' evaluation scenario: test the output of our segmenters in a melody classification experiment.

The classification task consists in predicting the cultural origin of each melody in a dataset of melodies, using subphrases as classification features. In this scenario 'good' segmentations should facilitate classification and thus result in high classification performance.

In the following subsections we describe the melodic corpora used for our classification experiment, the compared segmenters, the classifiers employed, and finally we list

[44] Since $\phi(s)$ can in principle be negative, to compute λ we consider negative $\Delta H(w)$ values to be zero when computing $\phi(s)$ to avoid the possibility of negativity.

evaluation metrics and present results.

Collection Name Abbreviation	Subset Name	Cultural Origin of Sample	Encoding	Number of Melodies	Average Melody Size in Notes	Number of Phrases	Average Phrase Size in Notes
MTC	FS	Dutch	**kern	4120	52.3 (22.5)	19935	9.1 (2.5)
EFSC	CHINA	Chinese	**kern	2201	62.8 (41.2)	11046	12.5 (4.7)
OHFT	-	Hungarian	EsAC	2323	38.6 (12.0)	9308	9.6 (3.2)

Table 6.1: Melodic Corpora. Numbers in parenthesis correspond to standard deviation.

6.5.1 Phrase Annotated Melodic Corpora

The melodic corpora used in our experiments is summarised in Table 6.1. The *Meertens Tune Collection* (MTC), FS subset, is a collection of vocal Dutch folk songs. The *Essen Folk Song Collection* (EFSC), CHINA subset, is a collection of vocal folk songs from Eurasia. The *Old Hungarian Folksong Types* collection (OHFT) is a collection of vocal folk songs from Hungary. All corpora is briefly described in §3.5.

All corpora used for our experiments have been annotated with phrase boundaries by expert ethnomusicologists.[45] We cleaned the collections by removing all melodies with overly short and overly long phrases. We considered a phrase to be overly short if it contains two notes or less (one pitch interval in our input representation). We considered a phrase to be overly long if it is longer than 30 notes in length.

6.5.2 Enculturated Segmenters

We evaluate three enculturated segmenters: NS, ST1, ST2. The NS segmenter uses a LTM model trained with non-selective acquisition, using the PPM-C algorithm (Pearce and Wiggins 2004). The ST1 segmenter uses a LTM model trained with the selective acquisition technique 1, Eq. 6.8. The ST2 segmenter uses a LTM model trained with the selective acquisition technique 2, Eq. 6.12. A sample of 1000 melodies from each collection is used to train the LTM models. The parametric settings for each enculturated segmenter are specified in Table 6.2.

[45] In the case of the EFSC-CHINA the origin of the phrase markings is uncertain. However, it is often assumed it corresponds to notated breath marks and/or to the phrase boundaries of lyrics. In the case of the MTC-FS phrase boundary markings where produced by two experts (which agreed on a single segmentation). The annotation process is detailed in (van Kranenburg et al. 2014). In the case of the OHFT the phrase boundary marking process is detailed in (Kodály and Vargyas 1982; Járdányi 1965). It also appears to be driven mainly by breath marks and/or to the phrase boundaries of lyrics.

6.5.3 Reference Segmenters and Baselines

We compared the performance of the enculturated segmenters to two local boundary detection segmenters (LBDM and PAT), and two naïve baseline segmenters (FIXLEN and RAND). The LBDM and PAT segmenters were selected for comparison because they have been used for subphrase level segmentation in the past (Orio and Neve 2005; Cambouropoulos 2006). The LBDM segmenter (Cambouropoulos 2001) computes subphrase boundaries by detecting large pitch intervals and inter-onset-intervals. Intervals sizes are given a score by comparing them to immediately surrounding intervals (the larger the difference the higher the score). High scoring intervals are taken as subphrase ends. The PAT segmenter (Cambouropoulos 2006) computes subphrase boundaries by detecting and scoring repetitions of pitch interval sequences within each phrase. The starting points of high scoring repetitions are taken as subphrase starts. The FIXLEN baseline segments a phrase into subphrases of constant size. The RAND baseline segments a phrase into subphrases of randomly chosen sizes. The parametric settings for each of the reference and baseline segmenters are specified in Table 6.2.

6.5.4 Features and Classifiers

As mentioned above, in our experiment we are interested in evaluating the effectiveness of subphrases as classification features. To use subphrases in the most transparent way, we represent melodies as a 'bag-of-subphrases'. That is, we use a vector space model representation,[46] where each vector element is weighted using the common term frequency - inverse document frequency ($tf * idf$) heuristic (Manning et al. 2008). Then, to classify melodies by cultural origin, we use two simple and well known classifiers: *k-means* and *k nearest neighbours* (kNN).

6.5.5 Test Set, Performance Measures, and Results

We constructed a dataset of 3000 melodies by randomly sampling 1000 melodies from each corpus. (All melodies used to train the enculturated segmenters were excluded from the sample.) For each of the 3000 melodies, the classifiers are required to predict whether the melody is of Hungarian, Chinese, or Dutch origin.

[46] In a vector space model, melodies are represented as a vector of size $|V|$, where $|V|$ is the number of unique figures occurring in the corpus. If a figure occurs in the melody, its value in the vector is equal to the number of times it appears in the melody. The frequency of occurrence of each figure is then used as a feature for classification.

Segmenter	Parameter Setting	Segmentation Results (for the best parametric setting)								
		Mean Number of Subphrases per Phrase			Mean Number of Subphrases per Melody			Total Number of Unique Subphrases per Corpus		
		C	H	D	C	H	D	C	H	D
NS	LTM training: PPM-C, with exclusion, 1000 melodies of each culture.	5.0	4.8	4.8	27.7	16.7	23.4	1300	1091	1197
St1	LTM training: convergence 10E-8 or 8000 phrases, 1000 melodies of each culture.	3.8	3.7	3.7	21.7	13.7	19.6	3437	2562	2204
St2	LTM training: convergence 10E-8 or 8000 phrases, 1000 melodies of each culture.	3.6	3.5	3.5	23.2	15.4	21.3	3566	2743	2311
LBDM	detection threshold $\{0.2, \mathbf{0.4}, 0.6\}$.	3.0	2.9	2.9	15.5	10.4	15.4	4497	3841	2999
PAT	detection threshold $\{0.2, \mathbf{0.4}, 0.6\}$.	3.6	3.4	3.4	19.3	11.4	16.7	3603	3139	2810
FIXLEN	constant size $CS = 3$ intervals.	4.0	3.8	3.8	22.0	13.5	18.9	1371	1179	1474
RAND	constant size $RS \in [2-4]$ intervals.	4.1	3.9	3.9	22.2	13.5	19.2	2827	2413	2551

Table 6.2: Parameter settings and segmentation results. C - Chinese, H - Hungarian, D - Dutch. Text in bold indicates best performing parametric settings. Experimental settings for LBDM and PAT are taken from suggested parametric configurations by Cambouropoulos (2001, 2006).

Validation technique. We used 10-fold cross validation to iteratively separate the melodic dataset into training and test sets.

Evaluation measures. Given a N_{total} of melodies per fold to be classified, we use tp to indicate the number of true positives, fp the false positives, and fn the false negatives. With these statistics we measure measure classification performance using accuracy $A = \frac{N_{correct}}{N_{total}}$, precision $P = \frac{tp}{tp+fp}$ and recall $R = \frac{tp}{tp+fn}$. These measures are then averaged over all folds.

Statistical testing. We used an an ANOVA test ($\alpha = 0.01$) with Bonferroni correction to test the statistical significance of the differences in accuracy for each segmenter.

Setting and optimising classifier parameters. The training sets were used to optimise the permutation labels of the k-means classifier and select the optimal number of nearest neighbours for the kNN classifier. The optimal number of nearest neighbours (selected from $k \in [1, 15]$) was set by optimizing cross-validated accuracy on the training data.

The results of our experiment are presented in Table 6.3. We discuss our results in the following section.

Segmenter	k-means (k=3)			kNN (k optimised)		
	$\overline{R}$	$\overline{P}$	$\overline{A}$	$\overline{R}$	$\overline{P}$	$\overline{A}$
NS	0.94	0.93	0.71	0.93	0.87	0.83
St1	0.90	0.95	**0.74**	0.93	0.94	0.87*
St2	0.92	0.93	0.71	0.92	0.96	**0.88**
Lbdm	0.47	0.50	0.47	0.75	0.84	0.76
Pat	0.74	0.76	0.58	0.83	0.87	0.79
FixLen	0.88	0.89	0.67	0.86	0.90	0.83
Rand	0.84	0.84	0.63	0.88	0.85	0.78

Table 6.3: Clasification results: average recall ($\overline{R}$), precision ($\overline{P}$), and accuracy ($\overline{A}$) (average is computed over the 10-folds used for crossvalidation). Text in bold highlights the highest performances. Asterisks indicate performances that are not significantly different from the highest performances.

6.5.6 Discussion

Note: Table 6.3 shows that all scores, even the ones by random baselines are relatively high. One could then argue that there is little 'room to improve'. It is then important to stress, once more, that the evaluation is not focused on assessing classification performance per se, but rather to use classification performance to compare

segmentations. Our focus is hence on discussing relative differences.

Selective vs. Non-Selective Learning Segmenters

Table 6.2 shows that the NS segmenter produces relatively short segments, resulting in an average of ∼4.9 subphrases per phrase, and an average of ∼1196 unique subphrases over all three corpora. Conversely, the ST1-2 segmenters produce larger segments, resulting in an average of ∼3.6 subphrases per phrase, and an average of ∼2767 unique subphrases over all three corpora. Using the k-means classifier with subphrases computed using ST1 we obtain a (statistically significant) 3% $\overline{A}$ improvement over the NS segmenter, which seems to be driven by a 2% improvement in $\overline{P}$. Using the k-NN classifier with subphrases computed using both ST1 and ST2 we obtain (statistically significant) 3-4% $\overline{A}$ improvements over the NS segmenter, which are again in pair with 7-9% increases in $\overline{P}$. These results show the larger segments produced by the ST1-2 segmenters allow better discrimination between melodies of different cultural origin, suggesting that selective learning leads to better models of prior listening experience than non-selective learning.

Selective Learning Segmenters vs. Local Segmenters

Table 6.2 shows that local segmenters prefer larger segments than the ST1-2 segmenters. Also, the local segmenters produce an average of ∼3481 unique subphrases over all three corpora, which is 741 subphrases larger than the average of unique subphrases produces by the ST1-2 segmenters. Table 6.3 shows that $\overline{A}$ results using the segments produced by ST1-2 are >8% better than $\overline{A}$ results using the segments produced by LBDM and PAT. The $\overline{A}$ performance improvements are in line with relatively large improvements in both $\overline{P}$ and $\overline{R}$. These results show that the larger segments produced by the local segmenters leads to an increase in unique subphrases, and that these unique subphrases are not discriminative of cultural origin. The relatively large improvements in $\overline{A}$ of the ST1-2 segmenters over the local segmenters supports the hypothesis that enculturated listening might be of importance for the segmentation of melodic phrases.

Selective Learning Segmenters vs. Baselines

Table 6.2 shows the baseline segmenters produce relatively short segments (of 2 or 3 intervals), resulting in an average of ∼3.9 subphrases per phrase, and an average of ∼1969 unique subphrases over all three corpora. When using the k-means classifier we

can observe significant and relatively large differences ($> 5\%$) between the $\overline{A}$ obtained using ST1-2 and those obtained using the baseline segmenters. These results show the larger segments produced by the ST1-2 segmenters allow better discrimination between melodies of different cultural origin than the shorter segments produced by the baseline segmenters, indicating once more the ST1-2 segmenters might be capturing important aspects of subphrase structure.

Scepticism

Any conclusions from our results are limited to classification schemes using 'bag-of-subphrases' representations of melodies. This representation limits the similarity assesment between any two subphrases to exact matches, which is most likely introducing an unwanted bias on the evaluation. To draw more definitive conclusions our experiment needs to be complemented with other use case evaluations.

6.6 Conclusions

In this chapter we introduce techniques for selective acquisition learning in the context of melodic segmentation, specifically the segmentation of melodic phrases into subphrases. Our aim is to show that enculturated listening is important for the segmentation of melodic phrases, and that selective rather than indiscriminative acquisition techniques are better to model the long term memory of enculturated listeners. We present two selective acquisition techniques: one in which an artificial learner selects the subphrases that give it the 'clearest' possible understanding of a phrase, and another in which the learner attempts to use subphrases it 'knows well' to expand its melodic vocabulary.

We test the usefulness of the subphrases produced by our template-based segmenter in a classification experiment. In the experiment automatically determined subphrases are used as features to estimate the cultural origin melodies. We compare to three reference segmenters and two naïve baselines. Results show that using subphrases produced by our segmenter results in a 5% to 8% increase in classification accuracy over the references and baselines. These results suggest that learning melodic figures is a goal-driven process in which segmentation and figure exemplar acquisition act conjointly.

Future Work

Extensions to Multiple Representations and Polyphony. In future work we plan to extend the current approach so that it can process multiple attribute representations of a melody, as well as polyphony. To meet these goals we plan an integration between our approach and the multipleviewpoint formalism of (Pearce 2005; Conklin and Witten 1995).

Applications in Ethnomusicology. Recent attempts at applying MIR tools to music from non-western traditions have not been particularly successful. As obseved by Sankalp et al. (2015): *"results suggest that (...) several melody-dominant music traditions of the world such as Flamenco and Indian art music need dedicated research efforts to devise specific approaches for computing melodic similarity"*. The adaptability of the approach presented in this chapter appears to be adequate to research these traditions. We hence plan to undertake experiments in this direction.

Testing Sensitivity to Cross-Learning. We also plan to conduct experiments to test the sensitivity of our selection techniques to cross-learning. That is, cases in which the learners have prior knowledge of one melodic tradition and are required to adapt their knowledge to the particularities of a different melodic tradition. Listening experiments with human subjects will be undertanken to check for similitude in the ability of cross learning.

Chapter 7

Multi-Cue Segmentation

In this chapter we tackle the problem of automatically segmenting complete melodies into phrases by combining multiple cues.

Chapter Contributions We formulate multiple cue segmentation as an optimisation problem, and introduce a cost function that penalises segmentations considering cues related to boundaries, segments, and the complete segmentation. Our segmenter differs from existing multi-cue segmenters in two respects. First, it is more complete, in that it has a wider coverage of cues. Second, it has a higher degree of autonomy, in that it has dedicated modules to estimate all needed cue information.

We evaluate our segmenter on the `FJ375` corpus. To have a comparison point we also evaluate GROUPER and two naïve baseline segmenters on the same corpus. GROUPER is currently at the state of the art of melodic segmentation. Results show that our segmenter achieves statistically significant $\overline{F1}$ improvements of ~8% in respect to GROUPER, and of over 20% in respect to the baselines.

This chapter is extends work presented in (Rodríguez-López et al. 2014a).

7.1 Introduction

In this chapter we tackle the problem of automatically segmenting complete melodies into phrases by combining multiple cues.

Motivation. The subconscious mind is thought to be capable of extensive parallel processing (Baars 1988, 1997; Minsky 1985; Ornstein 1986; Edelman 1987; Jackson 1987). We hence posit that melodic segment structure is the result of a parallel processing system, which not only detects segmentation cues, but also uses them to conceive and assess different segmentation hypotheses.

Modelling Tasks. Given a number of manually or automatically estimated cues, multi-cue segmenters have the task of evaluating the different segmentations the cues might suggest, to find those which are likely to be preferred by human listeners.

Contributions. We formulate multiple cue segmentation as an optimisation problem. We introduce a cost function that penalises segmentations considering cues related to boundaries, segments, and the complete segmentation. More specifically, our cost function discriminates between possible segmentations based on (a) estimated boundary confidence, (b) commonness of phrase contours, (c) quality of phrase segmentations, and (d) commonness of phrase length. (We ask the reader for patience on our use of rather vague terms such as 'commoness' or 'quality', they will be defined formally in the following sections.)

Our segmenter differs from existing multi-cue segmenters in two respects. First, it is more complete, in that it has a wider coverage of cues. Second, it has a higher degree of autonomy, in that it has dedicated modules to estimate all needed cue information.

We evaluated our segmenter on the `FJ375` corpus. To have a comparison point we also evaluate Grouper and two naïve baseline segmenters on the same corpus. Grouper is currently at the state of the art of melodic segmentation. Results show that our segmenter achieves statistically significant $\overline{F1}$ improvements of ~8% in respect to Grouper, and of over 20% in respect to the baselines. Our results also show clear benefits of using multiple sources of information for segmentation, which supports the view of human segmenting mechanisms as composed of parallel processing modules.

Chapter Structure. This chapter is organised as follows. In §7.2 we review previous work on multi-cue melodic segmentation. In §7.3 we describe our approach to multi-cue segmentation. In §7.4 we describe the experiments conducted to test our segmenter and discuss results. Finally, in §7.5 we present our conclusions and outline future work.

7.2 Related Work

In the 30-plus years of machine segmentation research we were able to find only six melody segmenters that combine cues. We summarise them in Table 7.1. In these segmenters cue combination is resolved either manually, via interaction with the user, or automatically, following a segmentation scoring approach. Below we review each segmenter with respect to these two approaches.

Segmenter Author(s)	Cues	Input	Output
Hamanaka et al. (2006)	{g,r,t}	{note diatonic pitch, quantised onset, offset; beats}	single parse of ... figures, phrases (arranged hierarchically)
Temperley (2001)	{g,a,t}	{quantised onset, offset; beats, bar, hypermetric; typical phrase length}	single phrase parse
Camilleri et al. (1990)	{g,r,a}	{note diatonic pitch quantised onset, offset, dynamics; beats, bars}	unranked list of... figure, phrase parses
Rowe (1992)	{g,a,c}	{note chromatic pitch, onset, offset, dynamics; beats}	single phrase parse
Baker (1989b,a)	{g,r,a,t}	{note chromatic pitch quantised onset, offset; beats, bars; template phrase forms; chord dictionary}	ranked list of... phrase parses

Table 7.1: Multi-cue melody segmenters. Cue class abbreviations: (g)ap, (r)epetition, (c)losure, (a)lignment, (t)emplate.[47] [48]

Cue Combination via User Interaction

Hamanaka et al.'s Atta (2006) and Rowe's Cypher (1992) fall in this category. Atta is a system for computer-assisted melody analysis based on Lerdahl and Jackendoff's

[46] To interpret the cue classes refer to Figure 2.3.

[46] When describing segmenter's outputs use the term 'parse' for brevity. However, its meaning varies for each segmenter. For the segmenters of Hamanaka et al. and Camilleri et al. the output is a data structure, containing boundary locations and hierarchy information. For the segmenters of Temperly, Rowe, and Baker, the output is not a data structure, but rather simply a list of boundaries, where each boundary is assumed to correspond to either the ending note or the starting note of a phrase.

GTTM (1983). It outputs both segmentation and time-span reduction analyses. For the segmentation analysis module, ATTA provides a visual interface so that users can adjust weights/scores to merge gap, alignment, and repetition cues. CYPHER, conversely, is a system for real-time human-machine improvisation. It is composed of two macro modules: one for listening (i.e. analysing musical input) and another for reacting (i.e. generating a musical response). The analysis module contains a phrase structure analyser, which uses gap, alignment, and closure cues to determine segmentations. Cue combination is tackled using a mixture of hard coded and automatically adaptable parameters. Hard coded parameters are set according to heuristic rules inspired from music theory, e.g. gaps occurring at beats are given more weight; dominant-tonic progressions mark the end of phrases. Automatically adaptable parameters are given an initial value by users, and then recurrently updated during improvisation by the system. To this end the system keeps a memory stack of previous segmentations, and updates parameters as a way to 'correct' segmentations (once more guided by heuristics), e.g. produced segments should have a length ranging between a user-defined maximum and minimum.

Critical Comment. ATTA requires manual tuning of over 25 parameters. Tuning time was estimated at ~10 mins per melody in Hamanaka's experiments. While the estimate is acceptable for case studies in CMMC, it is restrictive for most applications in MIR. What is more, ATTA's parameters are set at initialisation and not updated during processing. Thus the cue combination strategy is insensitive to changes in the relative importance of a given cue during the course of a melody. CYPHER is more autonomous. It has dedicated modules to estimate all metric and harmonic/tonal information required by its segmenter module. However, Rowe himself acknowledges than the modules dealing with key, chord, and beat estimation are limited, and perform optimally only for "well behaved" tonal input, hence making the system genre/style specific. Moreover, CYPHER's segmenter contains a considerable number of hard coded parameters (e.g. gap cue strength), which remain invariant through analysis.

Cue Combination via Segmentation Scoring

All other segmenters in Table 7.1 follow a segmentation scoring approach. That is, they define an initial set of candidate segmentations and score each of them to find one (or more) likely to be preferred by human listeners. For brevity, in this section we focus on the description of the two segmenters that we believe are more representative of this group, namely Bakers's GRAF (1989a) and Temperley's GROUPER (2001). For a comprehensive review of the other segmenters we refer to (Rodríguez-López and Volk 2012).

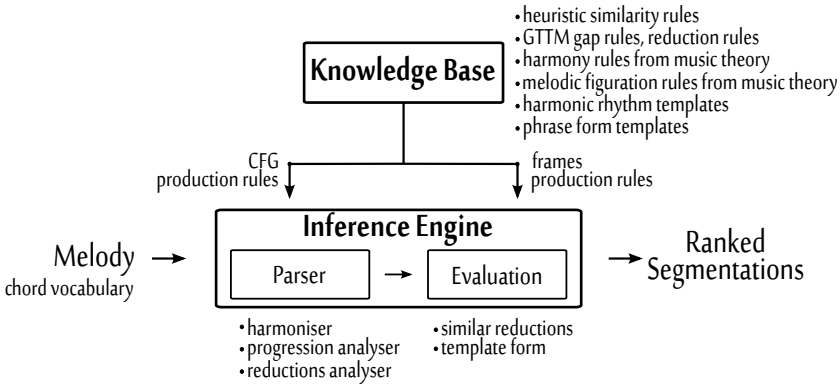

Figure 7.1: The GRAF segmenter.

Both GRAF and GROUPER were developed specifically for segmenting melodies into phrases, and combine gap, repetition, alignment, and template cues. GRAF is a rule-based 'expert system', augmented with a parsing module. It consists of a knowledge base and an inference engine – see Figure 7.1. The knowledge base contains a set of production rules, context-free grammars, and frames encoding relevant music knowledge. The inference engine consists of a parser and an evaluation module. The parser module produces candidate segmentations. In the candidate segmentations, phrase boundaries agree with GTTM gap rules, and also have tonal progressions which are 'well formed' from the point of view of Western music theory. Phrases are furthermore submitted to reduction analysis (using GTTM rules), the output of which is a tree where at each level only tonally important note pitches are kept. The evaluation module checks the level of similarity among reduced phrases (using heuristic rules), and the degree to which phrase structures match the template forms stored in the knowledge base. Segmentations are ranked by giving preference to those having similar reductions and more matches to template forms.

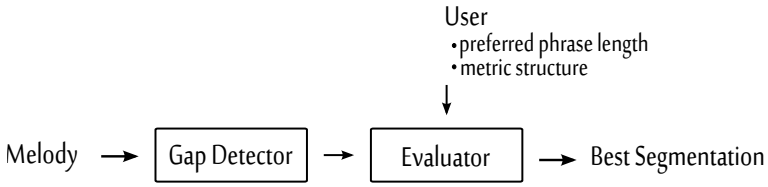

Figure 7.2: The GROUPER segmenter.

GROUPER is much simpler than GRAF – see Figure 7.2. It segments melodies considering gap, alignment, and template cues, but all related uniquely to temporal

information. GROUPER first defines the set of candidate segmentations as that whose boundaries are cued by temporal gaps. Then it uses dynamic programming to search for the segmentation which is most congruent with the melody's metric structure (alignment cue), and also where each segment shows the minimum deviation from and ideal segment length (template cue).

Critical Comment. GRAF requires a large amount of information to be supplied by the user. On the one hand, it requires the input melody to be encoded considering metric (bar, beat) information. On the other, it also requires genre/style specific information to be supplied to the knowledge base, such as phrase form templates, harmonic rhythm templates, figuration rules, harmony rules, and so on. While it might be feasible to locate and supply all this information for case studies in CMMC, this feature greatly restricts the applicability of GRAF to MIR. GROUPER, conversely, is more flexible. Its approach to cue combination is simpler, and thus easily adaptable and extendable. The amount of information required from the user is more manageable, and it has a mode of operation which runs in the absence of metric information. GROUPER has also been tested extensively. In fact, results of comparative studies put it at the state-of-the-art of melodic segmentation (Thom et al. 2002; Wiering et al. 2009; Pearce et al. 2010a,b). That said, GROUPER's main limitation is its over reliance on the gap detector to define candidate segmentations. It will fail to detect boundaries not cued by gaps, and hence it will never be able to locate the 'true' optimal segmentation. Moreover, from a cognitive point of view, using metric information above the beat level is questionable, as there is experimental evidence that suggests human listeners use segment structure for the conception of bars and hypermeter (Ahlbäck 2004).

7.3 Approach

We frame cue combination as a constraint satisfaction problem, and use an optimisation-based segmentation scoring approach to solve it. Our approach is conceptually similar to that used to develop GROUPER, yet its formalisation and implementation differ – in fact the formalisation is closer to that of work in audio segmentation research (Levy and Sandler 2008; Su et al. 2009; Sargent et al. 2011). Our approach presents two advantages over GROUPER. First, it is more complete, in that it has a wider coverage of cues (which rely both on temporal and pitch information). Second, it has a higher degree of autonomy, in that it has dedicated modules to estimate all needed cue information.

In the following we first give an overview of the approach, then formalise it, and lastly describe implementation specifics.

7.3.1 General Overview

Our approach is schematised in Figure 7.3. The main idea is to use automatically estimated boundaries to define the set of possible candidate segmentations. And to discriminate between candidate segmentations on the basis of information related to boundaries, phrases, and phrase structures. This information is estimated automatically, either by single cue segmenters (e.g. by using their 'boundary strength' estimations), or by statistical models inferred from a phrase annotated melodic corpus (e.g. estimated distributions for phrase length and phrase contour).

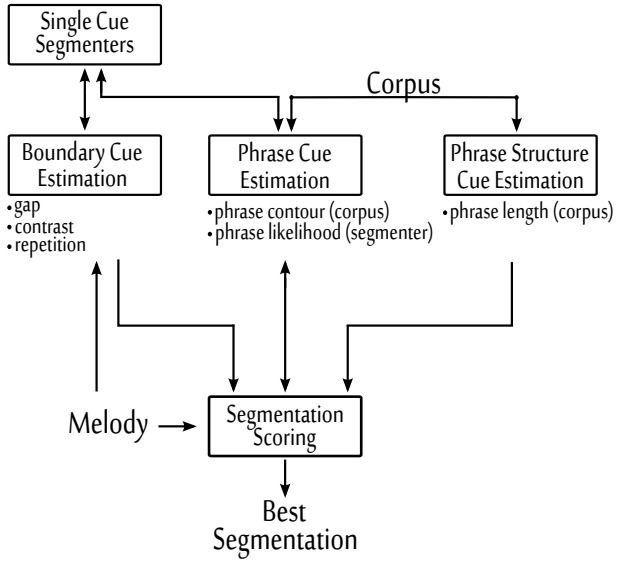

Figure 7.3: Schematic of our approach to multi-cue segmentation.

7.3.2 Formalisation

The segmentation scoring module has two inputs. The first is the melody, encoded as a sequence of notes $x = x_1 \dots x_n$ where each note is represented by a (pitch, onset) tuple. The second is a set of estimated boundaries $\hat{B} = \{(p_1, s_1), \dots, (p_m, s_m)\}$, where p is the boundary position index in x, and s is the machine estimated strength for the boundary. Moreover,

$$\hat{B} = \bigcup_{i=1}^{\{r,g,c\}} B_i \tag{7.1}$$

where B is the partial set of boundaries estimated using either g - gap, c - contrast, or r - repetition segmenters.

We construct a directed graph $G = (V, E)$ as a representation of the space of candidate segmentations – see Figure 7.4. In this graph, a vertex represents a note onset, and an edge between two vertices represents a candidate phrase. A path in G represents a candidate segmentation.

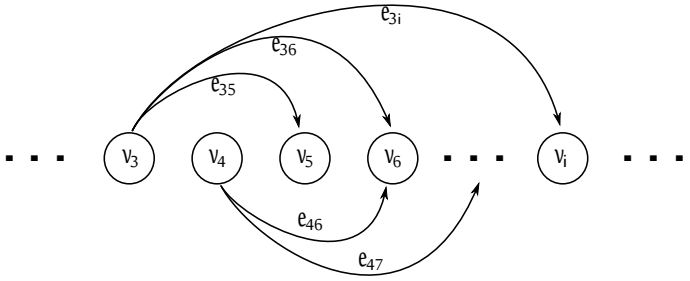

Figure 7.4: The directed graph G representing the space of segmentations of a melody.

We define the cost of a path $\mathcal{S}$ as the sum of the weights of its constituent edges and vertices,

$$C_{\mathcal{S}} = \sum_{v \in \mathcal{S}} w_v + \alpha \sum_{e \in \mathcal{S}} w_e, \tag{7.2}$$

where w_v and w_e are the weights of a vertex and an edge in B, and α is a user-defined mixing parameter. We take a path with high cost to represent an implausible segmentation. We thus formulate the problem of finding the most plausible segmentation as an optimisation problem

$$\mathcal{S}^* = \operatorname*{argmin}_{\mathcal{S}} C_{\mathcal{S}}, \tag{7.3}$$

where B^* is the list of boundary locations consituting the optimal solution. We frame this optimisation problem as a shortest path problem, and use Sargent et al. (2011) modified version of the Viterbi algorithm to solve it. The modified Viterbi algorithm employs a memory window, so that the number of possible paths to assess at any one time is reduced to those contained in the window. Reducing the number of paths has an obvious technical benefit, as it lowers the complexity of the algorithm. It also increases the plausibility of the algorithm as a model of cognition, as the added

locality and reduced number of hypothesised segmentation candidates are less likely to break known limits of human memory.

7.3.3 Defining Weights

In this section we define vertex (boundary) and edge (segment) weights required to compute path costs in Equation 7.2.

Vertex Weights

To determine the input $\hat{B}$ we use the repetition and constrast segmenters introduced in Chapters 4 and 5, respectively. We also use the LBDM gap segmenter of (Cambouropoulos 2001). These segmenters return a list of boundary positions, as well as estimations of boundary 'strength'. Estimated strengths are determinant of which locations are selected as boundaries, so it is often posited that their value is correlated to perceptual salience. To the best of our knowledge this has not been empirically validated. However, we could consider strength values as an indication of how 'confident' a segmenter is that a boundary occurs at a particular point in time, and hence create a cost based on it. Given that the aformentioned segmenters output strength values $s \in [0, 1]$, we compute boundary weights as

$$w_{v_{ij}} = \sum_{h \in S} C_{s_h} \,, \tag{7.4}$$

where $w_{v_{ij}}$ denotes the weight of a vertex i and vertex j (connected by an edge), $C_s = \frac{1}{(1+s)}$ is a cost based on strength s, and S is the set of all strength values predicted by different segmenters for positions i, j in the melody. (Vertices for which there are no boundary predictions are given a weight of zero.)

Edge Weights

We compute edge weights as

$$w_{e_{ij}} = C_l \, + \, C_c \, + \, C_p \,, \tag{7.5}$$

where $w_{e_{ij}}$ denotes the weight of a edge starting from vertex i and reaching vertex j, and $C_{l,c,p}$ are costs for phrase length, contour, and probability. Below we describe

each of these costs in turn.

Phrase Length Cost. We estimate a probability distribution for phrase length from phrase-annotated corpora. More specifically, we fit a Gaussian to a distribution of lengths obtained from 1000 phrases selected randomly from the `EFSC` and `MTC-INST` melodic databases – see §3.5 for a description. (The sample does not include phrases of any of the songs present in the `FJ375` test dataset.) The cost of phrase length is then computed as

$$C_l = -log(P(\lambda))\,, \tag{7.6}$$

where $P(\lambda)$ is the probability of the candidate phrase's length λ.

Cost C_l penalises 'atypical' phrase lengths. Imposing constraints on segment length can be motivated by a statistical fact: human annotated segments (particularly phrases and sections) have rather consistent lengths. For instance, Temperley observers that *"in the Ottman collection, the mean number of notes per phrase is 7.5; over 75% of phrases have from 6 to 10 notes, and less than 1% have fewer than 4 or more than 14"* (2001, p. 69). This observation can be generalised to phrases and sections lengths in both melodies and polyphonic music, as several studies have observed that segment length distributions are characterised by one or two clear peaks, and rather small interquartile ranges, e.g. (Spevak et al. 2002; Melucci et al. 2000; Bimbot et al. 2010a,b; McFee et al. 2015). The grounds for this consistency are to this day unclear, yet several speculative reasons can be given, e.g. human memory limitations (Miller 1956; Fraisse 1982), attentional span limitations (Clarke and Krumhansl 1990), or habituation due to motor/physiological constraints on music production (Godøy 2014).

Phrase Contour Cost. We estimate a probability distribution for phrase contour from the same 1000 sample of phrase used to derive length distributions. Phrase contours are computed using the polynomial contour extractor of the `Fantastic` toolbox (Müllensiefen 2009). A contour model with three nodes was selected. We fit a Gaussian Mixture Model (one Gaussian per node) to contour distributions obtained from the same sample of phrases taken to compute length distributions. The cost of phrase contour is then computed as

$$C_c = -log(P(\kappa))\,, \tag{7.7}$$

where $P(\kappa)$ is the probability of the candidate phrase's contour κ.

Cost C_c penalises 'atypical' phrase contours. Imposing constraints on phrase contour can be motivated by its alleged importance in music psychology, as contour is considered an integral aspect of melodic memory and recognition (White 1960; Dowling and Fujitani 1971; Edworthy 1985; Dowling et al. 1995; Cutietta and Booth 1996).

Phrase Probability Cost. Using our template segmenter (introduced in Chapter 6) we can estimate the most likely segmentation of a given phrase into figures.[49] The template segmenter not only outputs the segmentation, but also gives numerical estimates of its entropy and probability. Using the latter the cost of phrase probability is computed as

$$C_p = -log(P(\sigma^*)), \tag{7.8}$$

where $P(\sigma^*)$ is the probability of the candidate phrase's best segmentation σ^*.

Cost C_p penalises phrases with 'unlikely' segmentations. In the context of our template segmenter, a low cost implies that the candidate phrase contains many figures matching exemplars stored in the machine's long-term memory (LTM). In Chapter 6 we presented an approach to simulate the LAM of listeners 'enculturated' in a given melodic tradition. In this section, however, we prefer a set of exemplars simulating the LTM of average listeners, not biased towards any particular melodic tradition. We hence use the NS figure learner described in §6.5.2, which can be expect to do just that. (We use the same parametric settings and training data.)

7.4 Evaluation

We evaluate our segmenter in a traditional scenario, i.e. by comparing machine-estimated segment boundaries to human-annotated segment boundaries. The experiments described in this section have two main goals. First, benchmark our multi-cue segmenter against existing ones, to compare and validate our optimisation approach. Second, analyse the extent to which our approach might be integrating cues in a way similar to human listeners, to validate/reject the hypothesis of parallel processing mechanisms in human melody segmentation.

[49] Our template segmenter actually computes the segmentation with maximum *entropy*. That said, Berger et al. (1996) showed analytically that a model that satisfies maximum likelihood constraints is also the same that satisfies maximum entropy constraints.

7.4.1 Preliminaries

In the following subsections we motivate our choice of reference and baseline segmenters, describe our test corpus and evaluation metrics, and finally specify parametric settings.

Baseline Segmenters. We want to assess the performance of simple segmentations using the boundaries in $\hat{B}$. To that end we define two baselines: UNION and MIDDLE. UNION simply takes all boundaries in $\hat{B}$. MIDDLE samples $\hat{B}$ incrementally, one boundary at a time. If $p_{i+1} - p_i \leq 4$ notes p_{i+1} is not included in the sample.

Reference Segmenters. We take GROUPER as a reference. While it would certainly be interesting to test the other approaches reviewed in Table 7.1, most of them have complex architectures and are not specified in sufficient detail for implementation. The one exception is ATTA. However, a preliminary experiment showed that a random assignation of parameters results very low performances, and manually tuning the system for each melody in the test dataset proved unfeasible.

Test Dataset. All machine segmenters studied in this chapter are evaluated in the `FJ375` corpus – refer to §3.5 for a description.

Evaluation Measures. In this thesis we use the well known $F1$, precision P, and recall R measures, defined in Equations 3.1, 3.2, and 3.3, respectively. Phrase boundaries are considered as predicted correctly (a 'true positive') if the prediction identifies either the last event of an annotated phrase or the first event of the following phrase.

To take into account the possibility of fuzzy boundaries during evaluation (see discussion in §3.3.1), we also use the Boundary Edit Distance Similarity B, defined in Equation 3.4. One of the parameters of the B measure is a tolerance window (in notes). Within this tolerance window boundaries are given a partial score proportional to their relative distance. We tested the effect of soft disagreement by computing the B using a tolerance of four notes.

Parametric Settings. We use GROUPER with its default settings. That said, in many of the melodies of the `FJ375` corpus metric information is unreliable. Thus, for our experiment the input to GROUPER is the same as that feed to our segments: a list of pitch-onset pairs.

Our multi-cue segmenter, which during this section we call κUES for short, has only one parameter: α, which controls the mixing of vertex and edge weights in the graph. In our experiments $\lambda = 0.6$ worked best. κUES uses the repetition, contrast, and tem-

plate segmenters, introduced in Chapters 4, 5, and 6, respectively. These segmenters are used with the parameter settings that proved best for the experiments conducted in their respective chapters. KUES also includes the LBDM gap segmenter (Cambouropoulos 2001). LBDM is used with its default parameters. Only the template segmenter is required at runtime, so all the rest processed the corpus off-line.

7.4.2 Results and Discussion

In Table 5.4 we present mean recall $\overline{R}$, precision $\overline{P}$, and $\overline{F1}$ results. We tested the statistical significance of the paired $F1$, B, P, and R differences between the compared segmenters. For the statistical testing we used a non-parametric Friedman test ($\alpha = 0.05$). Furthermore, to determine which pairs of measurements significantly differ, we conducted a post-hoc Tukey HSD test.

Note 1: In our discussion, when we say a difference is significant, we mean it is significant according to our hypothesis testing settings above.

Note 2: The performance estimates for $\overline{B}$ are in all cases higher than the $\overline{F1}$ estimates. We believe the B estimates are more likely to match segmentation performance as judged by humans than those of the $F1$. However, to make the analysis easier to follow, in the following sections we focus on discussing $F1$, P and R performances. Moreover, due to annotation issues (discussed in §3.3), determining if a machine estimated boundary is a false positive or not is at present unfeasible from our test corpora. Therefore, our discussion tends to favour precision over recall. We focus on comparing only very related approaches to be able to have some grounds in assuming a segmenter with high precision is in fact 'better' than one with high recall.

In what follows we analyse the results shown in Table 7.2. We first discuss how KUES compares to GROUPER and the baseline segmenters. Then we discuss more specific aspects of performance: the possible benefits of our approach to multi-cue segmentation, and look for evidence to validate/reject the hypothesis of parallel processing mechanisms in human melody segmentation.

Performance Analysis of the References and Baselines

The $\overline{F1}$ of GROUPER over the three subsets is uneven. It appears specially adequate for the segmentation of jazz melodies, where it obtains an $\overline{F1}$ of 0.68. And less so for vocal and instrumental folk songs, where performance drops to the 0.50–0.55 range. This might initially seem counter intuitive, as one could expected the melodies of the jazz subset to have a more complex phrase structure than those of the folk subsets.

Database	Folk Vocal				Folk Instrumental				Jazz			
Segmenter	$\overline{R}$	$\overline{P}$	$\overline{\mathbf{F1}}$	$\overline{\mathbf{B}}$	$\overline{R}$	$\overline{P}$	$\overline{\mathbf{F1}}$	$\overline{\mathbf{B}}$	$\overline{R}$	$\overline{P}$	$\overline{\mathbf{F1}}$	$\overline{\mathbf{B}}$
KUES	0.63	0.68	**0.65**	**0.71**	0.59	0.67	**0.63**	**0.70**	0.61	0.65	0.67	**0.73**
s GROUPER	0.48	0.56	0.50	0.54	0.55	0.57	0.54	0.62	0.77	0.64	**0.68**	0.70
MIDDLE	0.59	0.43	0.42	0.47	0.57	0.36	0.39	0.44	0.55	0.41	0.45	0.48
UNION	0.73	0.44	0.43	0.41	0.71	0.41	0.46	0.43	0.82	0.39	0.46	0.41
RND10%	0.17	0.25	0.20	0.22	0.17	0.19	0.17	0.15	0.17	0.25	0.20	0.21
ALWAYS	1.00	0.10	0.17	0.16	1.00	0.02	0.04	0.03	1.00	0.10	0.17	0.16
NEVER	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 7.2: Performance of KUES variants, reference segmenters, and baselines (abbreviations are defined in the text). From left to right: mean recall $\overline{R}$, precision $\overline{P}$, $\overline{F1}$, and $\overline{B}$ Higher performances are marked in bold.

We have a factual explanation for the observed unevenness: the jazz subset has a bias to exhibit temporal gaps at boundaries. More than half of the melodies in the jazz subset are extracted from performed (multi-track) MIDI files. This entails that space for ornamentation (e.g. a bar or half-bar for a piano arpeggio) results in 'unnaturally' large temporal gaps when the melody track is separated from its polyphonic context. Hence, temporal gap algorithms producing many boundaries are likely to get at least a good $\overline{R}$ performance. This can be observed to be the case for GROUPER in our results. We can hence speculate that GROUPER's performance in the jazz subset is slightly overestimated, and expect the true performance to lay a bit closer to that of the other two subsets.

The baselines have performances that agree more with our expectations. The performances of UNION seem dominated by recall, which is easy to anticipate given the large number of predictions likely to be contained in the union set $\hat{B}$. On the other hand, MIDDLE evidences a drop in both $\overline{R}$ and $\overline{P}$ with respect the performances of UNION. The drop is also expected, given the naïve way in which boundaries from $\hat{B}$ are selected.

Comparing Our Segmenter to References and Baselines

κUES $\overline{F1}$ performances are relatively similar across subsets. This suggests that the way segmentation cues are estimated and combined generalises to all three subsets. κUES performances show significant improvements (of over ~20% $\overline{F1}$) in respect to the baselines. It also shows to outperform GROUPER in the folk song subsets by ~8%. In the jazz dataset GROUPER obtains the highest performance. However, it is found to be statistically equivalent to that of κUES. It is worth stressing that, as discussed above, the performance of GROUPER in the jazz subset might be overestimated, due to its over-reliance on recall. κUES performance in the jazz subset seems driven by precision, making it less likely that it is being overestimated.

Possible Benefits of our Approach to Multicue Segmentation Modelling

κUES performance in all subsets are driven by precision. This suggests that from the relatively large number of initial boundaries in $\hat{B}$, our approach to cue combination is able to correctly select boundaries that make for a cognitively plausible segmentation. During previous experimentation we tested by modifying edge weights. From the initial set of three costs (phrase length, contour, and probability), we tested configurations where only pairs of cots where used, and also using each cost function separately. In all cases performances dropped. The experiments suggested a larger contribution

of phrase length and contour costs than of probability cost. However, it was only when all three cues are integrated that the best performance was obtained, stressing the importance and benefit of considering multiple sources of information when making segmentation decisions. In the experiments conducted in previous chapters we have observed a similar benefit when using multiple sources of information. We can also observe that the performances of our multi-cue segmenter are not only the highest in the experiment conducted in this chapter, but are also higher than the performances obtained by any of the single-cue segmenters introduced in previous chapters – see evaluations in §4.6 and §5.5. These results support our hypothesis of human segmenteing mechanisms as being composed of multi-scale, multi-cue, parallel processing modules.

7.5 Conclusions

In this chapter we introduced a novel multi-cue segmenter, targeted to segmenting melodies into phrases. We proposed an approach to modelling cue combination based on optimisation. Our segmenter differs from existing multi-cue segmenters in two respects. First, it is more complete, in that it has a wider coverage of cues. Second, it has a higher degree of autonomy, in that it has dedicated modules to estimate all needed cue information. An added feature of our segmenter is the interpretability of its mechanisms, made possible by its modular approach and cue-defined cost function.

We evaluated our segmenter on the `FJ375` corpus. To have a comparison point we also evaluate GROUPER and two naïve baseline segmenters on the same corpus. GROUPER is currently at the state of the art of melodic segmentation. Results show that our segmenter achieves statistically significant $\overline{F1}$ improvements of ~8% in respect to GROUPER, and of over 20% in respect to the baselines. Our results also show clear benefits of using multiple sources of information for segmentation, which supports the hypothesis of human segmenting mechanisms as being composed of multi-scale, multi-cue, parallel processing modules.

Future Work

In the following we briefly outline our outlook of future work.

Output a Segmentation Raking. As presented in this chapter, our approach searches for the optimal segmentation. Yet, the formalisation of our approach, being framed as a graph search problem, is extensible to output a ranking of segmentations. To this end modifications of the algorithms by Dijkstra (1959) or Yen (1971) can be explored.

Alternative Models of Segment Length. We rely on a corpus-based estimate of phrase length probability. This requires a phrase annotated corpus. Parametric modelling alternatives, as those presented in (Rhodes et al. 2006; Abdallah et al. 2006), might be investigated to eliminate this dependency.

Phrase Form Templates and Metric Structure. Our framework can be extended to incorporate other cues. Two cues important cues missing are phrase form templates and alignment to beat structure. A suitable model for phrase form templates can be (Bod 2001, 2002). A suitable model to infer beat structure automatically is presented in Ahlbäck (2007).

Chapter 8

Conclusions

In this dissertation we tackled the problem of developing machines to segment melodies. Our aim in developing these machines has been augmenting the state of knowledge about segmentation within the disciplines of Music Information Retrieval and Computational Modelling of Music Cognition. We have done so in three ways:

1. Introduced a conceptual framework of segments and segmentation.
2. Proposed techniques and developed resources to evaluate machine segmenters.
3. Developed and tested four different, yet complementary machine segmenters.

In this final chapter we briefly recapitulate our main findings and give an outlook for future work.

8.1 Findings

In Chapter 2 we identified two factors hampering segmentation research: unclear terminology and unclear goals. We therefore introduced a conceptual framework to guide the development of machine segmenters, grounded on cognitive theories of music listening. The framework is composed of a conceptual model and a taxonomy. The conceptual model consists of working definitions for what a segmenter is (as a cognitive mechanism) and how it operates. The taxonomy classifies both the processing mechanisms (subcomponents) and information (cues) needed for the segmenter to operate. Moreover, we provided working definitions of segments and segment types, and defined computational modelling tasks. The conceptual framework was used to classify existing segmenters, identify niches of novel research, and motivate/guide the development of melody segmenters introduced in this dissertation.

In Chapter 3 we critically reviewed the evaluation chain of machine melody segmenters. At present automatic segmentations are evaluated by comparing them to manual segmentations. We identified three important limitations of this evaluation scenario: first, available segment-annotated databases lack stylistic diversity; second, currently used evaluation measures give no partial score to nearly missing a boundary; third, due to the low number of boundary annotations per melody, it is impossible to estimate how to penalise an insertion or full miss. Our contributions to tackle these limitations were threefold: we present a new benchmark corpus consisting of 125 jazz melodies which helps broadening the stylistic diversity of annotated corpora; we survey measures proposed in the field of text segmentation that can give partial scores to near misses; we propose an approach to help extending the annotations of existing corpora to allow better penalisation of insertions and full misses.

In Chapter 4 we tackled the problem of *repetition*-based melody segmentation. Our focus was investigating the role of location related information on the automatic identification of repetitions. We introduced three complementary scoring functions based on location information to identify repetitions. To test the ability of our functions to identify repetitions, we incorporated them in a state-of-the-art repetition based segmenter. The original and function-extended versions of the segmenter were used to segment melodies from the `FJ375` corpus. Results show that by using our scoring functions the segmenter achieves a statistically significant 14% average improvement over its original version. The improvement suggests that the location of similar fragments found within a melody is an important aspect for humans to judge/recognise one as being a repetition of the other(s).

In Chapter 5 we tackled the problem of *contrast*-based melody segmentation. We

focused on investigating the role of attention and multi-scale perception when determining contrasts. We introduced a novel approach to model automatic contrast identification based on statistical hypothesis testing. Attention modelling was tackled using methods from information theory and statistical model selection. Multi-scale modelling was tackled using methods from classifier combination. We evaluated our segmenter on the FJ375 corpus. Results show that our segmenter achieves a statistically significant 10-12% improvement in precision in respect to four reference segmenters. If our segmenter is combined with a gap segmenter the combination achieves a statistically significant 10-20% average F1 improvement over the reference segmenters. Our results suggests that the perception of local difference has two facets, one at narrow time-scale and another at a wider time scale. And that both of these facets lead to cues (gaps and contrasts, respectively) which have a seemingly complementary role in phrase boundary perception.

In Chapter 6 we tackled the problem of segmenting melodic phrases into subphrases based on template cues. More specifically, we focused on modelling segmentation influenced by the recognition of previously heard melodic figures. We introduced a statistical learning approach to learn melodic figures. And introduced a maximum entropy approach to model template-based segmentation. We tested the usefulness of the subphrases produced by our template-based segmenter in a classification experiment. In the experiment automatically determined subphrases were used as features to estimate the cultural origin melodies. We compared to three reference segmenters and two naïve baselines. The expectation being that more 'better' segmentations would lead to more discriminative features, and hence higher classification accuracy. Results show that using subphrases produced by our segmenter results in a 5% to 8% increase in classification accuracy over the references and baselines. These results suggest that acquiring a discriminative 'lexicon' of melodic figures is a goal-driven process in which segmentation and learning act conjointly, and moreover reinforce each other.

In Chapter 7 we tackled the problem of automatically segmenting complete melodies into phrases by combining multiple cues. We formulated multiple cue segmentation as an optimisation problem, and introduced a cost function that penalises segmentations by considering cues related to boundaries, segments, and the complete segmentation. Our segmenter differs from existing multi-cue segmenters in two respects. First, it is more complete, in that it has a wider coverage of cues. Second, it has a higher degree of autonomy, in that it has dedicated modules to estimate all needed cue information. An added feature of our segmenter is the interpretability of its mechanisms, made possible by its modular approach and cue-defined cost function. We evaluated our segmenter on the FJ375 corpus. To have a comparison point we also evaluated a state-of-the-art

multi-cue segmenter and two naïve baseline segmenters on the same corpus. Results show that our segmenter achieves statistically significant $\overline{F1}$ improvements of ~8% in respect to the state-of-the-art, and of over 20% in respect to the baselines. Our results also show clear benefits of using multiple sources of information for segmentation, which supports the hypothesis of human segmentation mechanisms as being composed of multi-scale, multi-cue, parallel processing modules.

8.2 Outlook

Automatic music segmentation is a fundamental, yet challenging problem. At present, performance estimates suggest that machine produced segmentations are not yet in par with those produced by humans. The work presented in this dissertation, however, brings us a step closer to achieving this goal. What is more, our results make it feasible to foresee a not-so-distant future where, at least for melodies, automatic and manual segmentations will be indistinguishable.

There are numerous directions for future work in melody segmentation. Many of these directions have already been discussed in chapter conclusions sections. Below we revisit and summarise the ones we believe to be most important.

Multicue Approaches. Through this dissertation we have maintained the view of human segmentation mechanisms as a network of parallel processing modules with a (semi)centralised decision making unit. Our experimental results have provided confirmatory evidence for this view. We therefore encourage future work on approaches suitable for modelling distributed/parallel problem solving. One alternative is to extend the approach presented in Chapter 7. Extensions can tackle the problem of how to incorporate alignment cues (metric alignment is strongly encouraged), and closure cues. Also it is necessary to incorporate algorithms which enable outputting a ranking of plausible segmentations instead of single optimal one as it is done at present. Alternatively, multicue segmenters could be modelled using Bayesian networks or neural networks. An interesting unsupervised word segmenter using Bayesian networks is (Goldwater et al. 2009; Pearl et al. 2010). An interesting supervised audio music segmenter using convolutional neural networks is (Grill and J. 2015; Ullrich et al. 2014).

Alternative Evaluation Strategies and Measures. Arguably the main obstacle hampering melody segmentation research is the lack and quality of segment annotated corpora. Luckily this appears to be changing, as a good number new segment annotated collections have been released in the last couple of years – see for instance (Abeßer et al. 2013; Karaosmanoglu et al. 2014; van Kranenburg et al. 2014; Rodríguez-López

et al. 2015). With these new collections we can test for generalisation of our segmenters to different styles, cultural traditions, and even 'type' of monophonies (the first cited collection is comprised of jazz solos). However, most of these collections have boundary annotations of boundaries at a single scale (mostly phrases), which more over have been produced by a single annotator. Therefore the problems related to segmenter comparison ambiguity remain – see Chapter 3. There is hence an urgent need to develop new evaluation strategies and measures that can ameliorate these problems. One possible strategy is to define 'toy' MIR tasks specifically designed to evaluate segmenters so that there is no need to compare to manually annotated boundaries – for a more extensive discussion on ideas of this kind (and some guidelines) refer to (Rodríguez-López and Volk 2015b).

Expand Existing Corpora Annotations. Producing manually segmented corpora is a time consuming and overall tedious task. It is hence unlikely that the current reality (low number of annotations per piece, and so on) will change in the short-term. It is then necessary for the community to make the most of existing annotated collection. This entails conducting manual and automated analyses to deepen our understanding of the annotations. We want to know things like: what type of cues where the annotators paying attention to? What to what music dimension (pitch, timbre, etc.) where they attending? Or how likely it is that the annotator might have made a mistake? To this end some analyses that could be taken as reference are (Smith et al. 2013; Smith and Chew 2013a,b; McFee et al. 2015). Another alternative is to define efficient revision strategies so that annotators can enrich existing annotated data. For instance have an annotator rate (on a binary scale) how much he 'agrees' with an existing choice of boundary locations. The produced binary judgements can be used as confidence weighting mechanism in performance measures.

Appendix A

Melody as an Attribute Sequence

In this dissertation melodies are represented as a piano roll. That is, a sequence of events $e = e_1 e_2 \dots e_n$ where each event has information of pitch, onset, and offset. We use the `MIDI` standard (MMA 1995) to encode note attribute information. Pitch is given a discrete representation specified using midi note numbers. Duration is derived from onset/offset times, which are in turn given a continuous representation specified in midi ticks. That said, test melodies either correspond to encoded scores (so that duration is quantised by default), or have been quantised manually. Quantisation is only required by one of our segmenters. Nevertheless, we used the same input for every segmenter during testing.

We denote pitch, onset, and offset information of an event as

$$p(i) = p(e_i) \tag{A.1}$$

$$on(i) = on(e_i) \tag{A.2}$$

$$of(i) = of(e_i) \tag{A.3}$$

Below we list the formulas used to compute attribute sequences based on basic event information.

Computation of Pitch Attribute Sequences

· *Pitch Class.*

$$cpc(i) = 1 + p(i) \bmod 12 \tag{A.4}$$

·*Chromatic Pitch Interval.*

$$cpiv(i) = p(i+1) - p(i) \tag{A.5}$$

We truncate intervals to be within an octave (12 semitones).

·*Step-Leap Pitch Interval.*

$$piv\text{-}sl(i) = \begin{cases} -2 & cpiv(e_i) < -3 \\ -1 & -3 \geq cpiv(e_i) \leq -3 \\ 0 & cpiv(e_i) = 0 \\ +1 & 1 \geq cpiv(e_i) \leq 3 \\ +1 & cpiv(e_i) > 3 \end{cases} \tag{A.6}$$

·*Pitch Contour.*

$$pctr(i) = \text{sgn}(cpiv(e_i)) \tag{A.7}$$

Computation of Duration Attribute Sequences

· *Inter-Onset-Interval.*

$$ioi(i) = on(i+1) - on(i) \tag{A.8}$$

· *Onset-to-Offset Interval.*

$$ooi(i) = of(i+1) - on(i) \tag{A.9}$$

· *Offset-to-Onset Interval.*

$$rest(i) = on(i+1) - of(i) \tag{A.10}$$

·*Inter-Onset-Interval Ratio.*

$$ioir(i) = \frac{ioi(i+1)}{ioi(i)} \tag{A.11}$$

·*Inter-Onset-Interval Ratio Contour.*

$$ioi\text{-}rc(i) = \text{sgn}(ioir(i)) \tag{A.12}$$

·*Inter-Onset-Interval Class.*

$$piv\text{-}sl(i) = \begin{cases} +2 & ioi_n > 3.3 \\ +1 & 1.8 < ioi_n \leq 3.3 \\ 0 & 0.9 < ioi_n \leq 1.8 \\ -1 & 0.45 < ioi_n \leq 0.9 \\ -1 & ioi_n \leq 0.45 \end{cases} \tag{A.13}$$

where $ioi_n = \frac{ioi(i)}{\text{Mo}_{ioi}}$, and Mo_{ioi} is the most frequently occurring ioi in the sequence

Appendix B

Cognitive Theories of Music Segmentation

In this appendix we first briefly review the two most popular theories of music segmentation, then discuss cognitive structuring processes that work parallel or in combination to segmentation, and lastly discuss the notion of 'musical surface'.

B.1 Theories of Segment Perception and Cognition

There are anumber of cognitive theories of music that touch on the topic of segmentation (Lerdahl and Jackendoff 1983; Wiggins and Forth 2015; Hanninen 2001; Narmour 1990, 1992; Deliège 2001; Ockelford 2004; Ahlbäck 2004; Godøy 2009). In this section we briefly review two of them: Leardahl & Jackendoff's *Generative Theory of Tonal Music* (GTTM) (1983), and Narmour's *Implication-Realisation* (IR) theory (1990).[50] We have chosen to focus on these two theories due to their strong influence both on music cognition research and on the development of computational models of music segmentation. Below we give a short description of the theories, focusing the aspects directly related to music segmentation.

[50] Both theories enjoy wide popularity, and have been thoroughly described and critically reviewed in a number of publications. For the sake of brevity, we omit a thorough description and refer the reader to the original publications for details. Also, we recommend (Pearce et al. 2010b) for a succinct, yet well-balanced summary covering the two theories.

GTTM: short description

The GTTM theory attempted to thoroughly (yet not formally) describe the cognitive principles a listener develops in order to acquire the musical grammar necessary to understand a particular musical idiom. The model is strongly influenced by the generative grammars of Chomsky (1965; 1957), and as such presents arguments for the universality and innateness of the principles proposed. These principles are assumed to represent the final state of understanding of an experienced listener of tonal music, rather than on-the-fly mental processes.

GTTM: grouping principles

A summary of the principles of the theory related to segmentation is presented in Table B.1. The segmentation principles, called 'Grouping Preference Rules' (GPRs) in GTTM, are generally classified into three types: The first type of rules, GPRs $2_{a,b}$ and $3_{a,b,c,d}$, are based on the Gestalt principles of proximity and similarity (change). The second type of rules, GPRs 5 and 6, are based on symmetry and motivic similarity. The third type, GPR 7, is based on grouping effects of pitch structure (time-span reduction and prolongation stability). In addition to the above mentioned groups, there are two extra rules, GPRs 1 and 4, which give, respectively, a general guideline concerning length of segments, and a suggestion on how to classify segment boundaries.

IR: short description

In the IR theory music listening is treated as a dynamical process. This theory has a narrower scope than that of GTTM, focusing solely on melody structure. Influenced by the writings of Meyer (1956), the theory proposes to understand melodic perception as a process of fulfilled and unfulfilled expectations. The theory formalises melodic structure perception as the mediation between two systems, one encoding the musical experience of a listener (top-down), and another acting as a set of innate rules (bottom-up).

IR: grouping principles

In the IR theory segment boundaries are hypothesized determined mainly by *melodic closure*. Melodic closure indicates points where an ongoing cognitive process of melodic expectation is disrupted (Pearce and Wiggins 2006b), i.e. points in a melody that

GPR	Name	Description
1	-	Avoid analyses with very small groups –the smaller the less preferable.
2	Proximity	Consider a sequence of four notes n1 n2 n3 n4. Ceteris paribus, the transition n2-n3 may be heard as a group boundary if:
	a. Slur/Rest	the interval of time from the end of n2 to the beginning of n3 is greater than that from the end of n1 to the beginning of n2 and that from the end of n3 to the beginning of n4.
	b. Attack-point	the interval of time between the attack points of n2 and n3 is greater than that between n1 and n2 and that between n3 and n4.
3	Change	Consider a sequence of four notes n1 n2 n3 n4. Ceteris paribus, the transition n2-n3 may be heard as a group boundary if:
	a. Register	the transition n2 to n3 involves a greater intervallic distance than both n1 to n2 and n3 to n4.
	b. Dynamics	the transition n2 to n3 involves a change in dynamics and n1 to n2 and n3 to n4 do not.
	c. Articulation	the transition n2 to n3 involves a change in articulation and n1 to n2 and n3 to n4 do not.
	d. Length	n2 and n3 are of different lengths, and both pairs n1, n2 and n3, n4 do not differ in length.
4	Intensification	Where the effects of Group Preference Rules 2 and 3 are relatively more pronounced, a larger level group boundary may be placed.
5	Symmetry	Prefer grouping analyses that most closely approach the ideal subdivision of groups into two parts of equal length.
6	Parallelism	Where two or more segments of the music can be construed as parallel, they preferably form parallel parts of groups.
7	Time-Span	and prefer a grouping structure that results in more stable time-span prolongation and/or prolongation reductions.

Table B.1: GTTM grouping rules (Lerdahl and Jackendoff 1983) as summarized by Frankland et al. (2004). GPR stands for 'Grouping Preference Rule'.

provide a listener with a sense of completion. The bottom-up rules proposed by Narmour to estimate the degree melodic closure are presented in Table B.2. These rules operate considering pairs (and in one case triplets) of contiguous melodic intervals.

	Name	Description
1	Rest Closure	An interval is followed by a rest.
2	Durational Closure	The second tone of an interval has greater duration than the first;.
3	Registral Direction Closure	A change in registral direction between the two intervals described by three successive notes.
4	Metrical Closure	The second note of an interval occurs in a stronger metrical position than the first.
5	Interval size closure	Three successive notes create a large interval followed by a smaller interval.
6	Tonal closure	The second note of an interval is less dissonant in the established key/mode than the first.

Table B.2: Melodic closure rules of the I-R theory (Narmour 1990) as summarized by Pearce et al. (2010b).

B.2 The Musical Surface

In the theories outlined above the assumption is that these processes occur over a 'surface level'. The *musical surface* can be thought of as the lowest level of detail that is of 'musical significance' to a listener (Lerdahl and Jackendoff 1983). An alternative definition, which is perhaps less controversial, is to think of the musical surface as the lowest level of detail that is *of interest* for a given task (Pearce 2005). The surface level is assumed to be comprised of acoustic events which are indivisible from a perceptual perspective, which constitute the primitives of cognitive representation. It is also assumed listeners can perceive these events as occurring sequentially or simultaneously. In Figure B.1 we present an example of common musical attribute descriptors and the surface level commonly used by machine segmenters.

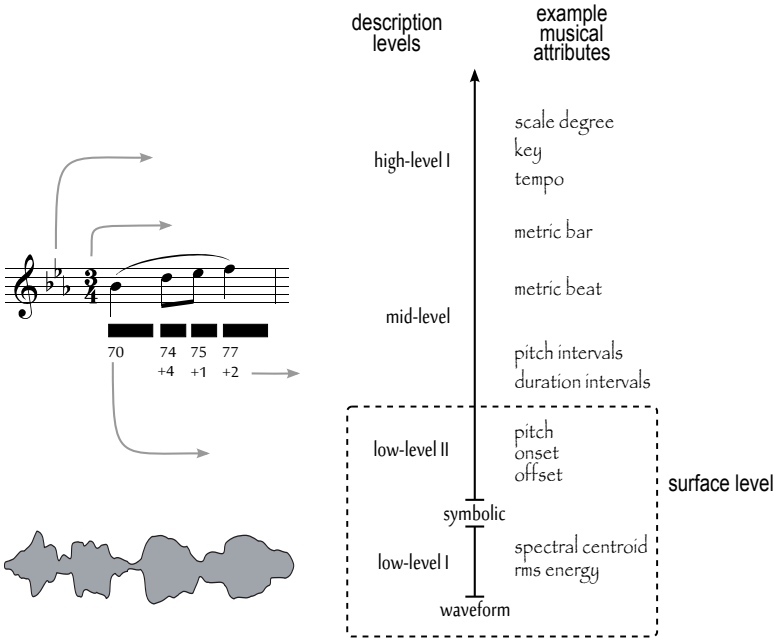

Figure B.1: Diagram illustrating levels of description and types of music structures. The description levels and example musical attributes have been organized following the taxonomization of musical attributes proposed by Lesaffre et al. (2003).

In this dissertation notes are assumed to be the primitives of cognitive representation of melodies. That is, their musical surface. This assumption is, however, subject of wide controversy and debate within the Music Cognition community. We briefly discuss three common points of criticism below.

First, notes are often assumed to be acoustic events where musical attributes such as pitch, timbre, loudness remain more or less constant, thus enabling human listeners to mentally represent them by using categorical systems. That is, a surjective mapping from sound events to a finite set of categories. The view of notes as an appropriate abstraction of sound events has been contested in modern theories of music perception. For instance, Wiggins and Forth (2015, p. 131) suggest that minds/brains are able to maintain dual representations of musical attributes that have both discrete and continuous properties, thus making the notion of note as used traditionally in music notation limited (and in cases completely inadequate) for scientific research related to music cognition.

Second, notes are often assumed to be rich enough to represent a vast amount of music. However, researchers in music cognition often argue that many instances of existing music can simply not be notated using notes, e.g. see (Godøy 2009, p. 68; Walker 1997).

Third, it has been hypothesised that the sound events used as the fundamental units in the construction of a segment structure are composed of units determined by imagining sound producing gestures (Godøy 2014). This has a mirror in Linguistics and Natural Language Processing research, where it is often argued that diphones or triphones have a more fundamental role as elementary structural units than single phones, due to the role of co-articulation.

Appendix C

Summary of Empirical Studies of Segmentation

In this appendix we briefly discuss listening studies performed to study music segmentation. We have two goals: (1) to provide a list of cues identified as significant for segment perception, and (2) to provide a summary, in the form of a list of findings, that represent the current state of understanding of segment perception and cognition from the perspective of empirical research.

C.1 Perceptual Studies: Setting and Potential Biases

Perceptual studies generally investigate segmentation cues by means of *annotation experiments*, i.e. having human subjects listen to a given set of musical pieces and 'annotate' the locations of the boundaries. The experimental setting of perceptual studies thus includes: a sample of music (stimuli), a sample of population (test subjects), a way to conceptualize and communicate the task to the subjects, and a way to have test subjects indicate boundary locations. In the following sections we discuss perceptual studies in respect to the characteristics of the stimuli used (§C.2), the level of agreement between test subjects (§C.3), and the types of cues seen as relevant for segment boundary perception (§C.4).

C.2 Stimuli Characteristics

In this section we first describe and discuss the stimuli used in respect to: *origin* (i.e. whether stimuli is sampled from existing music or artificially generated), and *format* (i.e. whether recordings, 'expressive' synthesis -including performance related information-, or 'dead pan' synthesis -as notated on a score- was used). Subsequently, we describe and discuss stimuli in respect to general attributes, namely: the *number* of music fragments or artificial tone sequences used, their *duration*, *texture*, and (if applicable) *style/genre*.

Origin and format of the stimuli

In Table C.1, three of the studies surveyed (Deliège 1987; Weyde 2003; Weyde et al. 2007) employed artificial tone sequences as stimuli. In these studies sequences of tones were synthesized using wave-table synthesis, providing control over amplitude, pitch, duration, and timbre (using the General-MIDI standard). A thorough survey of studies pre-dating those of Table C.1, and which also use artificial tone sequences to investigate segmentation, can be found in (Deutsch 1999). The tone sequences used in the studies surveyed by Deutsch were normally synthesized using additive synthesis, providing control over amplitude, fundamental frequency (pitch) and harmonics (timbre), onset-to-offset intervals (duration), offset-to-onset intervals (pauses), and presentation rate (tempo).

The rest of the studies in Table C.1 employed existing music. The columns labelled as *format* and Q in Table C.1 show that in most cases recordings and 'expressive' synthesis has been preferred over 'dead pan' synthesis. The synthesis technique employed in these studies is commonly wave-table synthesis, using the General MIDI specification for timbre.

The goal of studies employing real stimuli is to test the relevance of segmentation cues in a setting with high 'ecological validity', which in some cases is absolutely indispensable, e.g. the research of Spiro (2006) on performance-related cues for segmentation. On the other hand, artificial stimuli is preferred to investigate potential boundary cues in isolation, e.g. in (Deliège 1987) where GTTM grouping cues were tested. Or in combinations, to test for aspects such as dominance of a specific cue in relation to others, e.g. in (Weyde 2003; Weyde et al. 2007) where the role of pitch information was tested against duration information. The seemly unavoidable trade-off between ecological validity and experimental control is a well addressed issue in perceptual studies – see (Margulis 2012) for a discussion.

Gobal stimuli characteristics: stylistic diversity

Inspecting the columns *test set* and *duration* in Table C.1 we can see that, in general, artificial stimuli is comprised of relatively short monophonic sequences (10-20 tones per sequence). Artificial sequences are also normally isochronous and uniform, and the task given to the test subjects normally requires to distinguish relatively short segments (groups of 2-4 notes). The table also shows that experiments using existing music have used mostly small stimulus sets, and, just as with artificial stimuli, relatively short musical material (7 of the 9 studies commonly use excerpts lasting less than 1 minute). However, these studies normally show a balanced mix of tonal and non-tonal stimuli, with an equally balanced mix of polyphonic and monophonic textures. Despite the mix of textures and styles, observing at the column *genre* we can see a strong preference for western 'art' music.

The stimuli characteristics mentioned above clearly impair the investigation of temporally distant factors in segment perception, and, in the case of the real stimuli, also yield the conclusions of the studies as essentially style specific.

C.3 Subject Consistency

In the surveyed perceptual studies the amount of test subjects ranges between 7 and 45. Recruited test subjects normally possess different levels of 'musical training', to investigate the degree to which training influences the perception of segments. Following Spiro and Klebanov (2006), we have classified the subjects' level of musical training into three classes: degree-level musicians (DL), amateur musicians (M), and non-musicians (N).

In perceptual studies, subject consistency has been assessed in respect to the following:

> *Across-Trial-Agreement* (ATA): level of consistency of boundary locations chosen in consecutive experimental trials (listening of a stimulus) by a given test subject.
>
> *Across-Subject-Agreement* (ASA): level of consistency two or more test subjects have in respect to a given stimulus.
>
> *Across-Training-Group-Agreement* (ATGA): level of consistency between DL|M|N test subjects in respect to a given stimulus.

Most perceptual studies have only directly assessed ATGA. Studies more directly concerned which testing of automatic segmenters have been more interested in ASA, given that in these cases the segments boundaries identified by subjects are used as

a 'ground truth' or 'golden standard' against which the predictions made by machine segmenters can be compared. ATA has been rarely systematically assessed. Below we discuss each type of agreement in turn.

In respect to ATA, Spiro and Klebanov (2006) indicated that several subjects had given more than one phrase-level interpretation (in three experimental trials) of 4 European pieces of 'art' music belonging to the Common Practice Period. This finding also agrees with the study of Bruderer (2008, Ch.2), where subjects where found to be only moderately self-consistent when identifying phrase-level segment boundaries in six western popular songs.

In respect to ASA, we can divide our observations by segment granularity. Boundary locations with high ASA are normally reported to correspond to form-level boundaries. For phrase-level granularity, conversely, the level of agreement seems to be linked to the complexity of the melodies. As an example, high ASA was observed in the study of Wiering et al. (2009) where pop melodies where used for testing. On the contrary, the studies of Thom et al. (2002) and Pearce et al. (2010a), in which the test sets included jazz and classical melodies, report low ASA.

In respect to ATGA, the influence of musical training seems to not have a big influence with M and N type subjects. On the other hand, between DL and N the studies surveyed report contradictory findings. As an example we can mention Deliège (1987) who reports that non-musicians present a lower fit to GTTM principles than musicians, while Schaefer et al. (2004) reports the opposite. Also, in respect to the amount of segments produced by each type of subject Deliège (1987) observes that non-musicians perceive more boundaries than musicians, while Spiro (2006) reports the opposite.

C.4 Observed Segmentation Cues

Cues Observed in Short Artificial Tone Sequences

Deutsch (1999, ch. 3) surveys perceptual studies of grouping in artificial musical sequences.[51] The survey covers mostly early work ($1960s - 1980s$). The surveyed

[51] It must be noted that Deutsch treats the term *grouping* in a broader sense than the one used in this document. For Deutsch grouping describes the perceptual impression of connected tones, regardless of whether this connection separates a sequence of tones into contiguous sub-sequences or intertwined voices. In this dissertation a distinction is made so that the former is termed 'segmentation' or ''grouping', and the later is termed 'streaming'. The work surveyed by Deutsch deals, for the most part, with the study of streaming perception cues.

studies dealing with segmentation centre their attention on how 'temporal gaps' (i.e. silences between tones) might serve as segment boundary cues. The survey presents evidence that places temporal gaps as an important cue for the segmentation of short (3-4) tone sequences. Temporal gaps were even seen to override the perception of pitch patterns as a mechanism of segmentation, e.g. Handel (1973) found that, if in a sequence of 8 tones forming a pitch pattern a silence was introduced every 2 tones, subjects would report that it was easy to recognise the pattern, however, if in the same sequence a silence was introduced every 3 tones, subjects would report that recognising the pitch pattern was difficult). Interestingly, in the survey evidence is also presented for the fragility of the perception of temporal gaps (Deutsch 1999, p. 317). The surveyed studies therein suggest that if gaps are relatively brief, their perception is greatly affected by tempo and pitch interval size, to the extent that at fast tempos brief gaps might even not be perceived at all.

Two more recent studies (Weyde 2003; Weyde et al. 2007), included in Table C.1, focused on examining boundary cue combination and cue relevance. Weyde (2003) tested the common hypothesis that the combination of different cues for boundary perception can be modelled as a weighted sum. The cues studied were pitch interval sizes, pitch interval direction, loudness accents, and inter-onset-intervals. The task given to test subjects was to segment a sequence of 12 tones in either groups of 2 or 3 tones. The tone sequences were designed to have cue pairs, in a given parametric setting, suggesting groups of 2 tones and the same cue pair, in a different parametric setting, suggesting groups of 3 tones. The results obtained suggested that a weighted sum (linear model) is inadequate for modelling the combination of segmentation cues. Weyde et al. (2007) studied the relative relevance of 4 segmentation cues. The cues studied were again pitch interval sizes, pitch interval direction, loudness accents, and inter-onset-intervals. The task given to test subjects was to segment a sequence of 12 tones in either groups of 3 or 4 tones. The tone sequences were designed to have one cue suggesting groups of 3 tones and another cue suggesting groups of 4 tones. It was concluded that the size and direction of pitch intervals have less influence on segment perception than inter-onset-intervals and loudness accents.

Cues Observed in Existing Music: short-excerpts

Work using short excerpts commonly focuses on testing of theories presented in §B.1 (Schaefer et al. 2004; Deliège 1987; Brown et al. 2012). Most have focused on testing the GTTM rules.

Perhaps the main experimental verification of Lerdahl & Jackendoff's GPRs was that conducted by Deliège (1987). In the work, Deliège assessed GPRs $2_{a,b}$ and $3_{a,b,c,d}$, and

also suggested and examined two additional rules: change in timbre (instrumentation), and change in direction of melodic contour. She asked a group of musicians and non-musicians to segment 32 short recorded excerpts (3-16 notes) taken from the Common Practice repertoire (each excerpt containing only one GPR, 4 excepts per GPR). Deliège extended the previous experiment by generating 108 artificial melodic sequences to test for conflicts between rule pairs. In general, her findings support the existence of the grouping mechanisms proposed by Lerdahl & Jackendoff. As illustrated in Figure C.1, Deliège found significant differences in the salience of the different rules, with R. 7 (change in timbre) and R. 1,2 (slur/rest and proximity of attack point) being the most salient ones.[52] Her experiments also showed that the segmentation processes in musicians seem to be more directly linked to the GPRs than those of non-musicians. Also, she observed than non-musicians' segmentations to be more consistent with the GPRs when embedded in artificial sequences rather than existing music.

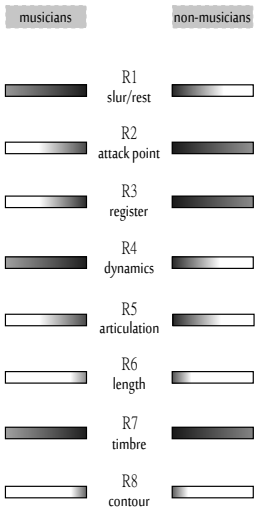

Figure C.1: Diagram of the relative strength (dark = strong) of the GPRs investigated by Deliège (1987). The study included GPRs $2_{a,b}$ and $3_{a,b,c,d}$, which in her work she renamed R1→R6. Additionally, Deliège included 2 extra rules (R7 & R8).

[52] Rule salience was measured weighting the number of times a boundary determined by a particular rule was chosen and the number of times the excerpt had to be repeated before subjects decided on a boundary location.

Cues Observed in Existing Music: whole-pieces

Work using longer excerpts follow a more open approach, instead of testing specific theories.

Deliège (1989) asked musicians and non-musicians to segment two complete pieces of contemporary music. Subjects heard the pieces 3 times and marked boundaries in real time. Deliège found that main segmentation cues were associated to silences/pauses, although only when observed in combination to other cues. The differences between boundaries marked by musicians and non-musicians was minimal.

Clarke and Krumhansl (1990) asked musicians to segment two complete pieces of classical music and to describe and rate the salience of selected boundaries. The obtained boundaries and their descriptions were found to be closely predicted by the rules of the GTTM. A sample of the the descriptions of the segment cues given by test subjects is presented in Figure C.2.

Stockhausen's klaviersück IX 10 strongest boundaries	CUES	Mozart's Fantasie 6 strongest boundaries
(chords-to-melody, pitch content, block chords) new material		new material (lyrical, dramatic, end of cadenza)
(chordal, chromatic run, coda) return of material		change of texture (thicker, thinner)
(expansion, jump) change of register		change of tempo
change of rhythm		change of register
change of dynamic		change of dynamic
(silence) pause		change of key
start of development		change of harmony
change of articulation		change of meter
change of texture		change of rhythm
change of pitch content		change of melody
relaxation of tension		
introduction of trill		
change of tempo		
(piano tone)change of timbre		

Figure C.2: Example of type of cues found by Clarke and Krumhansl (1990) for atonal (left) and tonal (right) piano pieces. The cues have been sorted in respect to a salience rating based on strength assigned to the boundary location and the total number of subjects using the cue. The strength the the boundary location was measured by counting the number of subjects which agreed on that location being a boundary. (The interested reader should refer to the publication for further details).

C.5 Concluding Remarks

1 None of the experiments has found strong evidence of stable and invariant cues. Where by stable we mean that is idiom-independent, and by invariant we mean that its presence (alone or in combination to other cues) can be determinant of segment boundary perception.

2 The number and type of cues have been observed to vary for each perceived boundary within a stimulus, and the importance of each cue involved has also been observed to vary: *"Cues can be of several kinds [...] it is the specific instance which is the determining factor"*(Deliège 1989, p. 228).

3 The most stable cues seem to be those associated to temporal proximity gaps (primarily rests or caesura, and to a slightly lesser degree prolonged note durations).

4 The average number of distinct cues reported to influence boundary perception for a single short piece of music is normally larger than 10.

5 Performance related cues influence boundary position in such a way that two performances of the same piece may vary in both the amount of boundaries and the position of those boundaries, in a manner proportional to the number of factors or cues present in the music (Spiro 2006).

6 The view that the most salient cues can be obtained from surface level information (focusing on change/sameness) (Clarke and Krumhansl 1990; Deliège et al. 1996), has now expanded to also acknowledge the importance of style-dependent factors (Spiro 2006; Schaefer et al. 2004; Bruderer and Kohlrausch 2009).

Table C.1: Perceptual studies of monophonic & polyphonic segmentation. **Fields** (columns from left to right): `Authors` - of the perceptual study, `Date` - of publication, `Focus` - main topic of experimental research, `Segs` - segment granularity addressed, `Test set` - number and type of music used as stimuli, `Texture` - Monophonic|Polyphonic|Homophonic texture of stimuli, `Genre` - of stimuli, `Duration` - of stimuli, `Format` - in which stimuli was presented, `Q` - onsets & durations are quantized (yes|no) `#` - number of subjects, `E` - degree of musical training (`DL`, `M`, `N`, I), `Cues` - number and type of segment cues described by subjects, `ATA` - reported level of agreement across different trials of experiment for a single subject, `ASA` - reported level of agreement across different subjects for stimuli, `Conclusions` - of study, **Abbreviations:** `record` - audio recording, `synth` - synthesised audio, `EX` - experiment number, `DL` - degree level musician, `M` - amateur musician, `N` - non-musician, `SR` - cues observable in a conventional western score, `PR` - performance-related cues, `ATA` - across trial agreement, `ASA` - across subject agreement. **Symbols:** ✓- yes, × - no, ⊖ - not clearly specified, ⊘ - does not apply.

Study			Stimulus						Subjects		Results			
Authors	Focus	Segs	Test set	Texture	Genre	Duration	Format	Q	#	Training	Cues	ATA	ASA	Conclusions
Deliège (1987)	experimental testing of GPRs $2_{a,b}$ & $3_{a,b,c,d}$ plus 2 extra rules	phrases	`EX1`: 32 phrases `EX2`: 108 artificial sequences	H M	baroque classical romantic early 20^{th} C	`EX1`: 3-16 notes `EX2`: 9 notes	`record` `synth`	⊘ ✓	60	`M`,`N`	8 `SR`	⊘	fair	`N` subjects less in agreement with GTTM, differences between `M` and `N` are "not abismal". Observed several conflicts among rules. Suggested the use of extra rules is necessary.
Deliège (1989)	recognition of form in complete pieces	form	2 pieces	P	mid 20^{th} C	≈ 7 mins ≈ 9 mins	`record`	⊘	`EX1`:36 `EX2`:32 `EX3`:24	`DL`,`N`	⊖	⊘	⊘	No difference between `M` subjects and NM subjects. No evidence of invariants. Caesura is a strong cue.
Clarke and Krumhansl (1990)	perception of temporal organization on relatively large pieces	form	2 piano pieces	P	mid 20^{th} C classical	≈ 10 mins (both)	`record`	⊘	`EX1`:23 `EX2`:24	`M` `M`,`DL`	14 `SR`,`PR` 10 `SR`,`PR`	⊘	⊘	Segmentation criteria broadly consistent with GTTM grouping rules.
Deliège et al. (1996)	identify salient elements in a piece (cues)	⊘	1 piece	P	early romantic	16 bars (30 sec)	`record`	⊘	7	`N`	9 `SR`	fair	⊘	Surface cues dominated segmentation of piece. Presence of cue does not warrants salience (context dependent).
Weyde (2003)	assess combination of cues for melodic segmentation	2-3 notes, phrases	`EX1`: 25 artificial sequences `EX2`: 20 artificial sequences	`M`	⊘	`EX1`: 12 notes `EX2`: 6-7 notes	`synth`	✓	`EX1`: ⊘ `EX2`: 6	`EX1`: ⊘ `EX2`: `M`	ex1: 3 `SR` ex2: 7 `SR`	⊘ ⊘	⊘ fair	Linear model inadequate for segment cue combination.
Schaefer et al. (2004)	exceptions to Gestalt principles	phrases	10 children's songs	`M`	folk	30.95 secs average	`synth`	✓	30	`DL`,`M`,`N`	⊘	⊖	⊘	Difference between `DL` and `N` is significant Structural grouping mechanisms are influenced by musical experience. `DL` rely on experience, `N` rely on Gestalts.
Spiro (2006)	study performance-related segmentation cues	phrases	5 pieces (2 performances of each)	`M`,P	baroque classical romantic	excerpts	`record`	⊘	45	`DL`,`M`,`N`	13 `SR` 2 `PR`	⊘	fair	`PR` do affect boundary location perception. Perceived locations where `PR` are active vary across performances.
Weyde et al. (2007)	assess role of pitch intervals for melodic segmentation	2-3 notes	150 artificial sequences	`M`	⊘	12 notes	`synth`	✓	10	`M`	ex1: 4 `SR`	⊘ ⊘	⊘ fair	size of pitch intervals have little influence in segmentation, when compared to timing and dynamics.
Bruderer and Kohlrausch (2009)	characterize temporal boundaries in western pop	phrases sections passages	6 songs	`M`	pop	5 mins average	`synth`	×	21	`M`,`N`	> 50 `SR` > 3 `PR`	fair	low	Salience ratings are consistent, indicated by voting agreement. Correlation to Gestalt based predictions is moderate.

Bibliography

Abdallah, S. and Plumbley, M. (2009). Information dynamics: patterns of expectation and surprise in the perception of music. *Connection Science*, 21(2-3):89–117. 21, 32, 105

Abdallah, S., Sandler, M., Rhodes, C., and Casey, M. (2006). Using duration models to reduce fragmentation in audio segmentation. *Machine Learning*, 65(2-3):485–515. 131

Abeßer, J., Frieler, K., Pfleiderer, M., and Zaddach, W. (2013). Introducing the jazzomat project-jazz solo analysis using music information retrieval methods. In *Proc. of the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 187–192. 135

Ahlbäck, S. (2004). *Melody beyond notes: A study of melody cognition.* PhD thesis, Goteborgs Universitet. 13, 18, 23, 32, 59, 62, 99, 120, 140

Ahlbäck, S. (2007). Melodic similarity as a determinant of melody structure. *Musicae Scientiae*, 11(1):235–280. 20, 24, 32, 59, 61, 62, 73, 131

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723. 84

Ash, R. (1997). Click migration and segmentation in gaelic melodies. In *Proc. of the 3rd Triennial Conference of the European Society for the Cognitive Sciences of Music (ESCOM)*, pages 257–262. 47

Attas, R. (2011). Sarah setting the terms: Defining phrase in popular music. *Music Theory Online*, 17(3). 25

Baars, B. (1988). *A cognitive theory of consciousness*. Cambridge University Press. 116

Baars, B. (1997). *In the theater of consciousness*. Oxford University Press. 116

Baker, M. (1989a). An artificial intelligence approach to musical grouping analysis. *Contemporary Music Review*, 3(1):43–68. 32, 117, 118

Baker, M. (1989b). A computational approach to modeling musical grouping structure. *Contemporary Music Review*, 4(1):311–325. 32, 117

Bartel, C. (2006). Can musical understanding be grounded in the phenomenology of musical experience? Talk summary, in 'Mind, Art and Beauty', University of Leeds. 99

Bartel, C. (2007). *The Perception of Music: An Essay on Musical Understanding, Phenomenology and the Contents of Musical Experience*. PhD thesis, King's College London, UK. 99

Benward, B. and Saker, M. (2008). *Music in Theory and Practice*, volume 1. McGraw-Hill, 8 edition. 25

Berger, A., Pietra, V. D., and Pietra, S. D. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71. 125

Biamonte, N. (2010). Musical representation in the video games guitar hero and rock band. In Biamonte, N., editor, *Pop-Culture Pedagogy in the Music Classroom: Teaching Tools from American Idol to YouTube*, pages 133–147. Scarecrow Press, Plymouth, United Kingdom. 3

Bigo, L. and Conklin, D. (2015). A viewpoint approach to symbolic music transformation. In *Proc. of the 11th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 56–70. 3

Bimbot, F., Le Blouch, O., Sargent, G., and Vincent, E. (2010a). Decomposition into autonomous and comparable blocks: a structural description of music pieces. In *Proc. of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 189–194. 124

Bimbot, F., Le Blouch, O., Sargent, G., and Vincent, E. (2010b). Decomposition into autonomous and comparable blocks: a structural description of music pieces. Technical Report 2102-6327, CNRS, IRISA. 124

Bod, R. (2001). Probabilistic grammars for music. In *Belgian-Dutch Conference on Artificial Intelligence (BNAIC)*. 32, 131

Bod, R. (2002). Memory-based models of melodic analysis: Challenging the gestalt principles. *Journal of New Music Research*, 31(1):27–36. 32, 131

Bozkurt, B., Karaosmanoğlu, K., Karaçalı, B., and Ünal, E. (2014). Usul and makam driven automatic melodic segmentation for turkish music. *Journal of New Music Research*, 43(4):375–389. 32

Brown, A., Gifford, T., and Davidson, R. (2012). Tracking levels of closure in melodies. In *Proc. of the 12th International Conference Music Perception and Cognition (ICMPC)*. 149

Bruderer, M. (2008). *Perception and modeling of segment boundaries in popular music*. PhD thesis, Technische Universiteit Eindhoven. 3, 18, 148

Bruderer, M., McKinney, M., and Kohlrausch, A. (2006). Structural boundary perception in popular music. In *Proc. of the 7th International Conference on Music Information Retrieval (ISMIR)*, pages 198–201. 59

Bruderer, M. nd Mckinney, M. and Kohlrausch, A. (2009). The perception of structural boundaries in melody lines of western popular music. *Musicae Scientiae*, 13(2):273–313. 152, 153

Byrd, D. and Crawford, T. (2002). Problems of music information retrieval in the real world. *Information processing & management*, 38(2):249–272. 2

Cambouropoulos, E. (1996). A general pitch interval representation: Theory and applications. *Journal of New Music Research*, 25(3):231–251. 59

Cambouropoulos, E. (1997a). In *Proc. of the 3rd Triennial Conference of the European Society for the Cognitive Sciences of Music (ESCOM)*, pages 533–538. 59, 60

Cambouropoulos, E. (1997b). Musical rhythm: A formal model for determining local boundaries, accents and metre in a melodic surface. *Music, gestalt, and computing*, pages 277–293. 32

Cambouropoulos, E. (2001). The local boundary detection model (LBDM) and its application in the study of expressive timing. In *Proc. of the International Computer Music Conference (ICMC)*, pages 232–235. 29, 32, 67, 69, 91, 109, 110, 123, 127

Cambouropoulos, E. (2006). Musical parallelism and melodic segmentation. *Music Perception*, 23(3):249–268. 20, 32, 61, 62, 67, 109, 110

Cambouropoulos, E. (2009). How similar is similar? *Musicae Scientiae*, 13(1 suppl):7–24. 60

Cambouropoulos, E., Crawford, T., and Iliopoulos, C. (2001). Pattern processing in melodic sequences: Challenges, caveats and prospects. *Computers and the Humanities*, 35(1):9–21. 60

Cambouropoulos, E. and Tsougras, C. (2004). Influence of musical similarity on melodic segmentation: Representations and algorithms. In *Proc. of the International Conference on Sound and Music Computing (SMC)*. 32, 60

Camilleri, L., Carreras, F., and Duranti, C. (1990). An expert system prototype for the study of musical segmentation. *Journal of New Music Research*, 19(2-3):147–154. 32, 117

Caplin, W. (1998). *Classical form: A theory of formal functions for the instrumental music of Haydn, Mozart, and Beethoven*. Oxford University Press. 2, 4, 24, 177

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., and Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696. 2

Chang, C. and Jiau, H. C. (2004). Representative music fragments extraction by using segmentation techniques. In *Proc. of International Computer Symposium*, pages 1156–1161. 3

Chang, C.-W. and Jiau, H. C. (2011). A numeric indexing and access mechanism for melody retrieval. *International Journal of Innovative Computing Information and Control*, 7:4083–4096. 3

Cheong, S., Stodghill, P., Schneider, D., Cartinhour, S., and Myers, C. (2009). Extending the recursive jensen-shannon segmentation of biological sequences. *arXiv preprint arXiv:0904.2466*. 86

Chew, E. (2006). Slicing it all ways: Mathematical models for tonal induction, approximation, and segmentation using the spiral array. *INFORMS Journal on Computing*, 18(3):305–320. 31, 32, 78

Chomsky, N. (1957). *Syntactic Structures*, volume 119. Mouton. 141

Chomsky, N. (1965). *Aspects of the Theory of Syntax*, volume 119. The MIT press. 141

Christenson, P. and Roberts, D. (1998). It's not only rock & roll: Popular music in the lives of adolescents. *Journal of Communication*, 49(4):122–130. 1

Cienniwa, P. (2014). *By Heart: The Art of Memorizing Music*. Self published. Website: `http://www.paulcienniwa.com/`. Last Accessed March 02, 2016. 2, 177

Claeskens, G. and Hjort, N. (2008). *Model selection and model averaging*. Cambridge University Press. 84

Clarke, E. and Krumhansl, C. (1990). Perceiving musical time. *Music Perception*, pages 213–251. 3, 18, 59, 124, 151, 152, 153

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46. 49

Conklin, D. and Witten, I. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1):51–73. 101, 114

Cook, N. (1994). *A guide to musical analysis*. Oxford University Press. 2, 177

Cooper, M. and Foote, J. (2003). Summarizing popular music via structural similarity analysis. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 127–130. 88, 89

Cope, D. (1992). Computer modeling of musical intelligence in EMI. *Computer Music Journal*, pages 69–83. 3

Copland, A. (1959). *What to Listen for in Music*. McGraw-Hill, New York. revised edition. 23, 24

Cox, G. (2010). On the relationship between entropy and meaning in music: An exploration with recurrent neural networks. In *Proc. of the 32nd Annual Cognitive Science Society*, pages 429–434. 32

Cuddy, L. (1993). Melody comprehension and tonal structure. *Psychology and music: The understanding of melody and rhythm*, pages 19–38. 59

Culpepper, S. (2010). *Musical time and information theory entropy*. PhD thesis, University of Iowa. 85

Cutietta, R. and Booth, G. (1996). The influence of metre, mode, interval type and contour in repeated melodic free-recall. *Psychology of music*, 24(2):222–236. 125

Deliège, I. (1987). Grouping conditions in listening to music: An approach to lerdahl & jackendoff's grouping preference rules. *Music perception*, pages 325–359. 3, 18, 76, 146, 148, 149, 150, 153

Deliège, I. (1989). A perceptual approach to contemporary musical forms. *Contemporary Music Review*, 4(1):213–230. 151, 152, 153

Deliège, I. (2001). Introduction: Similarity perception, categorization, cue abstraction. *Music Perception*, 18(3):233–243. 13, 140

Deliège, I. (2007). Similarity relations in listening to music: How do they come into play? *Musicae Scientiae*, 11(1):9–37. 20, 58

Deliège, I., Mélen, M., Stammers, D., and Cross, I. (1996). Musical schemata in real-time listening to a piece of music. *Music Perception*, pages 117–159. 152, 153

Desain, P., Honing, H., Vanthienen, H., and Windsor, L. (1998). Computational modeling of music cognition: problem or solution? *Music Perception*, pages 151–166. 2

Deutsch, D. (1999). *The psychology of music*, chapter 9: Grouping Mechanisms in Music, pages 299–348. Gulf Professional Publishing, second edition. 146, 148, 149

Dijkstra, E. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271. 130

Dowling, W. and Fujitani, D. (1971). Contour, interval, and pitch recognition in memory for melodies. *The Journal of the Acoustical Society of America*, 49(2B):524–531. 125

Dowling, W., Kwak, S., and Andrews, M. (1995). The time course of recognition of novel melodies. *Perception & Psychophysics*, 57(2):136–149. 125

Downie, J. S. (2003). Music information retrieval. *Annual review of information science and technology*, 37(1):295–340. 2

Downie, S. and Nelson, M. (2000). Evaluation of a simple and effective music information retrieval method. In *Proc. of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 73–80. 3

Dubnov, S. (2006). Analysis of musical structure in audio and midi using information rate. In *Proc. of International Computer Music Conference, ICMC*. 32

Dubnov, S. (2011). *Machine Audition: Principles, Algorithms, and Systems*, chapter 11: Musical Information Dynamics as Models of Auditory Anticipation, pages 371–397. IGI Global. 21

Edelman, G. (1987). *Neural Darwinism: The theory of neuronal group selection.* Basic Books. 116

Edworthy, J. (1985). Interval and contour in melody processing. *Music Perception: An Interdisciplinary Journal*, 2(3):375–388. 125

Eerola, T. and Toiviainen, P. (2004). MIR in Matlab: The MIDI toolbox. In *Proc. of the 5th International Conference on Music Information Retrieval (ISMIR)*. 53

Ehmann, A., Bay, M., Downie, J., Fujinaga, I., and De Roure, D. (2011). Music structure segmentation algorithm evaluation: Expanding on MIREX 2010 analyses and datasets. In *Proc. of the 12th Conference of the International Society for Music Information Retrieval (ISMIR)*, pages 561–566. 36

Ferrand, M., Nelson, P., and Wiggins, G. (2002). A probabilistic model for melody segmentation. In *Proc. of the 2nd International Conference on Music and Artificial Intelligence (ICMAI)*. 85

Ferrand, M., Nelson, P., and Wiggins, G. (2003a). Memory and melodic density: A model for melody segmentation. In *Proc. of the 14th Colloquium on Musical Informatics*, pages 95–98. 32, 76, 80, 88, 89

Ferrand, M., Nelson, P., and Wiggins, G. (2003b). Unsupervised learning of melodic segmentation: A memory-based approach. In *Proc. of the 5th Triennial Conference of the European Society for the Cognitive Sciences of Music (ESCOM)*, pages 141–144. 32

Fournier, C. (2013a). Evaluating text segmentation. Master's thesis, University of Ottawa. 43

Fournier, C. (2013b). Evaluating text segmentation using boundary edit distance. In *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1702–1712. 42

Fraisse, P. (1982). Rhythm and tempo. *The psychology of music*, 1:149–180. 47, 124

Frankland, B., McAdams, S., and Cohen, A. (2004). Parsing of melody: Quantification and testing of the local grouping rules of Lerdahl and Jackendoff's A Generative Theory of Tonal Music. *Music Perception*, 21(4):499–543. 32, 142

Friberg, A., Bresin, R., Frydén, L., and Sundberg, J. (1998). Musical punctuation on the microlevel: Automatic identification and performance of small melodic units. *Journal of New Music Research*, 27(3):271–292. 32

Frieler, K., Zaddach, W., and Abeßer, J. (2014). Exploring phrase form structures. part ii: Monophonic jazz solos. In *Proc. of the 4th Folk Music Analysis Workshop (FMA)*, pages 48–51. 47

Frith, S. (1998). *Performing rites: On the value of popular music*. Harvard University Press. 1

Futrelle, J. and Downie, J. S. (2003). Interdisciplinary research issues in music information retrieval: Ismir 2000–2002. *Journal of New Music Research*, 32(2):121–131. 2

Gardner, H. (2008). *Extraordinary minds: Portraits of 4 exceptional individuals and an examination of our own extraordinariness.* Basic Books. 15

Gobet, F., Lane, P., Croker, S., Cheng, P., Jones, G., Oliver, I., and Pine, J. (2001). Chunking mechanisms in human learning. *Trends in cognitive sciences*, 5(6):236–243. 17

Godøy, R. (2009). Chunking sound for musical analysis. In *Computer Music Modeling and Retrieval. Genesis of Meaning in Sound and Music*, pages 67–80. 140, 144

Godøy, R. (2014). Ecological constraints of timescales, production, and perception in temporal experiences of music: A commentary on Kon (2014). *Empirical Musicology Review*, 9(3-4). 124, 144

Gohlke, K., Hlatky, M., Heise, S., Black, D., and Loviscach, J. (2010). Track displays in DAW software: Beyond waveform views. In *Audio Engineering Society Convention 128*. 3

Goldwater, S., Griffiths, T., and Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54. 135

Green, A. (1997). Kappa statistics for multiple raters using categorical classifications. In *Proc. of the 22nd annual SAS User Group International conference*, pages 1110–1115. 49

Greene, R. (1986). Sources of recency effects in free recall. *Psychological Bulletin*, 99(2):221. 66

Grill, T. and J., S. (2015). Music boundary detection using neural networks on combined features and two-level annotations. In *Proc. of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, pages 531–537. 135

Grosse, I., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., Oliver, J., and Stanley, E. (2002). Analysis of symbolic sequences using the jensen-shannon divergence. *Physical Review E*, 65(4):041905. 84, 86

Hamanaka, M., Hirata, K., and Tojo, S. (2004). Automatic generation of grouping structure based on the GTTM. In *Proc. of the International Computer Music Conference (ICMC)*, pages 141–144. 32

Hamanaka, M., Hirata, K., and Tojo, S. (2005). ATTA: Automatic time-span tree analyzer based on extended GTTM. In *Proc. of the 6th International Conference on Music Information Retrieval (ISMIR)*, pages 358–365. 32

Hamanaka, M., Hirata, K., and Tojo, S. (2006). Implementing a generative theory of tonal music. *Journal of New Music Research*, 35(4):249–277. 32, 117

Handel, S. (1973). Temporal segmentation of repeating auditory patterns. *Journal of Experimental Psychology*, 101(1):46. 149

Hanninen, D. (2001). Orientations, criteria, segments: A general theory of segmentation for music analysis. *Journal of Music Theory*, pages 345–433. 2, 13, 140, 177

Harford, S. (2003). Automatic segmentation, learning and retrieval of melodies using a self-organizing neural network. In *Proceedings of International Conference on Music Information Retrieval, MD, Baltimore*. 32

Harford, S. (2006). *Content-Based Retrieval of Melodies using Artificial Neural Networks*. PhD thesis, School of Computing, Dublin City University. 32

Harris, M., Smaill, A., and Wiggins, G. (1991). Representing music symbolically. In *Proc. of the IX Colloquio di Informatica Musicale*, pages 55–69. 4

Harrison, L. and Loui, P. (2014). Thrills, chills, frissons, and skin orgasms: toward an integrative model of transcendent psychophysiological experiences in music. *Frontiers in Psychology*, 5(790):1–6. 1

Hébert, S. and Peretz, I. (1997). Recognition of music in long-term memory: Are melodic and temporal patterns equal partners? *Memory & Cognition*, 25(4):518–533. 101

Herrera, P., Serrà, J., Laurier, C., Guaus, E., Gómez, E., and Serra, X. (2009). The discipline formerly known as MIR. In *Proc. of the 10th Conference of the International Society for Music Information Retrieval (ISMIR), special session on The Future of MIR*. 2

Hewlett, W. and Selfridge-Field, E. (1998). *Melodic similarity: Concepts, procedures, and applications*, volume 11. The MIT Press. 20

Honing, H. (2006). Computational modeling of music cognition: A case study on model selection. *Music Perception*, 23:365–376. 34

Huron, D. (1996). The melodic arch in western folksongs. *Computing in Musicology*, 10:3–23. 48, 49

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. MIT press. 21, 67

ISO/IEC (2001). ISO standard 9126: Software engineering – product quality, part 1. Geneva, International Organization for Standardization / International Electrotechnical Commission. 35

ISO/IEC (2003). ISO standard 9126: Software engineering – product quality, parts 2 and 3. Geneva, International Organization for Standardization / International Electrotechnical Commission. 35

Jackson, J. (1987). Idea for a mind. *ACM SIGART Bulletin*, (101):23–26. 116

Janssen, B., de Haas, W., Volk, A., and van Kranenburg, P. (2013). Discovering repeated patterns in music: state of knowledge, challenges, perspectives. In *Proc. of the International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 225–240. 60

Járdányi, P. (1965). Experiences and results in systematizing hungarian folk-songs. *Studia Musicologica*, pages 287–291. 108

Jensen, K. and Hebert, D. (2015). Predictability of harmonic complexity across 75 years of popular music hits. In *Proc. of the 11th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 198–212. 3, 85

Juhász, Z. (2004). Segmentation of hungarian folk songs using an entropy-based learning system. *Journal of New Music Research*, 33(1):5–15. 32

Kaiser, F. and Peeters, G. (2013). Multiple hypotheses at multiple scales for audio novelty computation within music. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 231–235. 75

Karaosmanoglu, M., Bozkurt, B., Holzapfel, A., and Disiacik, N. (2014). A symbolic dataset of turkish makam music phrases. In *Proc. of the 4th Folk Music Analysis Workshop (FMA)*, pages 10–14. 135

Katz, R. (1981). On some criteria for estimating the order of a markov chain. *Technometrics*, 23(3):243–249. 84

Kirke, A., Dixon, B., and Miranda, E. (2015). Music and dementia: Two case-studies. In *Proc. of the 11th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 234–246. 1

Klaus, K. (1980). Content analysis: An introduction to its methodology. 49

Kodály, Z. and Vargyas, L. (1982). *Folk music of Hungary*. Da Capo Press. 108

Konishi, S. and Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer Science & Business Media. 84

Kuncheva, L. (2002). A theoretical study on six classifier fusion strategies. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (2):281–286. 85

La Rue, J. (1970). *Guidelines for style analysis: A comprehensive outline of basic principles for the analysis of musical style*. New York: W. W. Norton. 2, 4, 177

Lalitte, P., Bigand, E., Poulin-Charronnat, B., McAdams, S., Delbé, C., and D'Adamo, D. (2004). The perceptual structure of thematic materials in the angel of death. *Music Perception*, 22(2):265–296. 60

Lamont, A. and Dibben, N. (2001). Motivic structure and the perception of similarity. *Music Perception*, 18(3):245–274. 60

Large, E., Palmer, C., and Pollack, J. (1995). Reduced memory representations for music. *Cognitive Science*, 19(1). 32

Lartillot, O. (2007). Motivic pattern extraction in symbolic domain. *Intelligent music information systems: Tools and methodologies*, pages 236–260. 56, 60

Lartillot, O. (2010). Reflections towards a generative theory of musical parallelism. *Musicae Scientiae (Discussion Forum)*, 5:195–229. 20

Lartillot, O. and Ayari, M. (2014). A comprehensive computational model for music analysis, applied to maqâm analysis. In *Proc. of the 3rd International Workshop of Folk Music Analysis (FMA)*, pages 78–84. 30, 32

Lartillot, O., Cereghetti, D., Eliard, K., and Grandjean, D. (2013). A simple, high-yield method for assessing structural novelity. In *Proc. of the 3rd International Conference on Music & Emotion (ICME)*, pages 277–285. 75

Lattner, S., Chacón, C., and Grachten, M. (2015a). Pseudo-supervised training improves unsupervised melody segmentation. In *Proc. of the 24th International Conference on Artificial Intelligence*, pages 2459–2465. 32

Lattner, S., Grachten, M., Agres, K., and Chacón, C. (2015b). Probabilistic segmentation of musical sequences using restricted boltzmann machines. In *Mathematics and Computation in Music*, pages 323–334. Springer. 32

Lee, K. and Cremer, M. (2008). Segmentation-based lyrics-audio alignment using dynamic programming. In *Proc. of the 9th International Conference on Music Information Retrieval (ISMIR)*, pages 395–400. 3

Lefkowitz, D. and Taavola, K. (2000). Segmentation in music: generalizing a piece-sensitive approach. *Journal of Music Theory*, 44(1):171–229. 32

Lerdahl, F. and Jackendoff, R. (1983). *A generative theory of tonal music*. MIT press. 2, 13, 20, 23, 99, 118, 140, 142, 143, 177

Lesaffre, M., Leman, M., Tanghe, K., De Baets, B., De Meyer, H., and Martens, J. (2003). User-dependent taxonomy of musical features as a conceptual framework for musical audio-mining technology. In *Proc. of the Stockholm Music Acoustics Conference*, pages 635–638. 143

Levy, M. and Sandler, M. (2008). Structural segmentation of musical audio by constrained clustering. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):318–326. 120

Li, B., Burgoyne, J., and Fujinaga, I. (2006). Extending audacity for audio annotation. In *Proc. of the 7th Conference of the International Society for Music Information Retrieval (ISMIR)*, pages 379–380. 45

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151. 83, 84

Loeckx, J. (2015). Musical variation and improvisation based on multi-resolution representations. In *Proc. of the 11th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 87–96. 3

Love, S. (2011). *On phrase rhythm in jazz*. PhD thesis, University of Rochester. 22, 23

Macpherson, S. (1915). *Form in music*. London: Joseph Williams, ltd. 23, 25

Madsen, S. T., Typke, R., and Widmer, G. (2008). Automatic reduction of midi files preserving relevant musical content. In *Adaptive Multimedia Retrieval. Identifying, Summarizing, and Recommending Image and Music*, pages 89–99. Springer. 85

Manning, C., Raghavan, P., and Schütze, H. (2008). Scoring, term weighting and the vector space model. *Introduction to Information Retrieval*, 100. 109

Margulis, E. (2012). Musical repetition detection across multiple exposures. *Music Perception: An Interdisciplinary Journal*, 29(4):377–385. 59, 67, 146

Margulis, E. (2014). *On repeat: how music plays the mind*. Oxford University Press. 20

Margulis, E. and Beatty, A. (2008). Musical style, psychoaesthetics, and prospects for entropy as an analytic tool. *Computer Music Journal*, 32(4):64–78. 85

Mauch, M., MacCallum, R., Levy, M., and Leroi, A. (2015). The evolution of popular music: USA 1960–2010. *Royal Society Open Science*, 2(5). 3

McAdams, S., Vieillard, S., Houix, O., and Reynolds, R. (2004). Perception of musical similarity among contemporary thematic materials in two instrumentations. *Music Perception*, 22(2):207–237. 60

McFee, B. (2015). The role of structure analysis in music discovery. Talk summary, Machine Learning for Music Discovery Workshop. Website: `https://sites.google.com/site/ml4md2015/program`. Last Accessed October 02, 2015. 3

McFee, B., Nieto, O., and Bello, J. (2015). Hierarchical evaluation of segment boundary detection. pages 406–412. 36, 124, 136

Medin, D. and Schaffer, M. (1978). Context theory of classification learning. *Psychological review*, 85(3):207–238. 98

Melucci, M. and Orio, N. (2002). A comparison of manual and automatic melody segmentation. In *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, pages 7–14. 41

Melucci, M., Orio, N., and Gambalunga, M. (2000). An evaluation study on music perception for musical content-based information retrieval. In *Proc. of the International Computer Music Conference (ICMC)*, pages 162–165. 124

Meredith, D., Lemström, K., and Wiggins, G. (2001). A geometric approach to computing repeated patterns in polyphonic music. Document submitted to UK Patent office, application number GB 0200203.8. Website: `http://www.chromamorph.net/papers/public/siajnmr_submit_2.pdf`. Last Accessed August 21, 2015. 56

Meredith, D., Lemström, K., and Wiggins, G. (2002). Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research*, 31(4):321–345. 60

Meyer, L. (1956). *Emotion and meaning in music*. University of Chicago Press. 21, 141

Meyer, L. (1957). Meaning in music and information theory. *Journal of Aesthetics and Art Criticism*, pages 412–424. 102

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81. 124

Minsky, M. (1985). *The society of mind*. Simon and Schuster. 116

MMA (1995). 1.0 detailed specification v96.1. MIDI Manufacturers Association (MMA). 137

Müllensiefen, D. (2009). FANTASTIC: Feature ANalysis Technology Accessing STatistics (in a corpus): Technical report v1.5. Technical report. 124

Müller, M. (2011). New developments in music information retrieval. In *Audio Engineering Society Conference: 42nd International Conference: Semantic Audio*. Audio Engineering Society. 2

Müller, M. and Clausen, M. (2007). Transposition-invariant self-similarity matrices. In *Proc. of the 8th International Conference on Music Information Retrieval (ISMIR)*, pages 47–50. 64

Müller, M. and Grosche, P. (2012). Automated segmentation of folk song field recordings. In *Proc. of the ITG Conference on Speech Communication*, pages 1–4. 55, 57

Müller, M., Grosche, P., and Jiang, N. (2011). A segment-based fitness measure for capturing repetitive structures of music recordings. In *Proc. of the 12th Conference of the International Society for Music Information Retrieval (ISMIR)*, pages 615–620. 73

Müller, M., Jiang, N., and Grohganz, H. (2014). SM Toolbox: Matlab implementations for computing and enhancing similarity matrices. In *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*, pages 222–231. Audio Engineering Society. 68

Müller, M., Jiang, N., and Grosche, P. (2013). A robust fitness measure for capturing repetitions in music recordings with applications to audio thumbnailing. *IEEE Transactions on audio, speech, and language processing*, 21(3):531–543. 63, 65, 73

Müller, M. and Kurth, F. (2006). Enhancing similarity matrices for music audio analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages 9–12. 64

Murdock, B. (1962). The serial position effect of free recall. *Journal of experimental psychology*, 64(5):482. 66

Narmour, E. (1990). *The Analysis and Cognition of Basic Melodic Structures: The Implication–realisation Model*. University of Chicago Press. 2, 13, 21, 99, 140, 142, 177

Narmour, E. (1992). *The Analysis and Cognition of Melodic Complexity: The Implication–realisation Model*. University of Chicago Press. 2, 13, 99, 140, 177

Ockelford, A. (2004). On similarity, derivation and the cognition of musical structure. *Psychology of Music*, 32(1):23–74. 13, 140

Opdyke, J. (2010). A unified approach to algorithms generating unrestricted and restricted integer compositions and integer partitions. *Journal of Mathematical Modelling and Algorithms*, 9(1):53–97. 101

Orio, N. (2006). *Music retrieval: A tutorial and review*. Now Publishers Inc. 2

Orio, N. and Neve, G. (2005). Experiments on segmentation techniques for music documents indexing. In *Proc. of the 6th International Conference on Music Information Retrieval (ISMIR)*, pages 104–107. 3, 109

Ornstein, R. (1986). *Multimind*. Houghton Mifflin. 116

Palmer, C. and Krumhansl, C. (1987a). Independent temporal and pitch structures in determination of musical phrases. *Journal of Experimental Psychology: Human Perception and Performance*, 13(1):116. 23

Palmer, C. and Krumhansl, C. (1987b). Pitch and temporal contributions to musical phrase perception: Effects of harmony, performance timing, and familiarity. *Perception & Psychophysics*, 41(6):505–518. 23

Paulus, J., Müller, M., and Klapuri, A. (2010). Audio-based music structure analysis. In *Proc. of the 11th Conference of the International Society for Music Information Retrieval (ISMIR)*, pages 625–636. 30, 31, 36, 78

Pearce, M. (2005). *The construction and evaluation of statistical models of melodic structure in music perception and composition*. PhD thesis, City University. 101, 114, 143

Pearce, M., Müllensiefen, D., and Wiggins, G. (2008). A comparison of statistical and rule-based models of melodic segmentation. In *Proc. of the 9th International Conference on Music Information Retrieval (ISMIR)*, pages 89–94. 29, 32

Pearce, M., Müllensiefen, D., and Wiggins, G. (2010a). Melodic grouping in music information retrieval: New methods and applications. *Advances in music information retrieval*, pages 364–388. 36, 40, 42, 120, 148

Pearce, M., Müllensiefen, D., and Wiggins, G. (2010b). The role of expectation and probabilistic learning in auditory boundary perception: a model comparison. *Perception*, 39(10):1365. 3, 36, 89, 120, 140, 142

Pearce, M. and Rohrmeier, M. (2012). Music cognition and the cognitive sciences. *Topics in cognitive science*, 4(4):468–484. 3

Pearce, M. and Wiggins, G. (2004). Improved methods for statistical modelling of monophonic music. *Journal of New Music Research*, 33(4):367–385. 108

Pearce, M. and Wiggins, G. (2006a). Expectation in melody: The influence of context and learning. *Music Perception*, 23(5):377–405. 21

Pearce, M. and Wiggins, G. (2006b). The information dynamics of melodic boundary detection. In *Proc. of the 9th International Conference on Music Perception and Cognition (ICMPC)*, pages 860–865. 21, 32, 102, 141

Pearl, L., Goldwater, S., and Steyvers, M. (2010). Online learning mechanisms for bayesian models of word segmentation. *Research on Language and Computation*, 8(2-3):107–132. 135

Potter, K., Wiggins, G., and Pearce, M. (2007). Towards greater objectivity in music theory: Information-dynamic analysis of minimalist music. *Musicae Scientiae*, 11(2):295–324. 32

Rafael, B. and Oertl, S. (2010). MTSSM - a famework for multi-track segmentation of symbolic music. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 4(1):7–13. 27, 32, 60, 61, 62

Rafael, B., Oertl, S., Affenzeller, M., and Wagner, S. (2009). Using heuristic optimization for segmentation of symbolic music. *Computer Aided Systems Theory - EUROCAST 2009*, pages 641–648. 32, 61, 62

Rhodes, C., Casey, M., Abdallah, S., Sandler, M., et al. (2006). A markov-chain monte-carlo approach to musical audio segmentation. In *Proc. of the 3rd International Workshop on Speech abd Signal Processing (ICASSP)*, pages 797–800. 131

Roach, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology*, 104:192–233. 98

Roads, C. (2001). *Microsound*. MIT Press. 25

Roberts, P. (2001). *Images: The Piano Music of Claude Debussy*. Hal Leonard Corporation. 2, 177

Rodríguez-López, M., Bountouridis, D., and Volk, A. (2014a). Multi-strategy segmentation of melodies. In *Proc. of the 15th Conference of the International Society for Music Information Retrieval (ISMIR)*, pages 207–212. 115

Rodríguez-López, M., Bountouridis, D., and Volk, A. (2015). A novel music segmentation interface and the jazz tune collection. In *Proc. of the 5th Folk Music Analysis Workshop (FMA)*, pages 79–85. 33, 44, 135

Rodríguez-López, M. and Volk, A. (2012). Automatic segmentation of symbolic music encodings: A survey. Technical Report UU-CS-2012-015, Department of Information and Computing Sciences, Utrecht University. 5, 11, 12, 31, 78, 118

Rodríguez-López, M. and Volk, A. (2012). Melodic segmentation using the jensen-shannon divergence. In *Proc. of the 11th International Conference on Machine Learning and Applications (ICMLA)*, volume 2, pages 351–356. 74, 76, 80, 88, 89

Rodríguez-López, M. and Volk, A. (2013). Symbolic segmentation: A corpus-based analysis of melodic phrases. In *Proc. of the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 381–388. 33

Rodríguez-López, M. and Volk, A. (2015a). Location constraints for repetition-based segmentation of melodies. In *Proc. of the 5th International Conference on Mathematics and Computation in Music (MCM)*, pages 73–84. 55

Rodríguez-López, M. and Volk, A. (2015b). On the evaluation of automatic segment boundary detection. In *Proc. of the 11th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 234–246. 33, 136

Rodríguez-López, M. and Volk, A. (2015c). Selective acquisition techniques for enculturation-based melodic phrase segmentation. In *Proc. of the 16th Conference of the International Society for Music Information Retrieval (ISMIR)*, pages 218–224. 96

Rodríguez-López, M., Volk, A., and de Haas, W. (2014b). Comparing repetition-based melody segmentation models. In *Proc. of the 9th Conference on Interdisciplinary Musicology (CIM)*, pages 143–148. 55, 56, 62, 66, 69, 89

Rothstein, W. (1989). *Phrase rhythm in tonal music*. Schirmer Books New York. 4, 25

Rowe, R. (1992). Machine listening and composing with Cypher. *Computer Music Journal*, pages 43–63. 3, 31, 32, 117

Rubin, S., Berthouzoz, F., Mysore, G., Li, W., and Agrawala, M. (2013). Content-based tools for editing audio stories. In *Proc. of the 26th annual ACM symposium on user interface software and technology*, pages 113–122. ACM. 3

Rumelhart, D. (1980). Schemata: The building blocks of cognition. In *Theoretical Issues in Reading Comprehension*, pages 33–58. 98

Rusbridger, A. (2013). *Play it Again: An amateur against the impossible*. Macmillan. 2, 177

Salamon, J. (2013). *Melody Extraction from Polyphonic Music Signals*. PhD thesis, Universitat Pompeu Fabra. 23

Sankalp, G., Serrà, J., and Serra, X. (2015). Improving melodic similarity in indian art music using culture-specific melodic characteristics. In *Proc. of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, pages 680–686. 114

Sankalp, G., Serrà, J., Vignesh, I., Sertan, Ş., and Serra, X. (2016). Phrase-based rāga recognition using vector space modeling. In *Proc. of the 41st International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. In Press. 3

Sargent, G., Bimbot, F., and Vincent, E. (2011). A regularity-constrained viterbi algorithm and its application to the structural segmentation of songs. In *Proc. of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 483–488. 120, 122

Schaefer, R., Murre, J., and Bod, R. (2004). Limits to universality in segmentation of simple melodies. In *Proc. of the 8th international conference on music perception and cognition (ICMPC)*, pages 1–4. 148, 149, 152, 153

Schoenberg, A. (1967). *Fundamentals of musical composition*. Faber & Faber. Edited by: Strang, G. and Stein, L. 4

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464. 84

Serrà, J., Corral, Á., Boguñá, M., Haro, M., and Arcos, J. (2012). Measuring the evolution of contemporary western popular music. *Scientific reports*, 2. 3, 88, 89

Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., Gómez, E., Gouyon, F., Herrera, P., Jorda, S., Paytuvi, O., Peeters, G., Schlüter, J., Vinet, H., and Widmer, G. (2013). *Roadmap for Music Information ReSearch*. ISBN: 978-2-9540351-1-6. Creative Commons BY-NC-ND 3.0 license. 2

Shanahan, D. and Huron, D. (2011). Interval size and phrase position: A comparison between german and chinese folksongs. *Empirical Musicology Review*, 6:187–197. 52

Sleator, D. and Temperley, D. (2001). The melisma music analyzer. *Available online at* `www.link.cs.cmu.edu/music-analysis`. 53

Smalley, D. (1997). Spectromorphology: explaining sound-shapes. *Organised sound*, 2(02):107–126. 25

Smith, J. (2010). A comparison and evaluation of approaches to the automatic formal analysis of musical audio. Master's thesis, McGill University. 79

Smith, J. and Chew, E. (2013a). A meta-analysis of the MIREX structure segmentation task. In *Proc. of the 14th Conference of the International Society for Music Information Retrieval (ISMIR)*, pages 251–256. 36, 42, 136

Smith, J. and Chew, E. (2013b). Using quadratic programming to estimate feature relevance in structural analyses of music. In *Proc. of the 21st ACM international conference on Multimedia*, pages 113–122. 37, 78, 136

Smith, J. and Chew, E. (2015). Validating an optimisation technique for estimating the focus of a listener. Poster Session at the Mathemusical Conversations workshop, Singapore. 78

Smith, J., Chuan, C., and Chew, E. (2013). Audio properties of perceived boundaries in music. *IEEE Transactions on Multimedia*. 79, 136

Smith, J., Schankler, I., and Chew, E. (2014). Listening as a creative act: Meaningful differences in structural annotations of improvised performances. *Music Theory Online*, 20(3). 78

Snyder, B. (2000). *Music and memory: an introduction*. MIT press. 23, 26

Snyder, J. (1990). Entropy as a measure of musical style: the influence of a priori assumptions. *Music Theory Spectrum*, 12(1):121–160. 85

Spevak, C., Thom, B., and Höthker, K. (2002). Evaluating melodic segmentation. In *Proc. of the 2nd International Conference on Music and Artificial Intelligence (ICMAI)*, pages 168–182. 41, 124

Spiro, N. (2003). Various meanings of the term 'musical phrase'. In *Proc. of 5th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM)*, pages 674–677. 25

Spiro, N. (2006). Footprints of musical phrase structure in listeners responses. In *Proc. Of the 9th International Conference on Music Perception and Cognition (ICMPC)*, pages 1176–1183. 146, 148, 152, 153

Spiro, N. (2007). *What contributes to the perception of musical phrases in western classical music?* PhD thesis, Universiteit van Amsterdam. 18, 59, 76, 77

Spiro, N. and Klebanov, B. (2006). A new method for assessing consistency of real-time identification of phrase-parts and its initial application. In *Proc. of the 9th International Conference of Music Perception and Cognition (ESCOM)*, pages 793–800. 147, 148

Sridhar, R., Amudha, A., and Karthiga, S. (2010). Comparison of modified dual ternary indexing and multi-key hashing algorithms for music information retrieval. *International Journal of Artificial Intelligence & Applications*, 1(3):59–69. 3

Stein, L. (1979). *Structure & style*. Miami: Summy-Birchard Music Company. 4, 24, 25, 75

Stoffer, T. (1985). Representation of phrase structure in the perception of music. *Music Perception*, pages 191–220. 23

Sturm, B. (2014). The state of the art ten years after a state of the art: Future research in music information retrieval. *Journal of New Music Research*, 43(2):147–172. 2

Su, M., Yang, Y., Lin, Y., and Chen, H. (2009). An integrated approach to music boundary detection. In *Proc. of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pages 705–710. 120

Swaminathan, K. and Doddihal, V. (2007). Audio segmentation assisted synchronized lyrics editing for ce devices. In *Consumer Electronics, 2007. ICCE 2007. Digest of Technical Papers. International Conference on*, pages 1–2. IEEE. 3

Sylvan, R. (2002). *Traces of the spirit: The religious dimensions of popular music*. NYU Press. 1

Takasu, A., Yanase, T., Kanazawa, T., and Adachi, J. (1999). Music structure analysis and its application to theme phrase extraction. *Research and Advanced Technology for Digital Libraries*, pages 854–854. 32, 61, 67

Tax, D., Van Breukelen, M., Duin, R., and Kittler, J. (2000). Combining multiple classifiers by averaging or by multiplying? *Pattern recognition*, 33(9):1475–1485. 85

Temperley, D. (2001). *The cognition of basic musical structures*. MIT press. 30, 32, 59, 66, 67, 76, 117, 118, 124

Temperley, D. (2003). End-accented phrases: An analytical exploration. *Journal of Music Theory*, 47(1):125–154. 23, 25

Tenney, J. and Polansky, L. (1980). Temporal gestalt perception in music. *Journal of Music Theory*, pages 205–241. 32

Thakur, V., Azad, R., and Ramaswamy, R. (2007). Markov models of genome segmentation. *Physical Review E*, 75(1):011915. 83, 86

Thom, B., Spevak, C., and Höthker, K. (2002). Melodic segmentation: Evaluating the performance of algorithms and musical experts. In *Proc. of the International Computer Music Conference (ICMC)*, pages 65–72. 35, 51, 120, 148

Thornton, C. (2011). Generation of folk song melodies using bayes transforms. *Journal of New Music Research*, 40(4):293–312. 30, 32

Toiviainen, P. (2007). Similarity perception in listening to music. *Musicae Scientiae*, 11. Special Issue: Discussion Forum 4A. 20

Toiviainen, P. (2009). Musical similarity. *Musicae Scientiae*, 13. Special Issue: Discussion Forum 4B. 20

Turnbull, D., Lanckriet, G., Pampalk, E., and Goto, M. (2007). A supervised approach for detecting boundaries in music using difference features and boosting. In *Proc. of the 8th International Society for Music Information Retrieval Conference (ISMIR)*, pages 51–54. 80

Ullrich, K., Schlüter, J., and Grill, T. (2014). Boundary detection in music structure analysis using convolutional neural networks. In *Proc. of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, pages 417–422. 80, 135

van Balen, J., Burgoyne, J., Bountouridis, D., Müllensiefen, D., and Veltkamp, R. (2015). Corpus analysis tools fro computational hook discovery. In *Proc. of the 16th Conference of the International Society for Music Information Retrieval (ISMIR)*. 85

van Kranenburg, P., de Bruin, M., Grijp, L., and Wiering, F. (2014). The meertens tune collections. 108, 135

Velarde, G., Weyde, T., and Meredith, D. (2013). An approach to melodic segmentation and classification based on filtering with the haar-wavelet. *Journal of New Music Research*, 42(4):325–345. 32, 80, 88

Volk, A., Chew, E., Margulis, E., and Anagnostopoulou, C. (2015). Music similarity: Concepts, cognition, and computation. Workshop. Website: `http://www.lorentzcenter.nl/lc/web/2015/669/info.php3?wsid=669`. Last Accessed August 21, 2015. 20

Walker, R. (1997). Visual metaphors as music notations for sung vowel spectra in different cultures. *Journal of New Music Research*, 26(4):315–345. 144

Weyde, T. (2001). Grouping, smilarity and the recognition of rhythmic structure. *Proc. of the International Computer Music Conference (ICMC)*. 32

Weyde, T. (2002). Integrating segmentation and similarity in melodic analysis. In *Proc. of the 7th International Conference on Music Perception and Cognition (ICMPC)*, pages 240–243. 32

Weyde, T. (2003). Optimising parameter weights in models for melodic segmentation. In *Proc. of 7th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM)*, pages 130–133. 146, 149, 153

Weyde, T., Wissmann, J., and Neubarth, K. (2007). An experiment on the role of pitch intervals in melodic segmentation. In *Proc. of the 8th International Conference on Music Information Retrieval (ISMIR)*, pages 287–288. 146, 149, 153

White, B. (1960). Recognition of distorted melodies. *The American journal of psychology*, 73(1):100–107. 125

Wiering, F. (2006). Can humans benefit from music information retrieval? In *Adaptive Multimedia Retrieval: User, Context, and Feedback*, pages 82–94. Springer. 2

Wiering, F., de Nooijer, J., Volk, A., and Tabachneck-Schijf, H. (2009). Cognition-based segmentation for music information retrieval systems. *Journal of New Music Research*, 38(2):139–154. 36, 120, 148

Wiering, F. and Veltkamp, R. (2005). What is topical in cultural heritage: Content-based retrieval among folksong tunes (witchcraft). Project Description Report. Website: `http://www.cs.uu.nl/research/projects/witchcraft/projectDescriptions/witchcraftLong.pdf`. Last Accessed August 21, 2015. 1

Wiggins, G. and Forth, J. (2015). IDyOT: A computational theory of creativity as everyday reasoning from learned information. In *Computational Creativity Research: Towards Creative Machines*, pages 127–148. 4, 13, 21, 140, 144

Wilder, G. (2008). Adaptive melodic segmentation and motivic identification. In *Proc. of the International Computer Music Conference (ICMC)*. 32

Woelfer, J. and Lee, J. (2012). The role of music in the lives of homeless young people: A preliminary report. In *Proc. of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, pages 367–372. 1

Wołkowicz, J. (2013). *Application of Text-Based Methods of Analysis to Symbolic Music*. PhD thesis, Faculty of Computing Sciences, Dalhousie University. 32, 60, 61, 62

Wołkowicz, J., Kulka, Z., and Kešelj, V. (2008). N-gram-based approach to composer recognition. *Archives of Acoustics*, 33(1):43–55. 102

Yen, J. (1971). Finding the k shortest loopless paths in a network. *management Science*, 17(11):712–716. 130

Zanette, D. (2007). Segmentation and context of literary and musical sequences. *Complex Systems*, 17:279–293. 31, 32, 78

Summary

The work presented in this dissertation investigates *music segmentation*. In the field of Musicology, segmentation refers to a score analysis technique, whereby notated pieces or passages of these pieces are divided into 'units' referred to as sections, periods, phrases, and so on. Segmentation analysis is a widespread practice among musicians: performers use it to help them memorise pieces (Rusbridger 2013; Cienniwa 2014), music theorists and historians use it to compare works (Caplin 1998; Roberts 2001), music students use it to understand the compositional strategies of a given composer or genre (La Rue 1970; Cook 1994). In the field of Music Psychology it is posited that a similar type of analysis is performed by our auditory system when constructing mental representations of music. In fact, most theories consider segmentation to be a core listening mechanism, fundamental to the way humans recognise, categorise, and memorise music (Lerdahl and Jackendoff 1983; Narmour 1992, 1990; Hanninen 2001).

Digital music files often lack a segmentation analysis. Automatising segmentation has the potential of improving (or even enabling) situations where computers are used to search, browse, visualise, or summarise digital music collections. Moreover, investigation and modelling segmentation could lead to a better understanding of how music is perceived by humans.

In this dissertation we investigate segmentation via computer simulation. We focus on the analysis of melody. We provide a conceptual model of melodic segmentation, and use it to introduce and test four different segmenters. The results obtained by our automated segmenters extends previous research in the area, and allows us to foresee a not-so-distant future where manual and automatic segmentations will be indistinguishable.

Acknowledgements

I am using these last page to thank a number of people that directly or indirectly contributed to the completion of this dissertation.

First of all I would like to thank my supervisors Anja Volk and Remco Veltkamp. Both were incredibly helpful, supportive, and patient. And it is thanks to Anja's hard work that the MUSIVA project (which employed me as doctoral researcher) came to existence in the first place. Thanks also go to the music research group at UU: Frans Wiering, W. Bas de Haas, Jan Van Balen, Dimitrios Bountouridis, Anna Aljanaki, and Hendrik Vincent Koops. You are all knowledgeable, hard working, and kind people. The best team one could wish to work with. Lastly, thanks go to my parents and sisters, and to my love Virginia. Without your emotional support this dissertation would perhaps never have gotten off the ground.