Automatic Segmentation of Symbolic Music Encodings: A Survey

Marcelo Rodríguez López
Anja Volk

Technical Report UU-CS-2012-015
September 2012

Department of Information and Computing Sciences
Utrecht University, Utrecht, The Netherlands
www.cs.uu.nl

ISSN: 0924-3275

Department of Information and Computing Sciences
Utrecht University
P.O. Box 80.089
3508 TB Utrecht
The Netherlands

Automatic Segmentation of Symbolic Music Encodings: A Survey

Marcelo Rodríguez López and Anja Volk

1 Introduction

A challenge crossing diverse scientific domains is the segmentation of sequential data into sub-sequences that are significant for the analysis of the domain at hand. In this paper we survey computer models that aim to mimic music segmentation by human listeners. That is, models developed to segment sequential encodings of musical data into sub-sequences that humans would agree aid the cognition of the music the data describes. In other words, sub-sequences that are 'cognitively plausible'.

1.1 Scope

We restrict the scope of our survey to automatic segmentation of *symbolic* (score-like) representations of *monophonic* music. We focus on automatic *boundary detection* of segments resembling the musicological concepts of *phrase* and *section*. That is, computational models aiming to detect the time points separating contiguous phrases/sections.

1.2 Application Domains

Segmentation of music data is important for many areas concerned with automated music processing, such as:

- *Music Information Retrieval*, where automatically identified segments can be used to index large digital music collections [37], visualise and classify music [45, 62], and aid music summarisation [28].
- *Computational Musicology*, where automatically detected segments can be used to aid or complement theoretical analyses of music.
- *Audio Engineering*, where automatically detected segment boundaries can provide musically meaningful markers for the editing of music recordings [103], and segment identification can be used to develop of active listening stations [51] and lyric-to-music synchronisation algorithms [116, 49, 67] in consumer electronic devices .
- *Generative Arts*, where segment identification can be used to create melodic and harmonic schemata templates for automatic composition [29] and improvisation systems [102]. Also segment boundaries can be used as markers for music-to-video/text editing and synchronization [93].

In addition, computer models of segmentation are relevant to test theories and hypotheses in fields like Music Psychology, Music Cognition, and Musicology.

1.3 Contributions

This paper presents the following contributions in respect to previous surveys:

1. **Perception-centred Taxonomisation:** Previous research discussing computer models of segmentation have often classified models in respect to the computer modelling approach employed (e.g. 'rule-based' [89], 'state approaches' [92], or 'memory-based' [121]),[1] resulting in model classes where the aims, and perhaps more importantly, the possible interpretation of segments identified by the models might not be immediately clear. Conversely, In this survey we approach the taxonomisation and discussion of segmentation models primarily in respect to which 'perceptual cues'[2] they attempt to model. While our taxonomy does take into consideration technical aspects to organise models, such as the computer modelling approach employed, we believe our focus on targeted cues makes the aims of the models more explicit, hence revealing possibilities of integration between models, and highlighting which aspects should be taken into consideration for model evaluation and comparison.

2. **Survey of perceptual cues:** Previous research discussing computer models of segmentation often list 'principles' (perceptual cues) observed in music cognition research that might explain segment perception (e.g. [84, 88]). However, the list of principles is commonly incomplete, including only those cues directly related to the segmentation models being discussed. On the contrary, in this paper we provide a survey of the literature on segment perception, with the aim of establishing a more complete perceptual cue taxonomy, which we use as the foundation of the taxonomosations of segmentation models in subsequent sections.

3. **Survey of terminology:** The literature on music segmentation comprises domains as diverse as musicology, music cognition and perception, and music information processing and retrieval. As it can be expected, the terminology employed in these domains has a fair degree of overlap. However, there are still widely used terms that are inconsistent across domains and thus obscure rather than clarify the purpose and aims of segmentation models (e.g. the term 'structural segmentation', frequently employed in the literature of music information retrieval). In this paper we present a survey of terminology, aiming to (a) provide operational definitions concerning segments and segmentation, and (b) distinguish among the different sub-areas within music information retrieval that are concerned with segmentation.

4. **Coverage:** This survey is, to the best of our knowledge, the most extensive survey on symbolically-encoded music segmentation models to date.

1.4 Document Description

In §2 we define and discuss all the terminology and fundamental concepts used in this survey. In §3 we review experimental studies that have studied segmentation in an empirical setting, and also summarise the theories proposed to elucidate the cognitive factors influencing music listening (we focus on those concerned with segment perception). In §4 we taxonomise, describe, and discuss computational models of segmentation for symbolic encodings of music. In §5 we discuss evaluation methodologies, campaigns, and results of the surveyed computational segmentation models. Finally, in §6 we present our conclusions.

[1] One exception being [84], which is closer to the way to taxonomise models presented in this paper.

[2] The musical factors that have been observed to trigger the perception of a segment boundary. Perceptual cues are discussed in detail in §3.2.

2 Music Segments: Definitions & Terminology

In this section we define and discuss all the terminology and fundamental concepts used in this survey. In §2.1 we discuss basic terminology. In §2.2 we provide operational definitions concerning segments and segmentation, in §2.3 we define the task and sub-tasks associated to the computational modelling of segmentation, and finally, in §2.4 we taxonomise and describe the areas within music information processing that are concerned with segmentation.

2.1 Basic Terminology

In the following we define terminology from music perception and music theory used in this survey to describe and discuss musical segments.

2.1.1 Terminology from Music Perception

In [17, 61] notions in music perception such as musical description, category formation, and identity/similarity have been discussed. The concepts established have been used in a number of publications dealing with music analysis and music information retrieval. In the following we summarise concepts defined in the aforementioned studied that are of relevance to this survey.

Entity: In cognitive science *entities* denote perceptually complete and distinct "things" [17], often referred as *objects* in the visual domain and *events* in the auditory domain. In music theory event type entities can be notes, chords, melodies, phrases, motives, or even whole pieces. We refer to event entities more abstractly as either segments or streams. Segments refer to entities that are perceived to be organised sequentially, such as notes, figures, phrases, sections, etc. Conversely, streams refer to entities that can be perceptually distinguished from one another even though they occur simultaneously, such as different voices in a polyphony, melody plus accompaniment, etc.

To categorise the different levels in which an entity can be described, we use the terminology summarised in Table 1.

Term	**Definition**
property	Any premise that may be used to describe an entity [17].
attribute	Specification that defines a property of an entity [17].
feature	Attribute (or set of) that is *salient*, *distinctive*, and *meaningful* [61]. In this context *salience*: noticeability of an attribute, *distinctiveness*: relative salience of an attribute in a given context, and *meaningfulness*: attributes that are pertinent in describing a set of entities in a given situation.

Table 1: terminology used within music perception to differentiate levels of description of entities.

2.1.2 Terminology from Music Theory

To categorise segments we employ terminology from music theory used to describe structural units [98, 114, 113, 101, 119]. In Table 2 we collect definitions of basic structural units that are relevant to this survey. The definitions collected provide information on the intended structural function of the unit, as well as their approximate duration.

Term	**Definition**
note	Basic unit of musical structure, notational backbone of the western music. Specifies one musical event. Ranges from a fraction of a second to several seconds [98].
figure	Smallest musical unit with individual expressive meaning. Roughly 2-12 consecutive notes [114].
motive	Occasionally used as synonym of figure. Normally there is a distinction: the motive is a thematic particle (representative of the music), while a figure is used for accompaniment [114].
sub-phrase	Any unit smaller than a phrase [101], similar in length to a figure.
phrase	Aggregation of consecutive notes "expressing a complete musical thought" [114], or "containing significant tonal motion" [101]. Roughly 4-8 measures in length [119].
form	Overall architecture of a piece of music. Describes the layout of a composition as divided into sections [98].

Table 2: Relevant terminology of structural units used in music theory.

2.2 Musical Segment and Segmentation: A Working Definition

In this survey we consider a segment as an *entity that has a structural function in the perception and cognition of music* and segmentation as *the process of abstracting segment entities during (and after) the act of listening to music.*

In this paper we survey computer models mimicking this human capacity.

2.3 Computational Modelling of Segmentation: Task & Sub-Tasks

2.3.1 Task

In this paper we consider the *task* of segmenting digital music files as the process of automatically determining segments given a symbolic or sub-symbolic representation of a musical piece/melody/part.[3] That is, a computational segmentation analysis aims to automatically identify the structural constituents of a musical entity given as input. It is important to notice that a segmentation analysis does not intend to fully describe the internal structure of the constituents, nor to provide a complete description of the role of these constituents within the input musical entity.[4] Below we outline the type of segment attributes a computational segmentation analysis aims to capture.

2.3.2 Sub-tasks

The *sub-tasks* associated to segmenting digital music files consist of one or all of the following:

(a) *boundary detection*, i.e. locate to points in time which divide two contiguous segments.

(b) *boundary pairing*, i.e. identify which pair of boundaries encompass a segment (in case the perception of overlapping segments is assumed possible).

[3] The task of automatically detecting the boundaries of segment entities in digital music files has been commonly referred to as segmentation [127, 84], but also as grouping [69, 88] and chunking [50]. The term "grouping" suggests a part-to-whole modelling approach, where segment detection is the result of perceptually aggregating contiguous musical events that represent a given musical piece at some predefined atomic level. The terms "segmentation" and "chunking", on the other hand, suggest a whole-to-part modelling approach. Thus, starting from a musical piece described as a sequence of perceptually discrete events, segments can be identified by dividing the sequence into partitions/chunks of neighbouring elements. Despite the fact the terms grouping and segmentation/chunking by themselves do suggest a particular methodology to identify the sought-for segment entities, the computational models surveyed using one or the other to describe their models do not necessarily adopt the methodology suggested by the term. In this survey we consequently consider the terms as interchangeable.

[4] This type of analysis is has commonly been referred to as 'shallow' parsing in the field of natural language processing, to differentiate it from 'deeper', i.e. more exhaustive, types of structural analysis.

(c) *segment granularity level identification*, i.e. identify the granularity level of the segments (in case the perception nested/hierarchical segment organisation is assumed possible).

(d) *segment labelling*, i.e. categorize segments in respect to their relationship with other segments. For example, using a letter label (e.g. A, B, etc.) to designate all segments with similar musical characteristics, or using a semantic label (e.g. stanza, verse, antecedent, etc.) if structural stylistic conventions are assumed.

An example of the segmentation of a simple melody is depicted in Figure 1. In the figure, segment boundaries are represented as downward arrows. Two nested segment layers are illustrated using horizontal curly brackets, and letter segment labels are used to categorise segments at the two layers.

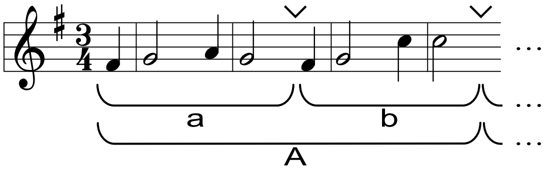

Figure 1: Example of segmentation in a simple melody. The analysis depicts a hierarchical ordering (bottom-to-top) of two levels of segment granularity.

2.4 Computational Modelling of Segmentation: Research Areas

In §2.4.1 we define which segment granularities have been targeted by the surveyed computer models of segmentation. Subsequently, in §2.4.2, we list and describe areas of research dealing with segmentation within music information retrieval.

2.4.1 Segment Granularity Targeted by Computer Segmentation Models

The *time-span* of a segment, has a strong influence on its perception and structural function, making it necessary to define a time scale or 'granularity' when attempting a segmentation analysis. To classify computer segmentation models in respect to segment granularity, we resort to terminology from music theory employed to designate structural units. In Figure 2 we present a parallel between structural units of western music theory and different aspects associated to the perception of musical time scales.

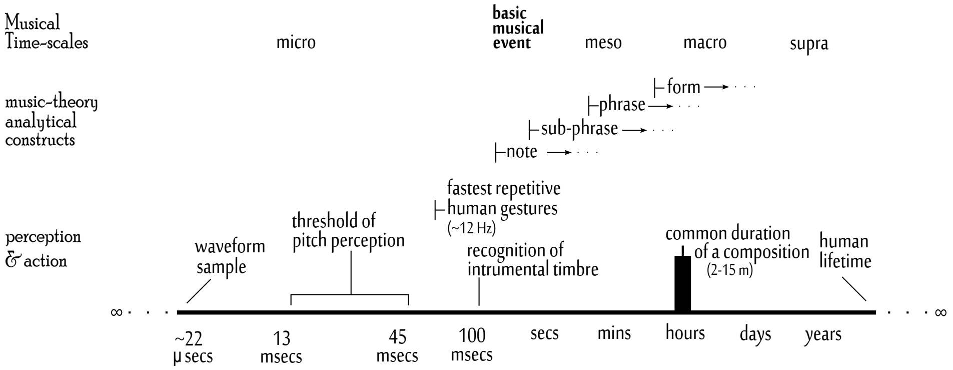

Figure 2: Broad view of musical time scales. The figure and definitions within have been adapted from [98, pp. 3–6].

The work on computer modelling of segmentation surveyed in this paper focuses mainly in the *meso* and *macro* time scales of music. As it can be seen in Figure 2, these time scales comprise segments ranging from note events to form. In this section we mainly discuss three segments granularity classes: relatively-short (referred as *subphrase-level*), medium-sized (referred as *phrase-level*), and relatively long (referred to as *form-* or *section-level*).

2.4.2 Areas of Research in MIR

The granularity classes discussed above have given rise to an equal number of segment inference areas. These areas can be organized by the level of detail and abstraction required in their specific application scenarios. Here we describe three different ones: *motif discovery* (often at subphrasel-level granularity), *melodic segmentation* (often at phrase-level granularity), and *structural segmentation* (often at form-level granularity),[5]. These areas are defined in Table 3.

Term	**Meaning**
melodic segmentation	Focuses on segment boundary detection on monophonic music (most often melodies). Segment granularities targeted are commonly at the phrase and sub-phrase level. Analysis assumes phrases do not overlap, and hence boundary pairing is not necessary. The resulting set of segments must exhaustively subsume all melodic events.
motif discovery	Focuses on segment boundary detection on monophonic and polyphonic music. Segment granularities targeted are commonly at the sub-phrase level. The analysis assumes segments do overlap, and hence boundaries need to be paired. The analysis also attempts to identify perceptually salient segments ('motives' in terms of music theory), and thus segment labelling and filtering is also required. The resulting set of segments is not required to subsume all the events events constituting the piece of music.
structural segmentation	Focuses on segment boundary detection and segment labelling mainly of polyphonic music. Segment granularities targeted are commonly at the section level. Analysis assumes sections do not overlap, and hence boundary pairing is not necessary. The resulting set of segments must exhaustively subsume all the musical events constituting the piece of music.

Table 3: Some areas of computational segmentation.

Other tasks in music information processing and retrieval might also be seen as ways to segment music: finding note onsets/offsets or chord boundaries as part of transcription systems, finding metric bars, identifying instants/regions with certain affective qualities (e.g. points of high 'tension' or sounding 'triumphant'). Yet, in this survey, we are concerned with segments which are sequentially coarser than individual notes or chords, and we consider that segments delimited by bars or bounding regions with an specific affective content do not always comply with our definition of segments (for a short description and discussion of metric induction and other perceptual structuring processes that work in parallel or in combination to segmentation see Appendix A).

[5] Computer segmentation models targeting the segmentation of music recordings commonly employ the term "structural segmentation", which in itself might imply many (or all) segment granularity levels, yet as observed in [84], the segment granularity and time scale of analysis has been approximately the same in surveyed publications, and refers to musical *form*.

3 Segment Perception

In this section we define, taxonomise, and comment on musical 'cues' hypothesised to influence segment boundary perception. Our main goal is to use the taxonomy of cues to classify computer models of segmentation in subsequent sections.

In §3.1 we first provide an operational definition of segment boundary cue, and subsequently, in §3.2, introduce a taxonomy of segment boundary cues. Both the definition and taxonomy are based mainly on theoretical work on segment perception as well as intuitions of scholars from music theory. In §3.3 we provide examples of phrase-level and form-level segment formation. In §3.4 we summarise and discuss observations made in respect to the cues included in our taxonomy, focusing on the number of cues observed, the ways in which cues seem to interact, and how stable they might be across different listeners. Finally, in §3.5 we outline section conclusions in respect to how these cues are relevant for computational modelling of phrase-level and form-level segmentation.

3.1 Segment Boundary Cues: Definition

In this survey we employ the term boundary *cue* to refer to the musical factors that suggest to the listener the temporal location of a given segment boundary. In the words of Deliège [34, p. 214]:

> "A cue should [...] facilitate the formation of [...] groups at various [...] levels and enable the totality of the work to be circumscribed. These cues are nothing other than input tags. Most of them are temporary and fleeting."

To provide an operational definition of boundary cue we resort to basic terminology of music cognition (such as entities, attributes, and so on, see Table 1, Appendix 2.1.1). Prior to our definition is important to reiterate that we focus on the *messo* and *macro* time scales of music, with segments granularities coarser than individual basic musical events (coarser than notes in music-theoretic terms).[6] At these granularities we can expect segments to have internal organisation dependent on constituent segment entities. That is, segments at one level of granularity can be decomposed into sequences of segments at finer granularity levels.

With the previous in mind, we define a boundary cue as a relationship perceived between (two or more) musical entities. The relationship can be 'categorical', e.g. the assessment of similarity or dissimilarity between two melodic motive entities, or 'functional', i.e. the functional association of an entity in respect to contextual factors, arising from the same piece or from previously heard music (such as hearing an entity as being stressed due to its metrical position, or the recognition of a specific tonal cadence). In Figure 3 we provide a visual depiction of boundary cues.

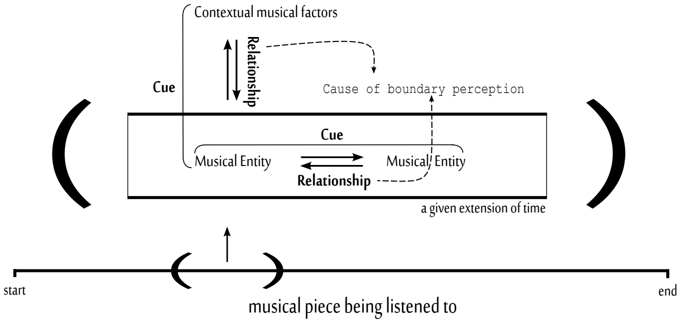
Contextual musical factors
Relationship
Cue
Cause of boundary perception
Cue
Musical Entity
Musical Entity
Relationship
a given extension of time
start
end
musical piece being listened to

Figure 3: Diagram depicting boundary cues.

[6] In this section we consider only music related cues, therefore excluding, for example, cues related to linguistic factors in vocal music, such as the possible influences that word co-articulation (or the phrase and syntactic structure of the text) have in vocal melody segmentation.

3.2 Segment Boundary Cues: A Taxonomy

In Figure 4 we present a taxonomy of segment boundary cues hypothesised to influence segment perception at the *messo* and *macro* time scales (from sub-phrases to sections). Our taxonomy draws from traditional (e.g [69, 82]) as well as more recent (e.g. [50, 56, 7]) theories of music and segment perception (all briefly described in Appendix B). In Figure 4 we present a segment boundary cue as a an {Entity, Relationship} dupple. We provide information of the source (perceived in the piece being listened to or in respect to previously heard music), the relationship-type (the cause or 'trigger' of boundary perception, highlighted using a bounding-box in the figure), and a general description of the musical entities involved (attribute-class of the relevant feature describing the entity, and the approximate time span of the entity). We elaborate on each of these aspects of our taxonomy in §3.2.1 and §3.2.2.

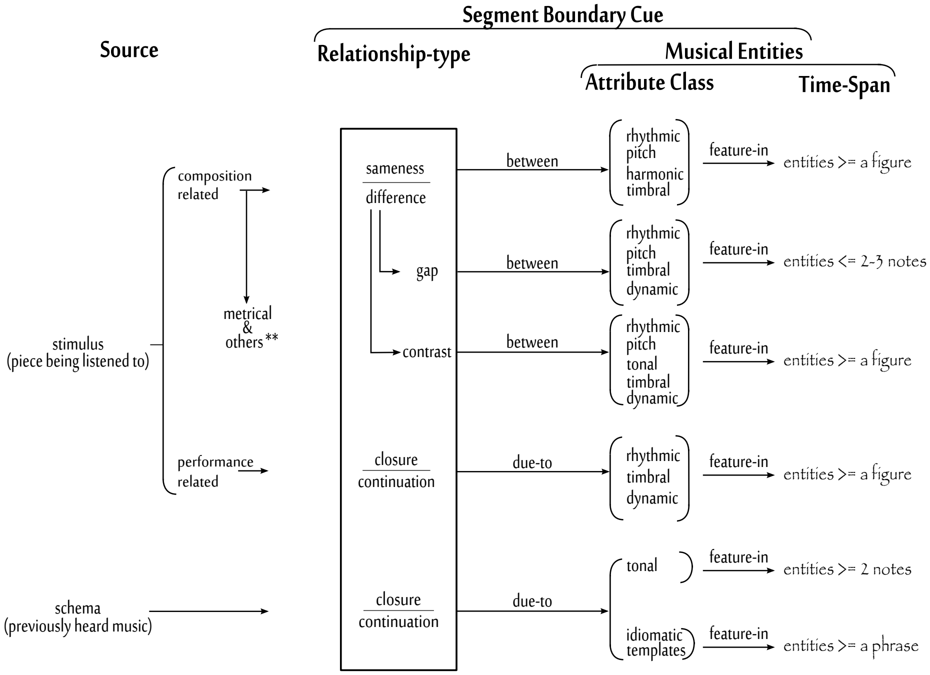

Figure 4: Taxonomy of cues influencing the segmentation of music. The taxonomy is organized left to right. Each taxonomisation class is highlighted in bold on the top. Concrete examples how cue causes affect the perception of segments are given in §3.3. All cues listed are hypothesised to occur in both polyphonic and monophonic music. Cues stressed with ** are not discussed in §3.2.1.

3.2.1 Entity Description and Source

Entity description: In our taxonomy we describe entities in respect to the attribute classes which contain the relevant features for comparison, and the time-span of the entities. The musical attribute classes considered in the taxonomy are described using standard terminology of music theory. However, one exception can be observed, a class we have termed 'idiomatic templates', which refers to segment structures found often enough in a given style to allow human listeners to form or abstract templates of these strucutures. Idiomatic templates are often at the level of form (e.g. verse-chorus form in pop, or strophic form in folk), but can also refer to global aspects of segments (e.g. prototypical pitch contours of melodic phrases). The time-span of entities considered is relatively brief (single intervals or at most 2-3 consecutive notes) or larger (such as figures, motives, phrases, or longer entities).

Source: We distinguish the information *source* from which the attributes describing the entities relevant for cue formation are abstracted. We classify the source of information as *stimulus* (internal) or *schema* (external) to the piece being segmented.[7] In the case of stimulus-related cues, the taxonomy distinguishes between *composition-related* and *performance-related* cues. The former refers to the case where the attributes used to describe entities are observable in conventional western music notation, and the latter to the case where the attributes are related to performance expression and body gesture.

3.2.2 Segment Boundary Cues

Below we describe the two aspects of cues as a {Entity, Relationship} dupple.

Relationship-types: We sort cues according to the terminology employed in the literature to verbalize the *cause* of boundary perception, i.e. the relationship-type between entities. In our taxonomy we employ four terms (presented as a pair of dualities) to denote the cause of a cue: *sameness/difference* [33, 6], and *closure/continuation* [81, 78]. Both sameness/difference and closure/continuation involve the assessment of a relationship between entities, the main distinction between the two is that the former places emphasis on the entities themselves, while the later places emphasis on the function of an entity in a given musical context.

Sameness/difference relationships: are related to semiotic paradigmatic analysis.[8] A paradigmatic analysis of a piece involves the examination of *similarities* (sameness) and *dissimilarities* (difference) between musical entities perceived within the piece [6]. We distinguish two types of differences, dependent on the time scale for the entities to be perceived, and the temporal proximity between entities. For brief and temporally-proximate entities differences are perceived as *gaps* in the flow of music, while for larger and not necessarily temporally-proximate entities differences are perceived as a *contrast* in musical material, i.e. a perceptually noticeable deviation of one or more attributes describing an entity discerned earlier within the current segment.

Sameness/difference examples: An example of a sameness cue is the recognition of a melodic figure listened to earlier in the piece. An example of gap type differences are musical rests or caesura, and, in the case of performance related in formation, vocal breaths. An example of a contrast type difference is the perceived contrast between two rhythmic motives characteristic on segment, e.g. a galloping rhythmic pattern and a 'shuffle' rhythmic pattern.

Closure/continuation relationships: are, on the other hand, related to semiotic syntagmatic analysis. A syntagmatic analysis of a piece involves the examination of *local functional relations* between musical entities found within the piece [6].[9] Functional relations between entities relevant to segmentation have been commonly formulated in respect to *musical expectation.* Points of conclusion are hypothesised associated to a disruption in expectation [90], and conversely points of continuation are associated to generation of expectation [59, 111].

Closure/continuation examples: An example of a point of closure is the identification of the end of a cadence. An example of a point of generation of expectation is the recognition of a chord progression common within a style.

As a final note, is important to stress that although we explain and depict sameness/difference and closure/continuation cues separately, this by no means is intended to suggest that these two

[7] The distinction between information from the piece being listened-to and information from previously heard pieces is recurrent in the literature of computational models of music perception, with Justus and Bharucha [64] referring to this as *veridical* or *schematic* knowledge, Crawley et al [31] classifying parsing modes into *stimulus-* or *scheme-* driven, and Pearce et al [88] proposing the use of *short-* and *long-* term models for data-driven segmentation. In this document we use the terminology of Crawley et al [31].

[8] Sameness has also been often referred to as *musical parallelism* in the literature of music cognition and symbolic music research [69, 15] (see GTTM's grouping rule 6 in Table 11), and *self-similarity* (or plainly *repetition*) in the literature of segmentation of musical audio [84].

[9] It seems pertinent at this point to stress that while the assessment of the 'relationship type' between two entities is local, the source of knowledge required is commonly non-local (related to the previous exposure to music of a given listener as well as entities that might be located far apart within the piece).

cue groups are *independent*. In fact, quite the opposite, it has been noted a number of times (e.g. in [74]), that local functional relationtips are most likely to influence our perception of similarity.

3.3 Boundary Cues in Musical Phrases and Sections

In Figures 5 and 6 we present two examples of how the cues listed in our taxonomy might help shaping phrase-level and form-level segments, respectively.

3.3.1 Boundary Cues in Phrases

First, in respect to **phrase-level segments**, Spiro [111] provides a classification of the structure of (mostly polyphonic) phrases in Common Practice Period music. Spiro distinguishes between simple (non-overlapping) phrases and complex (overlapping) phrases. In Figure 5 we present a diagrammatic view of a prototypical simple phrase, which we comment and explain below.[10]

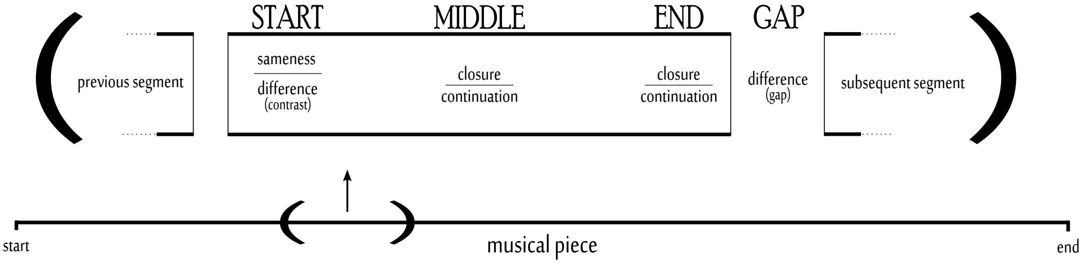

Figure 5: Anatomy of a 'simple' phrase-level segment (based on the work of Spiro [111, ch. 13]). In the diagram a musical piece is depicted as a bounded-time line, a slice of time of the piece is depicted using parenthesis, and three subsequent phrase-level segments within the time slice are depicted as rectangles.

In the following we elaborate on the sections identified in Figure 5 :

- START: The phrase start refers to a region including the 'first few' notes of a phrase. According to Spiro "The phrase start can establish the thematic and tonal centre of the phrase" [111, p.402]. In our diagram, we associate phrase starts to sameness/difference type cues. In respect to sameness, it has been proposed that recognizing the first few notes of an entity spotted earlier in a piece (e.g. a melodic figure), may help identifying, in retrospect, the starting point of a new phrase [2, 15]. Similarly, identifying contrasts between contiguous phrases (e.g. a change of key or a change of instrumentation), has been seen to, again in retrospect, help the identification of the starting points of phrases [27].

- MIDDLE: The phrase middle refers to the section of the phrase where an expectation for the end begins. This seems to be mostly determined by continuation type cues (e.g. voice leading), which give a sense of 'direction' to the evolution of music and thus might cause listeners to generate expectations of the overall length and point of conclusion of a phrase.

- END: The phrase end refers to a region including the 'last few' notes of a phrase. Spiro considers phrase ends to be determined mostly by closure/continuation type cues, hypothesising that the starting point of the END region is perceptually identified when there is a clear expectation for the end. This might be cued, for example, by the identification of the beginning of a cadential progression (a continuation type cue in our taxonomy). Spiro argues that the end point of simple phrases is most often cued by the music-theoretic concept of resolution[11] (which corresponds with a closure type cue in our taxonomy). We would add to her observations that the presence of performance-related cues might also be of significance for END region perception, e.g. it would agree with intuition for the combination of

[10] Spiro's complex phrases are not straightforward to summarise and are therefore left out of this example.

[11] Move from a dissonant or unstable sound in respect to the local key, to a consonance or stable sound in respect to the local key.

expressive tempo fluctuations (e.g. a ritardando) and a gradual change in dynamics (e.g. a decrescendo) to both generate expectations for the end of a segment, and give a sense of closure marking the end of a segment.

- GAP: Spiro proposes GAP as a delimiter which may be found after the end of a phrase. We identify GAP with our difference of type gap cue. Spiro argues that gap related cues do not inform the listener of aspects of a phrase, and thus can not be considered as determinant of the internal structure of a phrase.[12]

3.3.2 Boundary Cues in Sections

For our second example, this time in respect to **form-level segments**, we refer to Deliège [34] Deliège proposes form-level segment formation (in non-tonal music) as a bottom-up recursive process. The recursive process starts by identifying relatively brief segments (sub-phrases, phrases), and then organise the identified segments into larger segments (sections). Deliège argues that brief segments are mainly determined using gap related cues, and that larger segments are cued by sameness and contrast. Deliège suggests that sameness is used to perceptually 'cement' segments, and contrarily that contrast is used to 'demarcate' segments. In Figure 6, we present a diagram illustrating this paradigm. Even though Deliège hypotheses are tested only on non-tonal polyphonic music, there is some supporting evidence for her hypotheses holding in Common Practice Period music [27]. Also, a similar (yet more elaborate) paradigm has been proposed by Bimbot et al [6] for Pop music segment annotation.

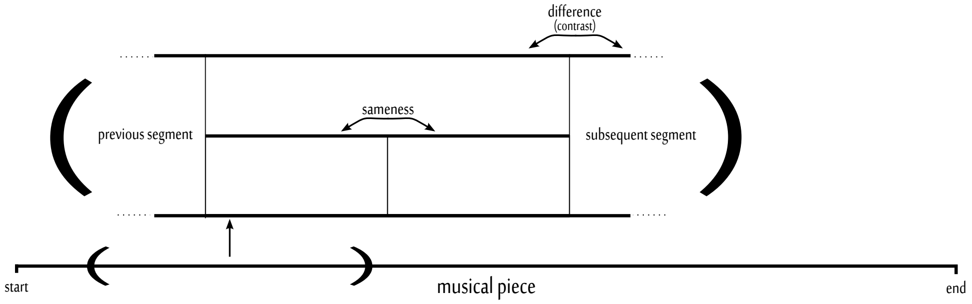

Figure 6: Form-level segment formation (based on the work of Deliège [34]). In this diagram a musical piece is depicted as a bounded-time line, a slice of time of the piece is depicted using parenthesis, and three adjacent form-level segments within the time slice are depicted as rectangles. The two smaller rectangles represent segments at one level (phrases), while the larger rectangle enclosing the two smaller ones represents segments at at a higher level (sections).

3.4 Empirical Work On Testing of Boundary Cues

In this section we discuss observations made in listening studies of segmentation with respect to segment boundary cues.

The perceptual studies surveyed in this section are summarised in Table 4, page 15. The table contains information of: (a) the *study* (focus, etc.), (b) the *stimulus* (duration, etc.), (c) the *test subjects* (number of subjects, etc.), and (d) the *results* (type of cues observed, etc.).

We centre our discussion of the surveyed studies in respect to segment boundary cues (§3.4.1), and how consistent human subjects are when segmenting music (§3.4.2). In our discussion we also refer to comparative studies of computational models of melody segmentation (Table 10, page 32), and three other survey publications [36, 115, 91] of music segment perception.

[12] Spiro sees gaps as the equivalent of 'white space' in written language, so that just as white-space is not informative of a word's morphology, she considers gaps not informative of a phrase's morphology.

3.4.1 General Observations in Relation to Boundary Cues

We approach our summary of observations of segment cues by attempting to give answers to the following questions:

How many cues are commonly at play when listening to music?

- The average number of distinct cues reported to influence boundary perception for pieces longer than 5 minutes is normally larger than 10 [27, 110, 12]. In short excerpts (≈ 30 seconds or less), normally 8-9 cues have been observed [34, 35].

(Cue Importance 1) Which cues seem to be used most often by listeners to determine segments?

To determine how often a cue is employed by a subject, perceptual studies either ask participants to give a description of the cues they employed for each boundary (e.g. [12]), or collect music where the locations of potential cues are known, and then during analysis the researchers check if the location of annotated boundaries coincides with the location of a potential boundary cue (e.g. [33]). In respect to this we can observe:

- Cues used more often seem to be those associated to *temporal proximity gaps*, primarily in respect to musical silence (rests, caesura, vocal breaths, etc.), and to a slightly lesser degree to prolonged note durations. Temporal proximity gaps have been observed to cue phrase-level segment boundaries in a range of (polyphonic and monophonic) musical genres, including Folk, Pop, Jazz, and Common Practice Period.
- Also, in experiments using Common Practice Period music [34, 27], cues associated to gaps in timbre and gaps in dynamics appear to play an important role.

(Cue Importance 2) is there evidence of a primacy effect of single cues over other cues?[13]

Cue combination and primacy in a given situation is difficult to test empirically due to the natural trade-off that arises between ecological validity (i.e. using 'real' music) and parametric control. Hence, studies that investigate cue combination and primacy effects often employ artificial 'music-like' stimuli. In these studies we can observe:

- A primacy of temporal gap cues (prolonged durations) and dynamic gap cues over pitch proximity gap cues in artificial monophonic tone sequences [126].
- A primacy of temporal gap cues (musical silence) over sameness cues (pitch patterns) in artificial monophonic tone sequences [36].

(Cue Importance 3) Is there evidence of primacy of a given cue class over other cue classes? (i.e. stimulus (composition, performance), or schema)

- In respect to stimulus (score) related cues and schema related cues, perceptual studies have found no definitive evidence which might suggest primacy of one class over the other. The class and number 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" [34, p. 228]. In computational research of segmentation, cues derived directly from the stimulus are often considered of higher importance than those reflecting scheme derived information [15, 43, 86]. Notwithstanding, in [106, 8] the authors show the strong importance of style specific listening experience in vocal folk songs, evaluating the influence of having internalised the segment structure of previously heard pieces.
- In respect to stimulus (performance) related cues, it has been observed that, while the positions of boundaries tend to be determined primarily by stimulus (composition) related and schema related cues, it seems that performance related cues can have an effect on the overall number of boundaries listeners perceive [112].

[13] Is a particular cue preferred when a conflict between two or more cues occur (where by conflict we mean that to cues might be suggesting different groupings).

3.4.2 Observations in Respect to Test Subject Behaviour

We approach our summary of test subject behaviour by giving answers to the following questions:

Are human listeners self-consistent in detecting segment boundaries and labelling segments?

- Subject annotation self-consistency, i.e. whether subjects agree with their own boundary annotations in subsequent trials of an experiment, has not been broadly investigated. However, some studies suggest that subjects are only moderately self-consistent when annotating segment boundaries for Pop [11, Ch.2] and Common Practice Period music [112].

Do human listeners agree with other listeners in detecting segment boundaries?

- Boundary locations with high inter-annotator agreement (IAA) are normally reported to correspond to form-level boundaries. For phrase-level granularity, on the other hand, the level of agreement seems to be linked to the complexity of the melodies. As an example, high IAA was observed in the study of de Nooijer [32, 127] where pop melodies where used for testing. Conversely, the studies of Thom [121] and Pearce [88], in which the test sets included jazz and classical melodies, report low IAA.
- Performance related cues influence IAA in such a way that perceived segmentation between two performances of the same piece may vary in both the *number* of boundaries and the *position* of those boundaries, in a manner proportional to the number of cues present in the music (i.e. the more boundary cues, the higher the variation) [110].

Does musical training influences segment perception?

- Musical training seems to not have a big influence between *musicians* (subject plays an instrument) and *non-musicians* (subject does not play any instrument). On the other hand, between *degree-level musicians* and non-musicians the studies surveyed report contradictory findings. As an example we can mention Deliege [33] who reports that non-musicians present a lower fit to GTTM principles than musicians, while Schaefer [106] reports the opposite. Also, in respect to the amount of segments identified by each type of subject, e.g. Deliège [33] observes that non-musicians perceive more boundaries than musicians, while Spiro [110] reports the opposite.

3.5 Section conclusions

3.5.1 Directions for Computational Segmentation Models

Studies have mostly investigated cues in non-tonal music, common practice period music, folk, and pop. In these genres, a number of boundary cues have been identified which we have grouped into families in our cue taxonomy (Figure 4). Perhaps more importantly, listening studies point to the diversity of cues, and diversity of situations in which a boundary cue might take precedence over others. Hence, these studies point to the need of models that are able to include more than one cue, and to combination mechanisms that are able to adapt to specific circumstances . That is, model cue 'prevalence' as a dynamic process that reacts and adapts as musical circumstances change. The modelling of this dynamical process needs the simulation of both real-time and retrospective listening, as well as information gathered from multiple listens of a piece and from previously heard music.

How many sources of information might we need to model boundary detection computationally? It seems that, for metrical music, beat level information has a big influence. For tonal music, tonal and harmonic[14] factors seem to be of importance. For both tonal and non-tonal music, subphrase-level segments (figures and motives) seem to be the entities upon which we perceive something being the same or different. Also for tonal and non-tonal music, conventions of style regarding segments, i.e. length, functional templates (phrase rhythm in jazz, sonata form for sections) - it seems that when a listener is aware that the music (s)he is listening to belongs to a specific style, the listener uses stylistic conventions to disambiguate possible segmentations that might be suggested by different cues.

[14] Both explicit (present in polyphonic music) or implied (suggested in monophonic music through voice leading and other methods).

3.5.2 Future Directions Perceptual Studies

Below we discuss future directions regarding perceptual studies. We divide our discussion in respect to studies focusing on testing a statistically significant number of participants (under ‘listener studies’), and those focusing on coverage (under ‘corpus analysis of segmentation cues’).

`In respect to listening studies`: In Table 4 we can see that, from the mid 1990s onward, perceptual studies have normally employed ‘ecologically valid’ stimuli (i.e. audio recordings or music synthesised including performance related information). However, using ecologically valid stimuli is non-trivial to test (a) how and when cues combine, and (b) in which situations one cue might take precedence over others. Consequently, there is a need to complement perceptual studies that employ audio recordings or expressive synthesis with perceptual studies in which stimuli can offer a greater degree of control (i.e. permit parametric manipulation, and allow to focus on a relatively small amount of musical dimensions). As an alternative, [94, 1] proposed to use minimalist music as a test bed for perceptual segmentation studies, arguing that the systematic compositional strategies used to create these works offer a good trade-off between ecological validity and the need for focused, controlled test environments. Moreover, the systematic processes employed for the creation of minimalist works allows to generate new musical material for testing.

`Corpus analysis of segmentation cues`: The number of studies investigating the generality of principles hypothesised to act as cues of boundary perception in freely accessible annotated corpora is not large (we are aware of only three studies [100, 108, 10]). While the fact that boundary indications in large corpora commonly represent the perception of only a handful of annotators does limit the validity of the conclusions of such experiments (on their own), if these experiments are seen as a complement to (the previously discussed) listening studies, corpus analysis can be used to refute or support observed segment perception behaviour. Some aspects that have been already studied in corpus based analysis are: statistical evidence for the presence of closure related cues at melodic phrase-ends in vocal melodies [10], statistical evidence for the presence of pitch and duration gap-related cues at melodic phrase ends in vocal melodies [100], statistical evidence for the presence of contrast related cues in polyphonic pop music [108]. Some questions that could be addressed in corpus based studies are: the role of pitch expectation in pitch related discontinuity perception at melodic phrase boundaries, or the role of motivic pattern repetition for both melodic/polyphonic form/phrase boundaries.

Table 4: **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	Date	Focus	Segs	Test set	Texture	Genre	Duration	Format	Q	#	Training	Cues	ATA	ASA	Conclusions
Deliège et al [33]	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 20th 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 [34]	1989	recognition of form in complete pieces	form	2 pieces	P	mid 20th 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 et al. [27]	1990	perception of temporal organization on relatively large pieces	form	2 piano pieces	P	mid 20th 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. [35]	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 [124]	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. [106]	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 [110]	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 accross performances.
Weyde [126]	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 et al [12]	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.

4 Segmentation Models of Symbolic Encodings of Music

In this section we taxonomise, describe, and discuss computational models of segmentation. In §4.1 we start by establishing the scope, motivation, and task definition of symbolic segmentation models. In §4.2 we continue by outlining the fundamental characteristics of a prototypical computational model of monophonic music segmentation, aiming to provide the reader with a general idea of its input/output behaviour, as well as pre/post processing stages. Finally, in §4.3 we classify and discuss models of segmentation of symbolically-encoded music.

All the models surveyed in this section have been summarised in Tables 5, 8, & 9.

4.1 Scope and Task Description of Symbolic Music Segmentation

Type of music: Most computer models of segmentation surveyed in this section have focused on the segmentation of **monophonic music**, i.e. music where elementary structural events are perceived as organized in a strictly sequential fashion. What is more, nearly all have constrained their scope to the segmentation of melodies. Melody segmentation models have been most often tested on vocal folk music.

Level of Music Representation: The segmentation models reviewed in this section also assume that the input is represented in **symbolic form**, i.e. any computer readable format where basic events correspond roughly to notes as notated on a score. This assumption is commonly justified arguing that monophonic music seems to be mentally represented in a way comparable to conventional western notation [80].

Task: To infer segment **boundary positions**, i.e. locate time points that divide adjacent segments. The models reviewed in this section assume segments do not overlap and hence is not necessary to distinguish between starting and ending boundaries. Also, in general no labelling of segments is required, yet some symbolic segmentation models also define a hierarchy of the resulting segments (e.g. [55, 130]).

4.2 I|O Description of a Melody Segmentation Model

In this section we describe the most generally employed pipeline in segmentation models starting from symbolic input (depicted in Figure 7). Due common constrains and limitations acknowledged in §4.1, the described pipeline corresponds to that of a *melody* segmentation model. We focus on the description of the input, output, and pre/post processing stages of the architecture. The boundary detection models are classified and described in §4.3.

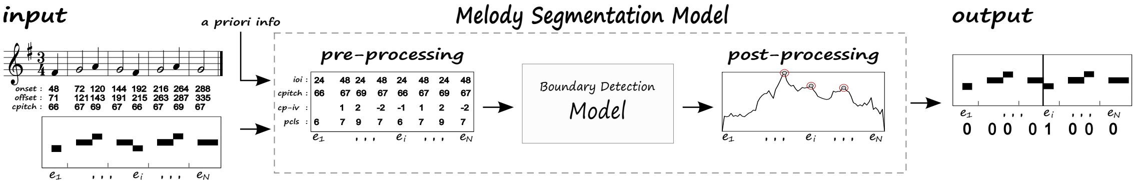
input
a priori info
Melody Segmentation Model
output
onset : 48 72 120 144 192 216 264 288
offset : 71 121 143 191 215 263 287 335
cpitch : 66 67 69 67 66 67 69 67
e_1 , , , e_i , , , e_N
pre-processing
ioi : 24 48 24 48 24 48 24 48
cpitch : 66 67 69 67 66 67 69 67
cp-iv : 1 2 -2 -1 1 2 -2
pcls : 6 7 9 7 6 7 9 7
e_1 , , , e_i , , , e_N
Boundary Detection
Model
post-processing
e_1 , , , e_i , , , e_N
e_1 , , , e_i , , , e_N
0 0 0 0 1 0 0 0

Figure 7: Input-output diagram of a prototypical computational model of melody segmentation. **From left to right**: input: melody and piano-roll representation, pre-processing stage: attribute computation, boundary detection model, post-processing stage: peak picking, output: boundary list where a 1 represents a border.

4.2.1 Input:

The input to a melody segmentation model typically consists of a sequence of temporarily ordered note events $e = e_1, \dots e_i, \dots, e_N$. In e each note event is commonly represented by a discrete pitch, an onset time, and an offset time, i.e. if we take ξ to be a discrete and finite event space and $\otimes$ as the Cartesian product operator, then $e_i \in \xi = \texttt{pitch} \otimes \texttt{onset} \otimes \texttt{offset}$.[15] This representation is commonly referred to as "piano roll". A depiction of a piano roll type input can be seen in the far-left of Figure 7.

4.2.2 Pre-processing:

During pre-processing the melodic sequence is normally converted into a number of parametric attribute profiles, hypothesized to hold higher perceptual relevance. The profiles are represented as a set of sequences of size M, $x^k = x_1, \dots, x_i, \dots, x_M$, where k denotes the attribute (e.g. chromatic pitch, inter-onset-interval, etc.), $i \in \{1, 2, \dots, M\}$, and $M \leq N$. A depiction of the melodic sequence represented as a set of parametric profiles can be seen as the first stage of model of Figure 7. A common choice of attribute profiles describe melodies as a sequence of chromatic pitch intervals and inter-onset-intervals. Some segmentation algorithms might also attempt an estimation of attributes corresponding to higher levels of description, such as representing pitch information in respect to a diatonic scale, or as a scale degree (e.g. in [23, 2]), or estimating a metrical rather than an event-based description of time (e.g. in [118, 5]). Given that the estimation of higher level attributes is generally non-trivial, some algorithms assume that this information is known *a priori* or that is supplied by the user (e.g. in [18, 23]).[16]

In Table 7, we define and classify melodic attributes employed by the computational segmentation models reviewed in this paper. The classification considers three levels of description (low, mid, high), based on the attribute taxonomisation proposed in [70]. The attributes are moreover organized into rhythmic, pitch, density, metric, harmonic, and tonal categories.

4.2.3 Post-processing:

As a post-processing step, many of the surveyed algorithms compute what can be described as a *segment boundary profile* (`sbp`) [13, 87, 122, 1]. A `sbp` is a vector of length N in which each element takes a value $= [0, a]$, where a is some finite upper limit that varies depending on the algorithm. (A depiction of a `sbp` is provided towards the right end of Figure 7.) The interpretation of the numerical value assigned to each note event position in a `sbp` is model-specific. As an example, in Camboroupoulus' LBDM [16, 14], numerical values correspond to the strength of duration- and pitch- related 'local discontinuities', while in the case of Chew's Argus [23] the values reflect the amount of tonal contrast in a selected local context, and in the case of Pearce's IDyOM [91, 87] the values are related to the amount of (information-theoretic) surprise. As a consequence, the aspects of the `sbp` that may indicate a segment boundary depend on the nature of profile itself, so that boundaries might be associated to local maxima [16, 13, 18, 91, 87, 94], local minima [91, 43], or points of inflection or discontinuity [44]. The process of selecting peaks/troughs/others as boundary indicators has been commonly implemented computationally using heuristics. In Appendix C we list commonly used heuristics.

4.2.4 Output:

The output of a segmentation algorithm generally constitutes a single list of boundary locations, yet in some occasions it can be a ranked set of lists [5, 4] (each reflecting a different segmentation), or simply a set of lists reflecting different possible interpretations of segment structure [2, 19].

A list of segment boundary locations is normally encoded in vector form as $\mathbf{s} = (s_1, \dots, s_i, \dots, s_N) \in \{0, 1\}^N$. In $\mathbf{s}$ value of 1 represents a segment boundary (whether the boundary corresponds to the starting or ending point depends on the model). In general s_1 and s_N are treated as trivial cases which correspond, respectively, to the beginning and ending notes of a melodic phrase.

[15] Is common for melody segmentation algorithms to assume quantized duration values, i.e. that the onsets-to-offset interval matches duration values used in score notation.

[16] Another type of a priori information that is on occasion required is that specific to melodic segment structure, as for example an estimate for preferred phrase-level segment length (e.g. in [118]).

4.3 Classification of Symbolic Segmentation Models

In this section we present and discuss two classification schemes of symbolic music segmentation models. In the first, Figure 8, we organise models in a two-dimensional plot in respect to: `(a)` whether they employ *knowledge-driven* or *data-driven* methodologies, and `(b)` whether they embody *lower* or *higher* levels of cognitive processing. In the second classification, Table 6, we present a taxonomy that classifies models in respect to which segment boundary cues they attempt to model.

4.3.1 Surveyed Segmentation Models

In Table 5 we list the segmentation models surveyed in this section. The list considers the names or acronyms provided by the authors of the models. If a name for the model was not provided, we constructed an abbreviation (in order to sort models with more efficiency in Figure 8).

Model	**Year**	**Acronym definition**	**Reference**
TPG	1980	*TemPoral Gestalt grouping*	[120]
AGA	1989	*Automated Grouping Analysis*	[5]
GRAF	1989	*GRouping Analysis with Frames*	[4]
ESMS	1990	*Expert System for Musical Segmentation*	[19]
Cypher	1992		[102]
RAAM	1995	*Recursive Auto-Associative Memory*	[65]
LBDM	1997	*Local Boundary Detection Model*	[16, 14]
SPIA	1998	*String Pattern Induction Algorithm*	[13]
MPM	1998	*Music Punctuation Model*	[47]
RPF	1999	*Representative Phrase Finder*	[117] [21]
PSS	2000	*Piece-Sensitive Segmentation*	[68]
Grouper	2001		[118]
DOP	2002	*Data Oriented Parsing*	[8]
ISSM	2002	*Integrated Segmentation and Similarity Model*	[125]
MDSM	2003	*Melodic Density Segmentation Model*	[43]
E4MS	2003	*Entropy for Melodic Segmentation*	[44]
SONNET	2003	*Self-Organizing Neural NETwork*	[57]
EME	2004	*Entropy-based learning for Melody Segmentation*	[63]
qGPRs	2004	*quantification of GTTM's GrouPing Rules*	[46]
Modus	2004		[2, 3]
ATTA	2005	*AutomaTic span-Tree Analyser*	[53, 54]
Argus	2006		[23]
PAT	2006	*PATtern boundary strength profile*	[15, 18]
IDyOM	2006	*Information Dynamics of Music*	[91, 94]
IR4S	2006	*Information Rate for Music Segmentation*	[38]
JS4S	2007	*Jensen-Shannon Divergence for Music Segmentation*	[130]
AMS	2008	*Adaptive Melodic Segmentation*	[128]
HOSS	2009	*Heuristic Optimization for Symbolic Segmentation*	[95]
PIR4S	2009	*Predictive-Information Rate for Music Segmentation*	[1]
E4SS	2010	*Entropy for Structural Segmentation*	[30]
MTSSM	2010	*Multi-Track Segmentation of Symbolic Music*	[97]
SASS	2013	*Structural Analysis of Symbolic Music*	[129]

Table 5: List of surveyed models of symbolically-encoded music segmentation. The models listed focus on subphrase-level segmentation or coarser-levels.

4.3.2 A Two-dimensional Classification

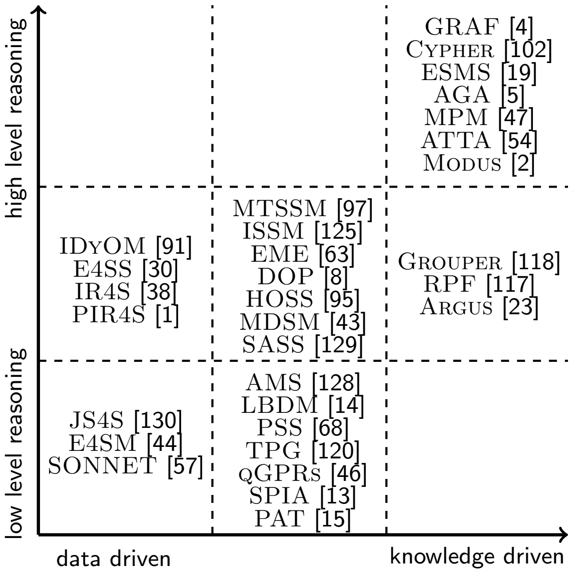

Figure 8: Taxonomy of symbolic segmentation models according to knowledge encoding and processing, as well as level of cognitive processing assumed required for segmentation (which is referred to as 'reasoning' in the figure for brevity). Model abbreviations and acronyms corresponds to those defined in Table 5.

In Figure 8 we present a two-dimensional classification of symbolic segmentation models according to knowledge encoding and processing, as well as level of cognitive processing assumed required for segmentation (which is referred to as 'reasoning' in the figure for brevity). In the figure models listed in Table 5 are grouped into quadrants in accordance to the following criteria:[17]

ordinate axis classification

The knowledge-driven and data-driven classes are used to discriminate models in respect to the strategies used to acquire, encode, and manipulate the knowledge required to segment music. Our criteria to rank models in the ordinate axis is:

- *knowledge acquisition*, i.e. whether automatic (−) or manual (+).
- *domain knowledge dependence*, i.e. if explicit encoding of music theory or music perception (+), e.g. by using production rules, was used in the model. Also whether domain information is assumed available (+), e.g. if the average segment length is assumed known a priori.
- *processing*, i.e. whether grammar-based (+), distance-based (±), statistics-based (−) approaches where used to process the data.

(we mark with a '+' if its a knowledge-driven characteristic, with a '−' if its a data-driven characteristic, and with a '±' if its somewhere in between.)

[17] All models within a quadrant comply with the same criteria, the relative position of model acronyms within a quadrant was chosen only to improve ineligibility and should not be seen as meaningful.

abscissa axis classification

Our classification along the abscissa sorts models in respect to the level of cognitive processing assumed necessary for the perception of segments in music. Lower-level models support the view of segmentation as a mostly universal, style-independent phenomenon. Conversely, higher-level models see segment perception as a (unconscious) reasoning process.

In the figure models are sorted from low to high levels of cognitive processing. Our criteria to rank models in the abscissa axis is:

- *modelling more than one cue*, i.e. the model attempts to combine influence of more than one cue (+).
- *multiple sources of information*, i.e. the model uses both stimulus (local and non-local) and schema related information (+).
- *adaptivity*, i.e. the model combines information from different attributes and different cues in a context-aware fashion at run time (+).

(we mark with a '+' if its a high-level characteristic.)

4.3.3 Taxonomisation Based on Cue Modelling

In Table 6 we classify models in respect to the perceptual cue(s) of Figure 4 they attempt to model, as well as the techniques used for modelling these cues. The cue classes in Table 6 are: difference (gap), difference (contrast), sameness, closure, continuation. We have also made a class 'others' to group models which have a different modelling strategy.

In the following we describe the surveyed segmentation models in respect to the aforementioned classes. We also have a class for models that handle multiple cues simultaneously.

Difference (gap) Modelling

$$\text{difference (gap)} \left\{ \begin{array}{c} \text{distance metrics} \\ \text{\& quasi metrics} \end{array} \right. \left\{ \begin{array}{l} \text{qGPRs [46], LBDM [14], AMS [128],} \\ \text{PSS [68],TPG [120], RPF [117],} \\ \text{MDSM [43],ESMS [19], Cypher [102],} \\ \text{Grouper [118], ATTA [54]} \end{array} \right.$$

These models have focused on the segmentation of melodies into phrases, and in some cases sub-phrases.[18] They most often approach melody segmentation by proposing quantifications for grouping principles of Gestalt psychology, generally proximity and similarity. However, most often the focus in on proximity (the emphasis is on the segregative rather than the unifying aspects of the principles).

Description: Difference (gap) models hypothesise that segment boundaries are cued by abrupt changes in a melody's dynamic evolution. Hence, most models of gap cues refer to the task of segmentation as 'discontinuity' detection. For discontinuity detection melodies are normally

[18] Quite often the music theory terms 'phrase' and 'subphrase' have been explicitly avoided, e.g. Tenney and Polansky [120] refer to segments as 'clangs' or 'sequences', and Cambouropoulos [16], who even avoids talking about segments altogether, speaks instead of models that detect 'local melodic boundaries'. Nevertheless, the models proposed in these studies have been **tested** in boundary annotated melodies which contain segments that comply with our definition of phrase and subphrase (see §2.4.1).

Model		Modelled cues	Modelling technique
TPG	[120]	gap	distance measure
AGA	[5]	gap, sameness, others	distance measure, grammars
GRAF	[4]	gap, sameness, closure/continuation	distance measure, grammars, frames
ESMS	[19]	gap, sameness	grammars
Cypher	[102]	gap, others	distance measure, grammars
RAAM	[65]	others	neural networks
LBDM	[16, 14]	gap	distance, similarity measure
SPIA	[13]	sameness	string pattern search
MPM	[47]	others	grammars, neural networks
RPF	[117, 21]	gap, sameness	similarity measure, grammar
PSS	[68]	gap, sameness	distance measure
Grouper	[118]	gap, idiomatic template, others	distance measure, dynamic programming
DOP	[8]	idiomatic templates	probabilistic grammars
ISSM	[125]	gaps, similarity	fuzzy neural network
MDSM	[43]	gap	distance measure
E4MS	[44]	closure/continuation	information theory
SONNET	[57, 58]	others	self organising maps
EME	[63]	idiomatic templates	information theory, optimisation
qGPRs	[46]	gap, sameness	distance measures
Modus	[2, 3]	gap, sameness	grammar
ATTA	[53, 54]	gap, sameness, others	distance measures, grammars
Argus	[23]	contrast	distance measure
PAT	[15, 18]	sameness	string pattern finding
IDyOM	[91, 94]	closure/continuation	markov models, information theory
IR4S	[38]	closure/continuation	information theory
JS4S	[130]	contrast	probabilistic distance measure
AMS	[128]	gap	distance measure
HOSS	[95]	sameness	genetic algorithms
PIR4S	[1]	closure/continuation	information theory
E4SS	[30]	closure/continuation	neural networks
MTSSM	[97]	sameness	string pattern extraction
SASS	[129]	sameness	shortest path finding

Table 6: List of surveyed models with information of modelled cues and modelling technique

represented as a sequence of intervals rather than a sequence of notes. Localising points of discontinuity often boils down to a search for large intervallic distances among smaller ones (i.e. local maxima) in different intervallic attribute parametrisations of a melody, often pitch intervals, inter-onset-intervals, or inter-offset-intervals. The temporal context defined to search for maxima commonly comprises at most three intervals (or equivalently four consecutive notes), hence the term 'local'.[19][20] Most measures of intervallic distance used for discontinuity detection are based on the L1-norm [120, 46, 16, 14, 68, 128, 55].

Discussion: At present LBDM and Grouper have performed best in comparative studies. These two models rely on discontinuity detection.[21] However, it has been observed that, when targeting phrase-level segmentation, these two models, as well as discontinuity detection models in general, have a tendency to over segment the input melodies. Moreover, recent empirical studies have questioned the relevance [124, 126, 100] of pitch related discontinuities to segment boundary detection, which is generally considered of prime importance in discontinuity detection models.

The approach to modelling discontinuity detection taken by most models can be criticised from

[19] The only exception to the local constraint described is MDSM [43], which measures pitch discontinuity as low 'cohesion' between all possible pitch intervals within a short-term memory window, rather than only between adjacent note events.

[20] The experiments on context size conducted in [127] suggest that longer contexts did not improve phrase-level boundary detection in pop melodies. In fact, it was observed that in some cases taking longer contexts was even detrimental to the performance of discontinuity detection models.

[21] Grouper does not only rely on discontinuity detection (it has a rule based on metric parallelism and a rule based on preferred segment length). However, the experimentation and results presented in [89, 127] suggest that the model's performance is driven by its discontinuity detection component.

three perspectives: (1) in respect to assumptions related to the choice of distance measure, (2) assumptions related to cross dependencies between attribute spaces, and (3) assumptions on how to aggregate discontinuities detected in different attribute dimensions. Regarding (1), intervallic distances are often measured using the L1 norm, hence assuming interval spaces are symmetric and non-directional. Yet, symmetry might be an oversimplification of how humans perceive intervallic distances in a loudness space (as mentioned in [16] the loudness interval perceived when passing from *pp* to *ff* should be larger than passing from *ff* to *pp*). Also, non-directionality might be inadequate to model the perception of intervallic distances in pitch spaces (experimental evidence has shown that human estimates of pitch interval size are larger for descending intervals than for ascending intervals [105]). Regarding (2), most models seem to assume total absence of cross dependencies in the perception of interval size between different attribute dimensions. However, there is experimental evidence showing such correlations are actually common, e.g. in [104] it is shown that changes in timbre can expand or contract the perceived size of pitch intervals for both musically untrained and trained listeners. Regarding (3), discontinuity detection models often assume the combination of discontinuities detected in different attribute dimensions can be achieved using a linear, non-adaptive model. That is, combination models consist of a weighted sum of the discontinuity profiles obtained for each attribute dimension, where the weights (representing the relative salience of different melody parametrisations) are set at initialization and remain constant throughout the analysis.[22] Yet, experimental studies suggest attribute combination might be non-linear and moreover vary in time [124, 126].

4.3.4 Difference (contrast) Modelling

$$\text{difference (contrast)}\begin{cases}\text{distance metrics}\\ \text{\& quasi metrics}\end{cases}\begin{cases}\textsc{Argus}\ [23]\\ \text{JS4S [130]}\end{cases}$$

To the best of our knowledge, only two models, Argus and JS4S, have pursued segmentation by modelling contrast related cues. These models have focused on the segmentation of polyphonic pieces into form level sections (JS4S), and 'tonally stable' passages (Argus).

Description: Contrast detection models are conceptually similar to models of discontinuity detection, employing distance measures to quantify difference. However, they differ in two respects: (1) the length of context needed (normally longer than 4 notes), and (2) the type of attributes used (normally of higher description level than intervals).

Regarding (1), Argus employs a sliding window approach with a minimum window size of 16 notes, yet in experiments the model performed best with larger window sizes (32-48 notes). In the case of JS4S, a global windowing approach is used. That is, a binary split at each time point of the musical sequence is performed, and all past and future note events comprising the piece are used to measure contrast.

Regarding (2), both JS4S and Argus consider segment boundaries occur when there is a significant break in the 'stability' of pitch content. In Argus stability breaks are taken to occur if the distance between the tonal centres of the left and right sections of the sliding window is large. The tonal centers are estimated using a geometric model of tonality (see [22]). In JS4S stability break are taken to occur if the (Jensen-Shannon) divergence between the pitch-class distributions of the left and right global windows is maximal.

Discussion: JS4S was tested on a classical piano sonata. The test showed a fair degree of correspondence between the boundaries predicted by JS4S and a form analysis of the sonata. Similarly, Argus was tested on three piano pieces of the romantic period, and the predicted boundaries matched to a great extent the key changes found in the scores. However, none of the models have been tested systematically and thus generalisation capacity is unknown. Moreover, These models

[22] No mechanisms to adapt to context specific circumstances have been proposed. Exceptions being [128, 68], yet none of the two approaches has been systematically tested.

assume a number of high level attributes to be known a-priori (e.g. pitch spelling was assumed known for the experiments with Argus, and tempo was assumed known for the experiments of JS4S). This restricts their applicability in scenarios where supplying these information manually is unfeasible.

4.3.5 Sameness Modelling

- sameness
  - string matching & extraction
    - text based
      - PAT [15], SPIA [13] , RPF [117], AGA [5], MTSSM [97]
    - grammar based
      - ESMS [19], Modus [3]
  - distance measure
    - ATTA [54]
  - optimisation
    - Grouper [118], HOSS [95]
  - similarity matrix
    - SASS [129]

These models have mostly focused on the segmentation of melodies into phrases and subphrases.[23] Some models rely entirely on sameness based cues (e.g. [15, 13, 3, 95]), while others incorporate sameness cues into multi-cue frameworks.

They operate under the assumption that repetitions of melodic material aid the perception of segment structure. In this context a ‘repetition’ is the recognition of a melodic fragment as been ‘similar enough’ to be considered an ‘instance’ of another melodic fragment listened elsewhere in the melody. As discussed in §3.3, some authors suggest the repetition of melodic figures might aid the perception of the *starting point* of melodic phrases [3, 15, 111].

Description: The modelling of melodic repetition for boundary detection has been attempted from a number of perspectives. Some have formulated the problem as a string pattern extraction problem [15, 13], where similarity is assessed within sliding local window, using a representation of inter-onset-interval contours and diatonic pitch interval information. Suffix trees are usually used to compute and store all string patterns. Others have tried optimization procedures (commonly genetic algorithms [95, 96] and dynamic programming [129]), using in some cases only pitch and in others multi-attribute graph representations [71] of melodies. Melodic self-similarity for segmentation has also been modelled using approaches based solely on generative grammars [3, 66].

Discussion: In most cases the testing of these models has been reduced to case studies. Despite successful application to the respective test cases, a number of authors acknowledge that evaluating when a similarity-based approach is determinant to segment structure is not trivial. In some cases authors have suggested combining similarity-based models with discontinuity-detection models [3, 15], yet results have been inconclusive.

4.3.6 Closure/Continuation Modelling

- closure/continuation
  - melodic closure
    - information theory
      - ES4MS [44], E4SS [30]
      - IDyOM [19],PIR4S [91, 94]
      - IR4S [38]
  - idiomatic templates
    - grammars, heuristics
      - Grouper [118],GRAF [4],
      - DOP [8]

The modelling of closure/continuation related cues has been carried out in two respects: melodic closure, and idiomatic templates.

[23] Two exceptions are SASS, which aims to infer form-level segments in polyphonic music (although it must be noted that during preprocessing SASS reduces polyphony to a monophonic representation), and MTSSM which takes as input multi-part polyphonic music.

Closure

Models focusing on musical closure have normally made use of information-theoretic measures [90, 38, 1], hypothesising that segment boundary closure cues might be correlated to the time-varying behaviour of these measures. Different information measures have been used to determine segment boundaries at different granularities, for example, form-level structural boundaries [1, 30] (in melodies and polyphonic pieces, respectively), and phrase-level structural boundaries [87]. Current systems developed following this idea are commonly composed of two main modules, one that scans the piece from beginning to end, estimating at each point a distribution over the set of possible note events, and another that analyses the resulting distributions using information-theoretic measures, aiming to construct a information profile from which to infer segment boundaries.[24] The former can be said to simulate a probabilistic listener, the output distributions a 'snapshot' of the listener's instantaneous belief of music continuation, and the latter a monitoring stage that traces and characterizes the evolution of these beliefs.

In most works using information-theoretic measures, methodologies for the extraction of boundary locations are not explicitly explored, often only presenting graphs that visually expose a relation between the obtained profiles and a given ground truth (as an exception see [87]).

The second limitation is related to data sparsity. Statistical models often assume that predictability in a sequence is strongly determined by context, so they attempt to statistically characterise patterns comprising a given observation (present) a number events preceding it (finite past). In [86], the statistical regularity of these patterns in musical corpora of folk and classical music was shown to be rarely high for contexts longer than 4 or 5 events, making these approaches mostly insensitive to long range dependencies, which in music analysis is naturally suboptimal.

Idiomatic Templates

To the best of our knowledge, the only published approach addressing the phrase-segmentation problem attempting to model idiomatic-temples is [8], which uses the method of Data-Oriented Parsing (a probabilistic grammar approach) to learn melodic phrase structural templates. DOP operates by creating a a phrase structure class for each distinct phrase in a melodic corpus with annotated phrase boundaries (The Essen Folk Song Collection). The phrase classes are soft (i.e. a phrase can be generated by more than one class), and a parsing algorithm is used to compute the phrase class sequence that fits an input melody with maximal probability.

The publication reports the highest segmentation performance to date on a large number of melodies (81% mean-F1 score on 1000 melodies). Yet it operated with absolute melodic attributes (i.e. chromatic pitch and onset-to-offset interval for duration). This raises the question of how much the algorithm is over-fitting to the set. The rules learnt by DOP from the Essen collection might be expected to hold little significance for the analysis of other styles. Given that annotated corpora for symbolic segmentation are at present hardly available, the approach can not be expected to be applicable to mixed melodic corpora in the short term.

[24] Here an 'information profile' is analogous to the `sbp` mentioned in §4.2.3.

Level	Type	Abbreviation	Meaning
low II	basic	`onset`	note onset time
		`offset`	note offset time
		`pitch`	note pitch
		`intens`	note dynamic
		`timbre`	note timbre
mid	rhythm	`rest`	duration of a musical rest: offset-to-onset (ooi) interval
		`dur`	note duration$_1$: onset-to-offset interval
		`ioi`	note duration$_2$: inter-onset interval
		`speed`	note duration$_3$: classification into slow, medium, or fast
		`dur-rat`	duration ratio of e_i relative to e_{i-1}
		`dur-rat-ctr`	contour of `dur-rat`$_i$ relative to `dur-rat`$_{i-1}$ (longer, same, shorter)
	pitch	`cpitch`	note chromatic pitch
		`cp-cls`	chromatic pitch class (chroma under octave equivalence)
		`cp-iv`	chromatic pitch interval of e_i relative to e_{i-1}
		`sl-cp-iv`	"step-leap" (step ($\pm s$), a leap ($\pm l$), or a unison ($\pm u$)) classification of `cp-iv`
		`p-ctr`	pitch contour of `cp-iv`$_i$ relative to `cp-iv`$_{i-1}$ (up, down, same)
		`p-reg`	pitch register classification of e_i relative to central c octave
	density	`v-dens`*	local/global vertical density (number & spacing between simultaneous events)
		`h-dens`	local/global horizontal density (number & spacing between successive events)
	metric	`beat`	*tactus* or main pulse
		`bar`	metric measures
high I		`tempo`	pace of music, in beats per minute (BPM)
	harmony	`chord`*	chord estimated to harmonize note
		`ch-lbl`*	chord label (e.g. major, minor)
		`ch-pos`*	chord position (e.g. root, inversions)
	tonal	`key`	local/global key
		`mode`	local/global mode
		`dpitch`	note diatonic pitch (requires pitch spelling)
		`dp-cls`	diatonic pitch class (assumes octave equivalence)
		`dp-iv`	diatonic pitch interval of e_i relative to e_{i-1}
		`sl-dp-iv`	"step-leap" (step ($\pm s$), a leap ($\pm l$), or a unison ($\pm u$)) classification of `dp-iv`
		`sc-deg`	note scale degree (e.g. I-VII)

Table 7: Descriptive *attributes* of a monophonic musical sequence of events $e = e_1 \dots e_i \dots e_N$ of length N. Here a note event e_i of a sequence e is here defined in respect to a multidimensional attribute space ξ, such that $e_i \in \xi = \{\texttt{onset} \otimes \texttt{offset} \otimes \texttt{pitch} \otimes \texttt{timbre} \otimes \texttt{intensity}\}$, where $\otimes$ is the Cartesian product operator. The attributes listed correspond to those used by the surveyed computational models of segmentation. The abbreviations proposed for sequence attributes at low to high levels of description used by reviewed segmentation algorithms (Tables 8 and 9). The attributes considered might be specific to an event (e.g. `cpitch`), two or four note event patterns (e.g. `cp-iv`, `p-ctr`), or longer spans which can be global (the whole sequence e) or local (a time window of predefined length). The attribute level of description is classified as low II, mid and high I according to the level classes established in [70]. The attributes are classified as low- mid- level descriptors if they are readily available in the symbolic description or can be obtained through simple numerical analysis requiring only a context of local temporal information (e.g. `cp-iv` $= e_i - e_{i-1}$, for $i = 2 \dots N$). The attributes are classified as high-level descriptors if they need to be estimated using more sophisticated computational analysis and require information spanning a larger or global temporal context (e.g. the key estimation algorithm described in [73]). The attributes are moreover classified as to whether they correspond to basic, rhythmic, pitch, density, metric, harmonic, or tonal information. Basic information describes a single note event. All other information might describe melodic sub-sequences of more than one event. In the case of the pitch and rhythm classes, the horizontal line separating the attribute list discriminates *absolute* from *relative* attributes. Finally, attributes with a $*$ are special cases in which the musical surface considered for segment analysis is polyphonic. This is always the case for `v-dens`, but not necessarily so for the harmonic attributes considered (in principle harmony could be estimated from a single melodic line by consider it as "implied").

Table 8: **Knowledge-driven models of symbolic segmentation**. **Fields** (columns from left to right): `Acronym` - by authors/taken from title (see Table 5), `Date` - of publication, `Q` - quantization required, `# of rules` - used for local analysis of boundaries, `Context window` - used by bottom-up heuristic rules, `Attributes` - of the input melody (low and mid level), `Additional` - high level attributes and a priori knowledge, `Pre/post` - processing steps employed, `Cues` - classes used to infer segments, `S` - "boundary strength" at position, `B` - segment at position yes/no (binary), `O` - other, `Description/Interpretation` - what the score/binary output suggests or, in case the output differs from a boundary strength profile, provides a description, `Test set` - used for case study, `Performance` - of model, `Claim` - or assumptive conclusion. **Abbreviations**: tp - true positive (found boundaries), fp - false positive (inserted boundaries), R - Pearson's correlation coefficient, $\overline{R}$ - mean recall, $\overline{P}$ - mean pression, $\overline{F1}$ - mean F-measure, attribute abbreviations are depicted in Table 7. **Symbols**: ⊖ - unspecified, ⊘ - does not apply, > - performs better than, ⋙ - take precedence over, ≈ - performance compares to, ⇆ - complements.

Model		**Input**				**Analysis**				**Output**				**Evaluation**		
Acronym	Date	Encoding	Q	Attributes	Additional	# of rules	Context window	Pre/Post	Cues	S	B	O	Description	test set	performance	claim
TPG [120]	1980	⊖		`cpitch` `dur` `intens` `timbre`		2 main	clang: 2 notes seque: 2 clangs		local		✓		clang/sequence boundaries	3 20^{th} century pieces of classical music	qualitative assessment	all other things equal, TPGs model segmentation.
AGA [5]	1989	MIDI	✓	`cpitch` `onset`	bar & beat chord lexicon harmonic rhythm figuration	4 main	⊖	tonal analysis reduction analysis pattern analysis	local parallelism tonal closure			✓	ranked list of tonally well-formed phrase sets	⊘	⊘	"tonal motion" & parallelism ⋙ GTTM GPRs $2_{a,b}$ & $3_{a,d}$
GRAF [4]	1989	MIDI	✓	`cpitch` `onset`	bar & beat chord lexicon key harmonic rhythm figuration genre	4 main	⊖	tonal analysis matching analysis	local tonal closure schema			✓	ranked list of tonally well-formed phrase sets	English nursery tunes	qualitative assessment	schematic knowledge ⋙ GTTM GPRs $2_{a,b}$ & $3_{a,d}$
Cypher [102]	1992	MIDI		`cpitch` `p-reg` `dur` `speed` `h-dens` `v-dens`	beat chord key phrase length	⊖	⊖	key & chord finder beat tracker harmonic analysis	local tonal closure heuristics	✓			real-time boundary	⊘	⊘	context necessary for phrase-segmentation
LBDM [16]	1997	MIDI		`cp-iv` `ioi` `ooi`		2 main	≦ 4 notes		local	✓			local proximity & identity phrase strength	52 excerpts of [47]	74% tp 49% fp	LBDM ≈ MPM-rule
SPIA [13]	1998	MIDI		`cp-iv` `dur` `onset`		3 main	⊘		parallelism	✓			boundary due to similar patterns	1 pop song	qualitative assessment	SPIA ⇆ LBDM
MPM-rule [47]	1998	⊖	✓	`cpitch` `dur`	chord label tonal function	7 main 6 revision	≦ 5 notes	⊖	local tonal closure	✓			insert comma marking at position	52 excerpts classical music (26 for tuning 26 for testing)	66% tp 34% fp	MPM-rule > MPM-NN
RPF [117]	1999	MIDI	✓	`cpitch` `ooi`		2 main 2 revision	⊖	pattern analysis	local		✓		discontinuity similarity phrases	96 songs Japanese pop	0.77 $\overline{R}$ 0.79 $\overline{P}$ 0.78 $\overline{F1}$	
Grouper [118]	2001	MIDI	✓	`onset` `offset`	bar & beat phrase length	3 main	2 notes	metric analysis	local parallelism		✓		proximity & similarity phrase boundary	65 E4SSC songs	0.76 $\overline{R}$ 0.74 $\overline{P}$ 0.75 $\overline{F1}$	phrase structure in folk largely inferable from quantized rhythmic info
Argus [23]	2006	⊖	✓	`dpitch` `dur`	beat pitch spelling	⊘	≧ 2 bars		local tonal discrepancy	✓			key modulation boundaries	3 romantic period pieces	qualitative assessment	
qGPRs [46]	2004	MIDI	✓	`cpitch` `dur`		4 main	≦ 4 notes		local	✓			GTTM phrase boundaries	4 nursery-rhymes 2 tonal melodies	GPR2b: R = 0.80 GPR2a: R = 0.54 GPR3a: R = 0.14 GPR3d: R = -0.09	only 2a,b needed to explain boundaries of test set
MODUS [2]	2004	MIDI	✓	`cp-iv` `p-ctr` `dur`	beat MPC (see [2])	> 15	⊖	metric analysis MPC (see [2])	local parallelism global			✓	similarity (start) discontinuity (end) phrase boundaries	> 200 western & non-western melodies	qualitative assessment	MODUS > LBDM
ATTA [53]	2004	MIDIXML	✓	`cp-iv` `ioi` `ooi`	beat	8 main	3 notes	metric analysis reduction analysis	local parallelism segm. length	✓			combined GTTM rules for phrase boundaries	100 extracts classical melodies	0.67 $\overline{F1}$	
PAT [18]	2004	MIDI		`p-ctr` `dur-rat`	tempo pitch spelling	4 main	10-12 secs	pattern analysis	parallelism	✓			similarity-based boundary strength	5 extracts of classical music	qualitative assessment	PAT+LBDM > LBDM
AMS [128]	2008	MIDI	✓	`cpitch` `p-ctr` `onset` `offset` `intens`		2 main 2 revision	4 notes	adjustable weight analysis	local global	✓			local discontinuity boundary strength	⊘	⊘	⊘

Table 9: **Data-driven models of symbolic segmentation**. Layers (row divisions): Upper - focus on motifs-level/reduced representations, Middle - focus on phrase-level, Bottom - focus on form-level or other, **Fields** (columns from left to right): Acronym - proposed by authors/derived from title (see Table 5), Date - of publication, Type - underlying learning algorithm, Learning type - (Un)Supervised | Incremental/Static, Learning source(s) - current piece and/or corpus of melodies, Learning corpus - details of the corpus (if applies), Pattern statistics - ngram patterns for which statistics are collected (if applies) also mentions if n-grams, fixed upper-bound, or adaptable upper-bound techniques are used, Encoding - used in publication, Q - ✓if quantization is required, T - **P**olyphonic|**M**onophonic texture, Attributes - of the input melody (higher level attributes in upper-case), S - "boundary strength" at position, B - segment at position yes/no (binary), O - other, Description - what the score/binary output suggests or, in case the output differs from a boundary strength profile, provides a description, Test set - used for case study, Performance - of model, Claim - or assumptive conclusion. **Abbreviations**: tp - true positive (found boundaries), fp - false positive (inserted boundaries), R - Pearson's correlation coefficient, $\overline{R}$ - mean recall, $\overline{P}$ - mean pressission, $\overline{F1}$ - mean F-measure, PG - probabilistic grammars, MMM - mixed order Markov models, VLMM - variable length Markov models, MM - Markov models (of fixed length), ME - Maximum Entropy, RNNs - Recurrent neural networks, NNs - neural networks. Attribute acronyms are depicted in Table 7. **Symbols**: ⊖ - unspecified, ⊘ - does not apply, > - performs better than, ≈ - performance compares to, ⊗ - interacting with (see [85]).

Model							**Input**				**Output**				**Evaluation**		
Acronym	Date	Type	Learning			Pattern	Encoding	Q	T	Attributes	S	B	O	Description[25]	Test set	performance	claim
			Type	Source	Corpus	statistics											
RAAM [65]	1995	NNs	U/S	piece	25 melodies (18 unique 7 variations)	⊘	⊖	✓	M	cp-cls p-ctr BEAT BAR			✓	reconstruction of salient fragments	3 melodies	qualitative assessment	RAAM reconstructions of structurally important fragments parallel those of human improvisers
EME [63]	2004	ME	U/S	corpus	2323 songs Hungarian folk	uni-grams bi-grams	MIDI		M	cp-iv			✓	stable motifs	⊘	⊘	stable motifs in folk correspond to pentatonic scales
MPM-NN [47]	1998	NNs	S/S	piece	⊘	⊘	⊖	✓	M	cpitch dur	✓			insert comma marking at position	52 excerpts classical music (26 for tuning 26 for testing)	39% tp 69% fp	MPM-rule > MPM-NN
DOP [8]	2002	PG	S/S	piece corpus	5251 songs from E4SSC	fixed upper bound (five-grams)	EsAC	✓	M	cpitch dur		✓		⊘	1000 folk songs form the E4SSC	0.81 $\overline{F1}$	previous exposure relevant for boundary identification
E4MS [44]	2002	MMM	U/I	piece	⊘	fixed upper bound (tri-grams)	MIDI		M	cp-iv dur-rat	✓			melodic predictive uncertainty	1 20^{th}-century performed flute piece	0.69 $\overline{P}$ 0.79 $\overline{R}$ 0.73 $\overline{F1}$	⊘
SONNET [57]	2003	NNs	U/I	piece	⊘	⊘	MIDI	⊖	M	cp-iv ioi BEAT			✓	reconstruction of melodic fragments	10 pop melodies	qualitative assessment	fragments learnt by SONNET correspond to melodic phrases
MDSM [43]	2003	⊘	U/I	piece	⊘	⊘	MIDI		M	cp-iv onset TEMPO	✓			local change in melodic interval density	7 mixed genre melodies	0.93 $\overline{P}$ 0.88 $\overline{R}$ 0.91 $\overline{F1}$	MDSM ≈ LBDM Yet, MDSM more precise
IDyOM [94]	2006-07	VLMM	U/I	piece corpus	900 tonal melodies	adaptable upper bound	MIDI	✓	M	cpitch cp-iv cp-cls cp-iv ⊗ cp-cls rests	✓			melodic surprise	2 minimalist music pieces	qualitative assessment	⊘
IR4S [38]	2006	MM	U/I	piece	⊘	tri-grams	MIDI		P	cpitch timbre[26]	✓			melody "interestingness"	4 performances of a Bach prelude	qualitative assessment	IR of symbolic and audio derived attributes relevant to interesting point detection
JSD4S [130]	2007	MM	U/I	piece	⊖	uni-grams	MIDI		P	cp-cls	✓			global contrast	1 classical piano sonata	qualitative assessment	inferred segments are similar to those in form analysis of the piece by experts
PIR4SS [1]	2009	MM	U/I	piece	⊘	bi-grams	MIDI	✓	M	cpitch	✓			melody "interestingness"	2 minimalist music pieces	qualitative assessment	PIR4SS > LBDM PIR4SS > GPR3a for *form-level* segments
E4SS [30]	2010	RNNs	U/S	piece corpus	6 J.Haydn Quartets	⊖	MIDI	✓	P	cpitch onset timbre	✓			sense of tension	1 J.Haydn Quartet	qualitative assessment	H-profile correlates to music-theoretic analyses of tension & release

5 Evaluation

Quantitative evaluations of symbolically encoded music segmentation have focused on monophony (most often melodies). Thus in this section we describe evaluation of melody segmentation. First, in §5.1 we discuss the main strategies proposed to evaluate melody segmentation models within music information retrieval. Next, in §5.2 we discuss common metrics used for evaluation. Finally, in §5.4 we describe the comparative studies carried out to date and discuss their results.

5.1 How to Evaluate Segmentation Models

To evaluate the output of segmentation models the question of *what constitutes a valid segmentation of music?* arises. In MIR this question has been commonly dealt with in two ways, each with its own pros/cons.

A COMPARING TO HUMAN ANNOTATIONS
In this setting the evaluation seeks to assess how 'perceptually valid' the output of a segmentation model is. Thus, the original question is generally reformulated to: *are automatically derived segment boundaries similar to those perceived by humans?* Generally, this type of evaluation compares the output of segmentation models to human annotated boundaries for a particular piece of music. This requires availability of a music corpus where each piece has been annotated with segment boundaries by human listeners. While giving a more or less direct link to human perception, this method of evaluation is constrained by the diversity of music in the boundary annotated corpora, the number of human annotators per piece, and level of detail of the annotations (e.g. location of a boundary, musical cues that influenced the perception of a boundary, 'notoriety' of the boundary, etc).

B USING SEGMENTATION MODELS IN AN MIR TASK
This type of evaluation seeks to assess the usefulness/relevance of a given segmentation model in practice. Thus, the original question is reformulated to: *What is the influence of automatically derived segment boundaries in a specific MIR task?* (e.g. query-by-humming). Generally, this type of evaluation is carried out by having models segment a number of pieces, and then using the output segments as indexes for classification/other MIR tasks. This type of evaluation can measure the effects of segments in the task at hand (e.g. how segmentation affects ranking in a classification task), or the effect that one perceptual process may have in another process (e.g. the role of segmentation in similarity computation). However, in this type of evaluation is often non-trivial to distinguish between effects due to segmentation and effects related to the evaluation chain.

5.2 Evaluation of Melody Segmentation Models

This section describes the steps taken to compare boundaries identified by humans to boundaries identified by segmentation models, when a binary vector encoding of boundary positions is used.

5.2.1 Comparing Segmentations of a Melody using Binary Vectors

To make the output of segmentation algorithms and human annotations comparable, all outputs/annotations are usually reduced to binary classification of each event in the melody (boundary at position yes/no). We can hence encode the output of a model that generates boundary markings, henceforth a *prediction*, as a vector $\mathbf{p} = (p_1, \ldots p_i, \ldots, p_N)$, and the phenomenal data collected, henceforth our *ground truth*, as a vector $\mathbf{g} = (g_1, \ldots g_i, \ldots, g_N)$, where $p_i, g_i \in \{0, 1\}, \forall i \in \{1, ..., N\}$.

Once this procedure is carried out, we can formulate the problem of comparing the predicted segmentation to a ground truth segmentation as the *computation of the similarity between two binary vectors.* Below we review ways in which binary vectors are created for prediction and ground truth in comparative studies of melodic segmentation.

Preparing the Prediction:

Processing of model outputs is needed when boundary vectors indicate boundary presence using some form of scoring or likelihood mechanism, so that $\mathbf{b} \in [0,1]^N$. In this case, processing is required to obtain binary boundary classifications. The most common assumption to convert boundary scoring into binary form is to take peaks in **b** as boundary indicators. Hence, converting model outputs into binary vectors is generally carried out using heuristic peak selection methods. The heuristics used in comparative studies of melody segmentation are described in Appendix C.

Preparing the Ground Truth:

The preparation of a ground truth set, i.e. a vector **g** for each melody in the test database, commonly consists of the following steps:

1. *Test Database Compilation*: Compilation of a representative sample of melodies or melody excerpts, for one or more musical genres of interest.

2. *Gathering Annotators*: Gathering a group of annotators, ideally composed of equal-size subgroups representing varying degrees of musical expertise. Have annotators indicate the temporal instants where they perceive segment boundaries.

3. *Inter-Annotator Agreement*: Measuring the degree to which annotators performed segmentation consistently, so that annotated boundary locations can be used as a reference for comparison.

4. *Boundary Selection*: Selecting the annotated boundary locations that are more likely to represent a 'valid' segmentation of each melody in the test database. The selected boundary annotations are finally encoded as binary flags in **g**.

The column field *Test Corpora* of Table 10 lists information describing how steps 1-3 described above were conducted in comparative studies of melody segmentation. The methods employed in comparative studies to carry out step 4, i.e. convert the human annotations into binary boundary decisions, are described in Appendix D.

5.3 Comparing Binary Vectors

Many measures have been proposed to compare equal length binary vectors. As an example, [24] presents a survey that reports on 69 different measures employed in various fields to determine the similarity/distance between binary vectors. However, as the survey indicates, the relevance of a specific measure depends on the task for which the comparison is needed and on the nature of the data. The measures described in [24] use as input statistics collected for co-occurrences between elements of **p** and **g**. Element co-occurrences are classified as either *true positives* (also *hits*), *true negatives* (also *correct negatives*), *false positives* (also *false alarms*), or *false negatives* (also *misses*). The statistics collected for this co-occurrences and their use in evaluation measures is described in Appendix E.

5.3.1 Evaluation Measures for Melodic Segmentation

The measures for binary vector comparison used in comparative studies of melody segmentation models are described in detail in Appendices E.2 and E.1.

In melodies the proportion of segments boundaries to note event positions is generally small, so our evaluation measure needs to minimize (or completely neglect) the influence of true negatives, otherwise a model that predicts no boundaries would still obtain a relatively high score. Hence, in most comparative studies [88, 89, 127, 121] the use of *precision*, *recall*, and $F1$ has become a standard, since these measures do not consider true negatives for the evaluation of segmentation.

5.3.2 Problems with Common Evaluation Measures

Most of the measures proposed for comparison of binary vectors, including the F1-measure, operate under the assumption that the dimensions of each vector are independent. It is only then that the statistics on co-occurrences can be taken as a complete and sufficient description of the binary vectors to compute similarity/distance [72]. In melody segmentation we deal with a sequence in time, so that event positions can not be considered independent dimensions. In case independence is assumed, a prediction vector $\mathbf{p} = (1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0)$ compared to a ground truth $\mathbf{g} = (1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0)$ could receive a relatively high score, even though is very rare for humans to perceive boundaries only one event apart, or in contiguous events. In other words, by assuming independence between boundaries aspects such as segment length and boundary position within the melody are discarded from the evaluation.

5.4 Comparative Studies of Melody Segmentation Models

In this section we first outline the results of evaluations of single models that have used large test sets. Subsequently we discuss comparative studies of melodic segmentation. This section feeds from Tables 8, 9, & 10.

5.4.1 Single Model Evaluation

Ahlbäck [2], collected and examined approximately 200 western and non-western melodies, and Bod [9] tested his DOP model in 1000 melodies of the Essen Folk Song Collection. The examination carried out by Ahlbäck is qualitative and on a case-by-case basis, so it is difficult to summarise and does not allow to get an overall idea of the performance of the model. Conversely, Bod's DOP evaluation is mostly quantitative, reporting on a mean-F1 of 0.81. This is the highest performance on large-scale melodic corpora to date. However, to hit the 0.81 mark, DOP was trained on ∼85% of the corpus (5251 melodies), using moreover absolute melody attributes consisting of chromatic pitch and quantised duration. If ∼50% (3000 melodies) of the corpus is used for training, the performance of DOP drops to 73.2%. This drop in performance suggests that the model might have been overfitting the corpus. Since DOP has not been included in comparative studies, its generalisation capability is unknown.[27].

5.4.2 Comparative Studies

Below we group comparative studies in small-scale and large-scale. The former refers to studies that use small datasets for evaluation, focusing on stylistic diversity and on statistically significant number of human annotators to determine the ground truth (∼20 annotators per melody). Conversely, large-scale studies refer to studies where large melodic corpora have been used to compare the performance of models. In these studies the focus is on generalisation capability of the models.

In the following each case is discussed in turn.

`Small-scale:` The small-scale studies considered here are [121, 127, 88, 12]. These studies consider small test sets (6-15 melodies) comprising a mix of styles (folk, pop, classical, jazz). The studies include segment boundary annotation by participants with musical training, and by degree-level musicians. Human segment annotations are used as a ground truth for comparison. Inter-annotator agreement of was ranked high in [127], and low in [121, 88, 12]. In [88] an effort was made to find the optimal clusters of annotation agreement to use it as a ground truth, as a result the mean F1 values reported are higher than in the other studies. The number of models evaluated are 2 in [121], 6 in [127], and 6 in [88]. Performance evaluations commonly have at top ranking

[27] In [127] The implementation of DOP that was used corresponded to that of [42], which, as we have come to learn from the authors of [42], was made only for demonstration purposes and hence can not be taken as reliable enough for comparison.

Grouper and LBDM. Grouper's performance ranges between mean $F1 = 0.73 - 0.83$ when folk and pop melodies are considered, and drops to mean $F1 = 0.61$ when classical and jazz melodies are included. LBDM, on the other hand, performs with a mean $F1 = 0.66 - 0.78$ when folk and pop melodies are considered, and drops to mean $F1 = 0.54$ when classical and jazz melodies are included.

`Large-scale:` The large-scale studies surveyed here are [89, 83]. These studies consider test sets of 1004 pop and 1705 (EFSC) vocal folk melodies, respectively. The evaluation in each study has different goals, and so they will be reviewed in turn.

In [83] segment boundary estimation is evaluated in a MIR scenario: melody retrieval based on index terms. In the evaluation the index terms correspond to the melodic segments bounded by the estimated boundary locations. The melody retrieval engine was based on a vector space model, using the $tf \cdot idf$ measure to compute similarity between melodic query and stored melody. Performance was evaluated as the percentage of queries with matching documents within the first k positions (with $k \in \{1, 5, 10, 20\}$), and a special percentage class representing 'not-found' queries. Experimental results showed that simple ngram extraction constituted the best choice of index terms, outperforming the indexes obtained using the TPG and LBDM segmentation models. The ngram indexes where obtained by extracting all note sequences of lengths 1-4 notes, allowing sequences to overlap. The inferior performance obtained using LBDM and TPG based indexes was hypothesized as the result of local mismatches between query and stored melodies. In that respect ngrams indexes proved more robust due to the abundance of possibilities resulting from allowing ngrams to overlap.

In [89] segment boundary estimation is evaluated comparing predicted boundaries to annotated boundaries using the precision, recall, and the F1 measure. The test set corresponded to the `Erk` database of the Essen Folk Song Collection, which has phrase boundary annotations for each melody. (In the following mean-F1 values are given in parenthesis.) Experimental results show that Grouper (0.66) and LBDM (0.63) perform best. They are followed by the IDyOM model (0.58) and the quantification of the GTTM rule GPR2a proposed by [46] (0.58). Statistical tests show that all pairwise differences between models F1 performances are significant, except those between LBDM and GPR2a. Models IDyOM and GPR2a exhibited the highest precision performance, while Grouper and LBDM had the highest recall. The high average F1 performance of GPR2a confirms the large influence that rests have as boundary predictors in the germanic folk songs collected in the Essen collection.

A logistic regression based hybrid model (including Grouper, LBDM, IDyOM, and GPR2a), shows an average F1 performance equal to that of Grouper. However, statistical signed tests showed that the hybrid model achieved better F1 scores in significantly more melodies than any other model included in the study. This suggests that, although marginally, there is a benefit in considering not only different cues, but also different quantification/modelling of a specific type of cue (e.g. temporal gap distances as computed by Grouper and LBDM).

Table 10: **Comparative Studies of Melodic Segmentation**. Fields (columns from left to right): `Study - Authors` and `Year` of study, `Models - Model` Acronyms (specified in Table 5) and parameter `Setting` used, `Test Corpora - Melodies`: F: # of files & format, G: genre(s), T: texture, C: compilation procedure, O: other details, `Ground Truth`: A: # of annotators, E: annotator's level of expertise (DL - degree level, M - musical training, N - non musician) IAA: inter-annotator agreement, `Results - Rank` of 5 top models, `Performance` obtained, `Measure` of performance, `Range` between performance of the best and the worst model, Symbols: $\ominus$ - unspecified, R - Pearson's correlation coefficient. Abbreviations: R - Pearson's correlation coefficient, $\overline{R}$ - mean recall, $\overline{P}$ - mean pression, $\overline{F1}$ - mean F1-measure,

Study		**Models**		**Test Corpora**		**Results**			
Author	Year	Models	Setting	Melodies	Ground Truth	Best models	Performance	Measure	Range
Thom et al. [121]	2002	Grouper LBDM	$k = 0.5$ $\ominus$	**set 1** F: 10 MIDI S: folk, classical, jazz T: Monophonic C: Manual	A: 19 E: DL & M IAA: low	**set 1** Grouper LBDM	0.61 0.54	$\overline{F1}$	0.54 – 0.61
				set 2 F: 2612 MIDI files S: folk C: Subsets of the E4SSC	A: ESFC annotations E: DL IAA: high	**set 2** Grouper LBDM	0.62 0.50	$\overline{F1}$	0.50 – 0.62
Orio et al. [83]	2005	fix-win NGRAMs TPG LBDM	= 3 notes ≦ 5 notes clangs $k = 0.5$	F: 1004 MIDI S: pop T: Polyphonic O: made use of a melody extraction algorithm	36 annotated queries (tested on a VSM -based retrieval engine)	NGRAMs fix-win TPG LBDM	12.6 15.1 5.6 2.8	% of queries with correct document at top rank	2.8 – 12.6
Wiering et al. [127]	2009	TPG Grouper LBDM MODUS DOP[28] IDyOM	segs length 8 max(4,3) low, end $k = 0.6$ $\ominus$	F: 10 MIDI S: pop T: Monophonic C: sampled (internet)	A: 20 E: DL & M IAA: high	Grouper LBDM IDyOM MODUS	0.73 0.66 0.45 0.42	$\overline{F1}$	0.23 – 0.73
Bruderer et al. [12]	2009	GPRs Grouper LBDM	$\ominus$ $\ominus$ $\ominus$	F: 6 MIDI S: pop T: Monophonic C: manual	A: 21 E: DL & M IAA: low	LBDM* GPRs* Grouper* (* + timbre rule)	0.50 0.45 0.40	R	0.20 – 0.50
Pearce et al. [88]	2010	Grouper LBDM GRP 2a,b,d & 3a TP PMI IDyOM	default $k = 0.5$ $k = 0.5$ $k = 0.25$ $k = 0.2$ $k = 1$	F: 15 MIDI S: folk, pop T: Monophonic C: Manual	A: 25 E: DL IAA: low O: used clusters with high IAA	Grouper LBDM IDyOM GRP2a	0.83 0.78 0.64 0.58	$\overline{F1}$	0.11 – 0.83
Pearce et al. [89]	2010	Grouper LBDM GRP 2a,b,d & 3a TP PMI IDyOM Hybrid	$k = 0.5$ $k = 0.5$ $k = 2.5$ $k = 0.25$ $k = 0.2$ $k = 2$	F: 1705 MIDI S: folk T: Monophonic C: Subset of the E4SSC	A: E4SSC annotations E: DL IAA: high	Hybrid Grouper LBDM GRP2a IDyOM	0.66 0.66 0.63 0.58 0.58	$\overline{F1}$	0.22 – 0.66

6 Conclusions

In this paper we have presented a survey of computer models developed to identify cognitively plausible segments in digital music files, with a focus on models segmenting melodies into phrases, and whole polyphonic pieces into sections. In our survey we have discussed segmentation models processing symbolic input, and have centred our discussion on a fundamental aspect of segmentation: *boundary detection* (i.e. localising the time points bisecting contiguous segments).

Since the conclusions of this survey are multiple we discuss them separately: In §6.1 we summarise general conclusions, in §6.2 we outline specific areas in which current computer segmentation models can be improved, and finally in §6.3 we draw conclusions and outline future work possibilities in respect to evaluation methodologies of segmentation.

6.1 General Conclusions

1. The automatic segmentation of symbolic encodings of music into phrases and sections has been an active topic of research for over three decades. Yet, relative to research of musical audio segmentation, the number of models and overall research effort that has gone into segmenting symbolic music encodings is small. Given the fundamental nature of segmentation for music perception and cognition [69, 81, 56], and hence for music information processing, research efforts towards modelling music segmentation should integrate both audio and symbolic domains. As a step on this direction, in this paper we present a *boundary-cue-centred taxonomisation* of segmentation models, and furthermore extensively discuss aspects related to *terminology* divergence across domains. The organisation and description of models in this survey reveals possibilities of integration not only between models designed for symbolic segmentation, but also in respect to models designed for audio segmentation.

2. Most models proposed for the segmentation of symbolic encodings of music have been designed-for and tested-on melodies. Results of comparative studies show that, at least for the segmentation of melodies into phrases, a fully automatic solution to the problem of segment boundary detection has not been yet achieved. Thus, in §6.2 and §6.3 we propose alternatives to improve model performance in phrase and section level segmentation of monophonic pieces and melodies. We also discuss what would be needed to move into segmenting polyphonic pieces, and suggest ways in which polyphonic segmentation could be tackled.

6.2 Suggestions for Computational Segmentation Model Design

`Combining information sources in phrase and form level segmentation:` The analysis of performance results in systematic evaluation studies points to the inability of current phrase and form level segmentation models to make 'musically sensitive' decisions in respect to (a) the situations in which a given musical attribute (harmony, pitch intervals, etc.) should take precedence over others, and (b) the situations in which a given cue model (discontinuity detection, contrast detection, repetition detection, etc.) should be employed. Hence, development of inferential engines for *context-aware* decision is urgently needed.

`Cognitive plausibility and adaptability in phrase and form level segmentation:` Symbolic music segmentation models rely mainly on local segmentation strategies when targeting phrase-level segments, and off-line segmentation strategies when targeting form-level segments. However, human music perception is a dynamic process with access to both local and global information, and segment perception does not seem to be exempt of these characteristics. Models relying only on local information are missing cues related to both longer time spans within the music being listened to (such as similarity between musical entities located far apart in a piece), and the effect of previously listened music (such as phrase and form level structure templates that a listener might have learnt during his/her musical listening experience). On the other hand, models that use purely off-line processing mechanisms (such as similarity matrices), might violate human memory constrains (e.g. a similarity matrix might be said to represent a listener with eidetic memory, i.e. a ‘perfect’ recall of all similarities present in a piece of music). Therefore, some alternatives that might prove beneficial for future modelling of segment perception are: (1) models that are able to make use of global and local information simulating real-time listening of music, (2) models that are able to dynamically keep track of identified segments and include this information in the identification of subsequent segments, (3) models that are able to re-evaluate past events in the presence of new information (react in retrospection), and (4) models that are able to ‘forget’ information as they carry out a segmentation analysis.

`High-level descriptors in phrase level melody segmentation`:
Models segmenting symbolically encoded melodies commonly do not use information from other structuring processes, e.g. metric (bar, beat), or harmony/tonal (cadence, key). That is, with some exceptions [5, 4, 47], computational models of melody segmentation rely mainly on low-level description attributes when processing melodies. However, the modest performances obtained in comparative studies using large test databases suggests that the inclusion of higher level descriptors needs to be re-visited. Moreover, opening to the use of higher-levels of description might give ways to tackle polyphony in symbolic segmentation, e.g. for certain types of music polyphony could be reduced to a melody-plus-harmony representation.

6.3 Suggestions for Computational Segmentation Model Evaluation

6.3.1 Evaluation Campaigns for Symbolic Segmentation

A relatively high number of models have been proposed to tackle the problem of segmentation in symbolic encodings of music (> 30), with a large part of these models focusing on segmenting melodies into phrases. So far evaluation of symbolic music encoding segmentation has covered only a small portion of the models proposed. Evaluation campaigns, such as the MIREX structural segmentation track,[29] are thus urgently needed to both stimulate development and have a better idea of the performances of these (and new) models.

6.3.2 Annotated Corpora for Evaluation

Below we discuss future directions regarding the current state of development and music coverage of boundary annotated corpora.

`In respect to phrase-level segmentation`: at present large annotated corpora are mainly comprised of vocal folk songs (largest corpora available contains ∼6000 melodies). There is hence an urgent need to develop corpora of mixed styles, instrumentation, and textures (polyphonic, homophonic, and monophonic music).

`In respect to form level segmentation`: recent annotation studies on audio data have developed corpora of mixed styles (pop, jazz, classical, and world music) and in sizeable quantities (largest corpora available contains a total of ∼1300 pieces). However, similar corpora for symbolic music encodings are at present non-existent.

For both phrase and form level annotations it is also necessary to expand the number of annotations per melody/piece, given that in most cases boundary and label annotations are made at most by two human listeners.

6.3.3 Deepening Analysis in Evaluations:

The question of which measures to use to evaluate segmentation models is an ongoing topic of debate [88, 89, 109]. However, a more fundamental aspect of evaluation has been missing from

[29] MIREX is a campaign for evaluating music information retrieval algorithms where the evaluation tasks are defined by the research community. As part of this campaign the first evaluation of music segmentation models was conducted in 2009 (`http://www.music-ir.org/mirex/2009/index.php/Structural_Segmentation`), and has been carried out every year since.

comparative studies, which is the systematic analysis of causes of failure in the models. That is, the examination of which specific pieces or melodies have proven difficult for segmentation models to parse 'correctly'. This type of analysis could provide insights into which specific musical traits might be challenging the models.

Also, systematic analysis of annotated corpora is needed to better understand the nature of the annotated segments therein, i.e. understand why the human annotators decided to segment the music in a particular way. This type of analysis can be used to enrich the original boundary annotations, e.g. provide estimates of what types of cues might be driving the perception of specific boundaries, or estimates of which types of musical attributes might be more determinant for the perception of a particular boundary or the global segmentation of a given piece. Some work along these lines has been conducted on annotated databases of music recordings (e.g. [108]), and to a lesser extent on databases of symbolically encoded melodies (e.g. [100]).

APPENDIX

A About Musical Structures

Here we briefly mention cognitive structuring processes that work parallel or in combination to segmentation. In Figure 9, we illustrate an analysis of a musical piece using the cognitive structuring processes proposed in the Generative Theory of Tonal Music (GTTM) [69]. 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*.

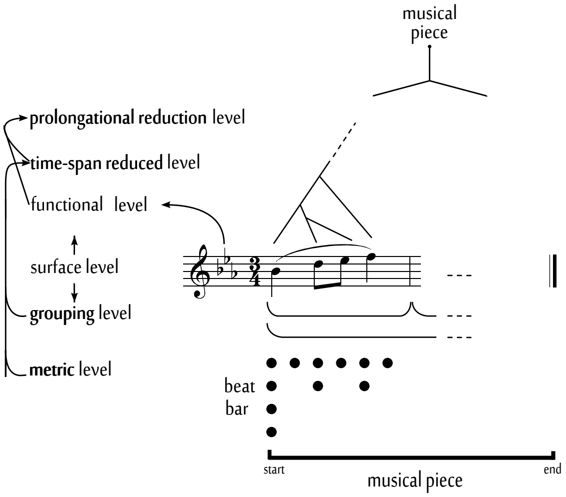

Figure 9: Music analysis depicting GTTM structuring processes (in bold). The analysis illustrates the levels of representation resulting from each structural analysis and considers dependencies among processes.

Short Description of Musical Structuring Processes

According to GTTM the analysis of a piece can be seen as the construction of a tree, where the the root of the tree (uppermost level) represents the entire piece, the intermediate branches the result of analyses of hierarchy between nodes, and the terminal nodes, the 'leaves', represent notes as notated on a score. On the left we mention the levels of structural description resulting with each analysis. The arrows indicate the interrelationships between the depicted levels of structural description. The **grouping analysis** results in a nested set of segments (represented by horizontal curly brackets), ordered hierarchically 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).

The Musical Surface

In Figure 9 we assume that the structuring process occurs over a 'surface level', yet we have not properly introduced the term. The *musical surface* can be thought of as the lowest (representational) level of detail that holds 'musical significance'[69]. 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 [85]. In Figure 10 we present an example of common musical attribute descriptors and the surface level commonly used for analysis by computational segmentation models.

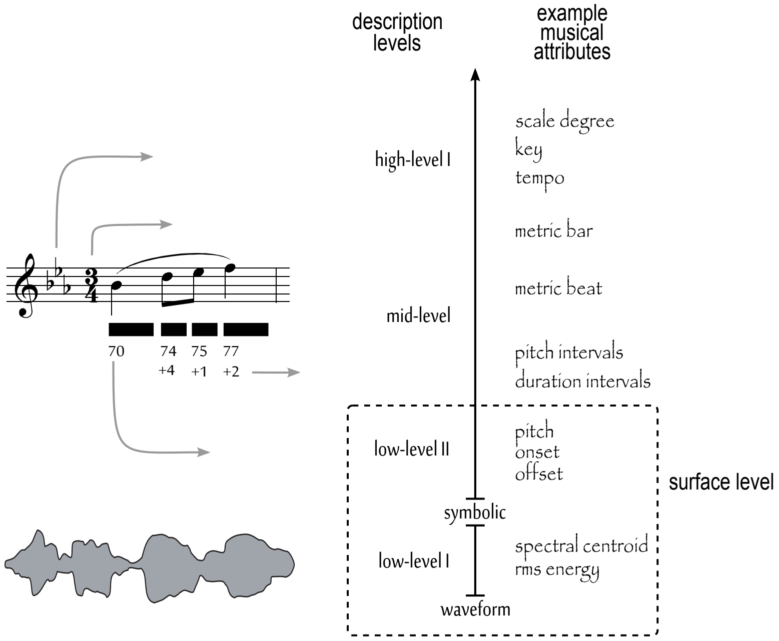

Figure 10: Diagram illustrating levels of description and types of music structures. The description levels and example musical attributes have been organized following the taxonomization of musical attributes proposed in [70].

B Theories of Segment Perception

In this section we comment on two theories of music perception used to construct our taxonomy: Leardahl & Jackendoff's *Generative Theory of Tonal Music* (GTTM) [69], and Narmour's *Implication-Realisation* (IR) theory [81, 82].[30] We have chosen to use these 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.

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 [25, 26], 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" [69, pp. 3-4] 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 11. The segmentation principles, called 'Grouping Preference Rules' (GPRs) in GTTM, are generally

[30] 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 [89] for a succinct, yet well-balanced summary covering the two theories.

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.

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 11: GTTM grouping rules [69] as summarized by Frankland & Cohen [46]. GPR stands for 'Grouping Preference Rule'.

IR: short description

In the IR theory music listening is approached 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 [77], 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 [91], i.e. points in a melody that 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 12. 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 12: Melodic closure rules of the I-R theory [81, 82] as summarized by Pearce [87].

C Peak Picking Heuristics

Below we list common heuristics used by symbolic segmentation models to decide on the location of a boundary from a segment boundary profile sbp.

- *Simple thresholding* [16, 14, 18, 46], where $T \in \mathbb{Z}$ is a threshold selected by the user.

$$\mathtt{s}(i) = \begin{cases} 1 & \text{if} \quad \mathtt{sbp}(i) > T \\ 0 & \quad\;\;\text{everywhere else} \end{cases} \tag{1}$$

- *Highest peak within a symmetric window* [23, 127], where $a \in \mathbb{Z}$ is window length selected by the user.

$$\mathtt{s}(i) = \begin{cases} 1 & \text{if} \quad \mathtt{sbp}(i-a) < \mathtt{sbp}(i) > \mathtt{sbp}(i+a) \\ 0 & \quad\;\;\text{everywhere else} \end{cases} \tag{2}$$

- *Highest peak within an asymmetric window* [127], where $a, b \in \mathbb{Z}$ are parameters selected by the user.

$$\mathtt{s}(i) = \begin{cases} 1 & \text{if} \quad \mathtt{sbp}(i-a) < \mathtt{sbp}(i) > \mathtt{sbp}(i+b) \\ 0 & \quad\;\;\text{everywhere else} \end{cases} \tag{3}$$

- *Highest peak in respect to a weighted mean context* [87, 88, 89].

$$\mathtt{s}(i) = \begin{cases} 1 & \text{if} \quad \mathtt{sbp}(i) > \mathtt{sbp}(i-1) \\ & \quad\;\;\mathtt{sbp}(i) > \mathtt{sbp}(i+1) \\ & \quad\;\;\mathtt{sbp}(i) > W(i, \mathtt{sbp}) \\ 0 & \quad\;\;\text{everywhere else} \end{cases} \tag{4}$$

Where,

$$W(i, \mathtt{sbp}) = k\sqrt{\frac{\sum_{j=1}^{i-1}\left(w_j \mathtt{sbp}_j - \overline{\mathtt{sbp}}_{1,\ldots,i-1}\right)^2}{\sum_{j=1}^{i-1} w_j}} + \frac{\sum_{j=1}^{i-1} w_j \mathtt{sbp}_j}{\sum_{j=1}^{i-1} w_j}$$

D Computing a Ground Truth From Multiple Annotations

Below we list criteria used to determine a benchmark boundary set or "ground truth", from human boundary-annotated melodies. The idea is to have, for each melody in the annotated corpus, a single reference segmentation rather than the multiple annotations obtained from humans.

Notation: In this section a melody is taken to be a sequence of note events $e = e_1, \ldots, e_i, \ldots, e_N$, where each note event e_i consists of a finite set of attributes, and each attribute may take a numerical value drawn from an alphabet ξ. The methods assume that a set of binary vectors $\mathcal{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_j, \ldots \mathbf{a}_M\}$ is available as the result of having M human listeners annotate segment boundaries for each melody in the corpus. In $\mathcal{A}$ an annotated melody is represented in binary vector form as $\mathbf{a} = (a_1, \ldots a_i, \ldots, a_N) \in \{0,1\}^N$, so boundary marking is encoded using a 1, and boundary absence is encoded as a 0. The task is to merge annotations of a $\mathcal{A}$ into single binary ground truth vector $\mathbf{g} = (g_1, \ldots g_i, \ldots, g_N)$.

1. Aggregating boundary annotations and establishing a threshold [12, 43, 76, 127]. In this case, the resulting histogram H of human boundary judgments per note in e is used to create binary boundary decisions by thresholding the number of annotators that agree a certain position i is a boundary, that is:

$$\mathsf{H}(i) > \mathsf{THR} \tag{5}$$

Where,

- $\mathsf{H}(i) = \sum_{j=1}^{M} \mathbf{a}_j(i)$
- THR is a numerical threshold.

2. Stochastic boundary selection criterion proposed in [76]. In this criterion boundary selection is made by requiring that the probability that a boundary is perceived, given that annotators have selected boundaries on the immediate neighborhood of position i (one note event to the left/right of i) , is greater than 0.5:

$$Pr(B_i = 1|X_i) > \frac{1}{2} \tag{6}$$

Where,

- $Pr(B_i = 1|X_i) = \sum_{x \in A} Pr\left(X_i = x_i\right) = \prod_{j=1}^{M} \prod_{k=0}^{2} p_{ik}^{x_{ijk}} \left(1 - p_{ik}\right)^{x_{ijk}}$

  $B_i = 1$ indicates boundary presence at position i

  $X_i = \{\{\mathbf{a}_1(k)\}, \ldots, \{\mathbf{a}_j(k)\}, \ldots, \{\mathbf{a}_M(k)\}\}$, with $k = i + \delta - 1$ and $\delta \in \{0,1,2\}$

  $A = \{(0,0,1), (0,1,0), (1,0,0), (0,1,1), (1,0,1), (1,1,0), (1,1,1)\}$

  $p_{ik} = \frac{\sum_{j=1} M x_{ijk}}{M}$

3. Evaluate models by comparing to an 'optimal' cluster of annotations [89]. To do so the following steps are needed: (a) cluster annotations to obtain $\mathcal{C}$, (b) create a binary ground truth $\mathbf{g'}$ for each cluster $\mathcal{A}'$ using the k-means algorithm, (c) for each melody, select the highest F1 performance of step (b) to compare to the other models.

Step (b) is computed as follows:

$$\arg\min_{S} \sum_{k=1}^{K} \sum_{j \in S_k} \|\mathcal{A}'_j - \mu_k\|^2 \tag{7}$$

Where,

- $\mathcal{C} = \{\mathcal{A}'_1, \ldots, \mathcal{A}'_h, \ldots, \mathcal{A}'_H\}$ is a set of clusters of $\mathcal{A}$ constructed using hierarchical agglomerative clustering, so that each annotated ground truth $\mathbf{a}_j$ included in $\mathcal{A}'_h$ is taken from $\mathcal{A}$.
- The k-means algorithm takes $K = 2$, basically classifying each note as {boundary, no-boundary}.

E Measures Used for Binary Evaluation

Below we list measures used to evaluate computational segmentation models of melodies. The measures assume that both model predictions and annotated ground truth are encoded as equal-size binary vectors **p** and **g**, respectively.

		g_i	
	i	1	0
p_i	1	tp	fp
	0	fn	tn

Table 13: Outcomes of comparing prediction and ground truth binary data. Table classifies co-occurrences of binary values of p_i and g_i as a *true positive* tp, *true negative* tn, *false positive* fn, or a *false negative* fn.

Table 13 shows how co-occurrences between elements in the binary vectors **p** and **g** are commonly classified. In the table co-occurrences of binary values of p_i and g_i are classified as a *tp*, if both ground truth and prediction indicate a value of 1, and *tn*, if both values are 0. Conversely, fn and fn indicate instances where ground truth and prediction differ. In the first case, the prediction contains a 1, while the ground truth contains a 0. The later case indicates the reverse situation. Using this classification of co-occurrences we can then determine:

- $\text{TP} = \mathbf{p} \cdot \mathbf{g}$
- $\text{FP} = \mathbf{p} \cdot \overline{\mathbf{g}}$
- $\text{FN} = \overline{\mathbf{p}} \cdot \mathbf{g}$
- $\text{TN} = \overline{\mathbf{p}} \cdot \overline{\mathbf{g}}$

Where $\cdot$ is the dot product operator of two binary vectors a, b, and $\overline{a}$ denotes the complement operation of a binary vector a.

In the list the *true positives* (TP) value is the number of *tp* occurrences, the *true negatives* (TN) value the number of *tn* occurrences. Likewise, *false positives* (FP) and *false negatives* (FN) correspond to the number of fp and fn occurrences.

We moreover use the notation G^+ and G^- to denote, respectively, the *number of occurrences* of 1s and 0s in **g**. Likewise, P^+ and P^- are used to denote the *number of occurrences* of 1s and 0s in **p**.

E.1 Measures Used When Ground Truth is a Single Binary Vector

The measures presented below assume that, for each melody in the test database, a single binary vector encodes a prediction and a single binary vector encodes the ground truth.

1. The *F1* measure as computed in [109, 127, 88, 89, 87]

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} = \frac{1}{1 + \frac{\text{FN}+\text{FP}}{2\text{TP}}} \in [0, 1] \tag{8}$$

Where,

- $precision = \frac{\text{TP}}{\text{TP}+\text{FP}}$
- $recall = \frac{\text{TP}}{\text{TP}+\text{FN}}$

2. The *accuracy* measure as computed in [88]

$$accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{G}^{+} + \mathrm{G}^{-}} \in [0, 1] \tag{9}$$

3. The *kappa statistic* κ as computed in [88]

$$\kappa = \frac{Pr - Pr_e}{1 - Pr_e} \in [0, 1] \tag{10}$$

Where,

- $Pr = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{G}^{+}+\mathrm{G}^{-}}$
- $Pr_e = Pr_1^2 + Pr_0^2$
- $Pr_1 = \frac{\mathrm{G}^{+}+\mathrm{P}^{+}}{2\left(\mathrm{G}^{+}+\mathrm{G}^{-}\right)}$
- $Pr_0 = \frac{\mathrm{G}^{-}+\mathrm{P}^{-}}{2\left(\mathrm{G}^{+}+\mathrm{G}^{-}\right)}$

4. The *sensitivity index* d' measure as computed in [88]

$$d' = z(\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}) - z(\frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}) \tag{11}$$

Where,

- $z(p), p \in [0, 1]$ is the inverse of the cumulative Gaussian distribution.
- Effective limit values for d' are $[-3.61, 4.65]$ for $z(0.01) - z(0.99)$ and $z(0.99) - z(0.01)$, respectively.

5. The *co-occurrence agreement probability* (CoAP) measure as computed in [76]

$$P(agreement) = \sum_{1 \leq i \leq j \leq N} D(i, j)\delta_G(i, j)\delta_P(i, j) \in [0, 1] \tag{12}$$

Where,

- i, j are two notes from the melody
- G, P stand for ground truth segments and predicted segments, respectively
- $\delta_{G,P}(i, j) = 1$, if i, j belong to the same segment, and 0 otherwise
- $D(i, j) = 1$, if i, j are k notes apart and 0 otherwise
- k is taken to be half the average length of a segment in the annotated corpus

The *co-occurrence agreement error rate probability* (CoAER), can be measured simply as $1 - P(agreement)$. However, to obtain a more detailed view of the error, in [76] the CoAER measure is decomposed in the aggregation of two conditional probabilities:

$$P(error\ rate) = P(miss)P(segment) + P(false\ alarm)(1 - P(segment)) \in [0, 1] \tag{13}$$

Where,

- $P(miss)$ is probability of missing a boundary in P
- $P(false\ alarm)$ is probability of wrongly estimating a boundary P
- $P(segment)$ is the prior probability of having different segments according to G

This allows an interpretation close to that of precision and recall (where $TP = hit$, $FP = false\ alarm$, $FN = miss$, and $TN = correct\ negative$). In the publication the average length of a segment was measured to be 15.08 notes, k was then set to 8 notes, and $P(segment) = \frac{8}{15.08} = 0.532$.

E.2 Measures Used When Ground Truth is a set $\mathcal{G}$

Below we list measures used evaluate computational segmentation models of melodies. The listed measures assume that the prediction is a binary vector, and that a set of binary vectors $\mathcal{G} = \{\mathbf{g}_1, \ldots, \mathbf{g}_M\}$ is available as ground truth. The two first measures correspond to probability distributions and the last measure is normalized, so they all output values in the range $[0, 1]$.

All measures below where proposed in [109].

1. The Position Model (PM) measure

$$PM = Pr(\mathbf{p}|H_s)^{\frac{1}{k}} \propto (\prod_{i=2}^{k} H_s(p'_i))^{\frac{1}{k}} \tag{14}$$

Where,

- H_s histogram of boundary indication (yes/no) per position i for each $\mathbf{g} \in \mathcal{G}$
- k is the total number of segments in prediction $\mathbf{p}$
- $\mathbf{p}' = (p'_1, \ldots, p'_j, \ldots, p'_k)$ is a *boundary index* vector, where $p'_j = i$ iff $p_i = 1$

2. The Position Length Model (PLM) measure

$$PLM = Pr(\mathbf{p}|H_s, H_l, H_n)^{\frac{1}{k}} \propto H_n(k)(H_l(l_k) \prod_{j=2}^{k} H_s(p'_j)H_l(l_{j-1}))^{\frac{1}{k}} \tag{15}$$

Where,

- k is the total number of segments in prediction $\mathbf{p}$
- H_s histogram of boundary indication (yes/no) per position i for each $\mathbf{p} \in \mathcal{G}$
- H_l histogram of the number of notes per segment k for each $\mathbf{p} \in \mathcal{G}$
- H_n histogram of the number of segments per melody in the corpus for evaluation
- $\mathbf{p}' = (p'_1, \ldots, p'_j, \ldots, p'_k)$ is a *boundary index* vector, where $p'_j = i$ iff $p_i = 1$
- $\mathbf{l} = (l_1, \ldots, l_j, \ldots, l_k)$ is a *segment lengths* vector, where $l_j = p'_{j+1} - p'_j$

3. The Average F Score Model (AFM) measure

$$AFM = \frac{1}{\mathcal{G}} \sum_{\mathbf{g} \in \mathcal{G}} F1(\mathbf{g}, \mathbf{p}) \tag{16}$$

For more details on the computation of histograms used in the PM and PLM measures, please refer to [109].

F Description Summary Tables

In this appendix we provide more detailed descriptions of the segmentation models surveyed in this paper. We have separated the approaches in machine learning driven, and non-machine learning driven ('knowledge' based) using Tables 14 & 15, respectively. The summaries provide information of the main goals of the models (purpose), a brief account of their inner-workings (description), and critical commentary (discussion). For convenience the models names have been abbreviated, the meaning of the abbreviations is listed in Table 5, page 18.

Table 14: Description of knowledge-based segmentation models

Acronym	**Year**		**Description**
TPG	1980	[120]	`Purpose`: TPG aims to identify segment boundaries at two time-scales, 'clangs' (∼subphrases) and 'sequences' (∼phrases). `Description`: The *TemPoral Gestalt grouping* model was the first to propose a formalization of visual Gestalt principles to predict musical segment boundaries. The model provides a quantification of the Gestalt principles of proximity and similarity. The quantification focuses on the segregative rather than the unifying aspects of the principles, i.e. they focus on measuring 'change' in a local context along a given parametric representation of a melody. Distance metrics are proposed to formally assess change in two melody parametrisations: chromatic pitch intervals and inter-onset-intervals. To identify clang boundaries, the models searches for local maxima, i.e. an interval being larger than the intervals immediately preceding and following it. If a local maxima is detected it correponds to the starting point of a clang. The computation of clang boundaries uses the $L1$ norm: $d[i] = ioi[i] + \lvert pitch[i+1] - pitch[i] \rvert$, for $i = 1 \ldots N-1$, with i the time index of the melodic sequence, and N the total length of the sequence. To identify sequence boundaries, again the L1-norm is used, this time to measure the distance between the mean pitch and duration content of contiguous clangs. If a local maxima is detected it corresponds to the starting point of a sequence. `Discussion`: In [16], Cambouropoulos argues that, by considering intervals as symmetrical and non-directional (use of the L1 norm), the TPG model can frequently miss grouping boundaries due to the natural asymmetry of musical entities (examples are given in the publication). Also, in [68] Lefkowitz critiques the inflexibility and (apparent) arbitrariness of the weighting system used in TPG to combine information measured in the two (pitch, duration) melodic parametrisations considered.

Continued on next page...

Table 14 continued from previous page

Acronym	Year		Description
AGA	1989	[5]	`Purpose`: The *Automated Grouping Analysis* system was created to identify melodic phrases. The main aim was to combine aspects of harmony, time-span reductions, and local heuristics to segment melodies. `Description`: The system has a serial modular architecture consisting of: a lexical analyser, a chord function parser, a reductions analyser, and an evaluation module. The final output is a ranked list of possible phrase parses for an input melody. AGA proceeds by first computing an estimate of a chord for each metric unit in the melody. Then it generates sets of possible chord parse trees of the melody. The parser employs a context-free grammar based on music theoretic rules of tonal harmony. Chord parses are taken to identify phrase boundaries. The combinatorial explosion on the number of possible chord progression parses is constrained by only allowing parses producing phrase boundaries that agree both with local heuristics (GPRs $2_{a,b}$, $3_{a,d}$) and genre specific constraints (a phrase-level structure template referred as 'normal form') . Subsequently the melody is time-span reduced (non important pitch values are eliminated). Then the melodic contour of the time-span reduced melody is computed. The resulting list of phrase parse hypotheses is finally ranked giving a high score to solutions where phrases show high similarity. Similarity is measured as the maximal match of melodic contour between estimated (time-span reduced) phrases. `Discussion`: The author acknowledges two main limitations: (1) poor generalization capacity, and (2) psychological implausibility. In respect to (1) the system assumes that the 'normal form' (phrase-level structure template) of the melody, which drives the grammar, is known. Also the ranking and evaluation mechanism seems highly sensitive to the properties of the music genre/style. In respect to (2) the system assumes complete and exhaustive perception of possible harmonizations, which violates limits of psychological processing and known human memory limitations.

Continued on next page...

Table 14 continued from previous page

Acronym	Year		Description
GRAF	1989	[4]	`Purpose`: The *GRouping Analysis with Frames* system was created to identify melodic phrases. `Description`: GRAF was developed as an extension to the AGA system [5]. In AGA 'schematic' knowledge is coded in the form of processing constraints (GTTM local grouping rules and music theoretic rules). In GRAF this is extended by using a system of "frames" [79]. GRAF's frames represent possible phrase-level structure templates found in a given genre/style of music (called "normal forms" in the publication). Frames are stored in a knowledge-base of styles/genres referred to as a 'context space'. As in AGA, the output of GRAF is a list of hypothesised grouping structures for an input melody. The system in this case processes the input sequentially (from beginning to end, a note at a time). It evaluates at each step if a boundary is likely to happen according to GPRs $2_{a,b}$, $3_{a,d}$. If the GPRs indicate the current note is a boundary, GRAF attempts to match the hypothesized segment's tonal progression (using AGAs parser) to the normal forms stored in frames. If there is a match the frame is retrieved from long-term memory (context-space) and stored in working memory. Frames are scored according to the degree of match between melody and frame. The system takes into consideration working memory restrictions, maintaining only a limited set of frames in memory. If working memory is exceeded the frames with the lowest scores are eliminated. The final output is ranked according to the phrase structures that most resemble a 'normal form' structure. `Discussion`: GRAF seems to make for a more cognitively plausible system than AGA, taking careful consideration of human memory restrictions, yet it does so at the expense of increasing the amount of knowledge that is assumed available (key, meter, and genre information is required). Moreover, as in the case of AGA, no systematic evaluation of the model is presented.

Continued on next page...

Table 14 continued from previous page

Acronym	Year		Description
ESMS	1990	[19]	`Purpose`: The *Expert System for Musical Segmentation* was created as a preprocessing step to tonal analysis of melodies, and had the purpose of testing possible interaction among grouping mechanisms. `Description`: The system quantifies GTTM grouping principles (with an special focus on the metric/rhythmic aspects of the theory), and adds heuristic rules based on experimental psychological studies of segment perception. ESMS weights the role of GTTM rules across several attribute parametrizations of the melody. ESMS uses forward/backward inference strategies (common in expert systems) for the combination/selection of rules. The strategies are dynamically activated during processing, by recalling control statements. Visual display is provided to the user to provide control statements at any point of processing. The output of the system is a list of grouping hypothesis which have been activated by different cues. `Discussion`: Only an abstract and short descriptions of this model were found at the time this survey was conducted. Thus the model can not be discussed in detail.
Cypher	1992	[102]	`Purpose`: The *Cypher* system was created for real-time machine improvisation. Cypher is composed of two macro modules: one for listening (i.e. analysing musical input) and another for reacting (i.e. generating a musical response). Cypher accepts monophonic and polyphonic MIDI input, and, as part of the listening sub-module, contains a phrase segmentation module. `Description`: The segmentation sub-module assigns phrase boundaries in real-time (analysing MIDI stream chunks of roughly 10 seconds). Cypher takes pitch, velocity, duration, and onset times directly from the MIDI data, and computes a (absolute) classification of these events into registers, note speed (horizontal density), number of chord notes (vertical density), and loudness classes. Also, Cypher automatically derives information of key, beat, and functional harmony degree. To detect phrases, Cypher checks these attribute dimensions for the presence of local discontinuities. Since Cypher uses a metrical representation of time, it checks for discontinuity presence on *each beat*. If a discontinuity is detected for a given attribute, the attributed is flagged with a 1, otherwise the flag stays as 0 (the default). If an attribute has a flag of 1, a (hard-coded) weight value is assigned, supposedly reflecting the perceived strength of the discontinuity. A beginning-of-new-phrase signal is fired if the sum of weights across all attributes for a given beat is greater than a threshold. The threshold adapts to the input in real-time by checking if the generated phrases are either too short or too long, trying to maintain a balance. For high-level (harmonic, key and metric) attributes, strength weights are treated with special care, increasing the weight value according to heuristics rules derived from music theory (e.g. tonic, dominant events as well as events on strong beats are given more weight). For lower level attributes the manner in which the discontinuities are detected, and also what the weights reflect, is not explicitly described in the publication. `Discussion`: The author acknowledges than the analysis agents dealing with key, chord, and beat processing are limited, and perform optimally only for “well-behaved” tonal input, hence making the approach genre/style specific.

Table 14 continued from previous page

Acronym	Year		Description
MPM	1998	[47]	`Purpose`: The aim of the *Music Punctuation Model* is to identify the boundaries of 'small structural units' (∼subphrases). It must be noted that in MPM boundary identification is seen from the perspective of music performance, i.e. the task is to provide advice for a performer (by adding 'comma' symbols to a score) rather than directly model the grouping mechanisms of human listeners. `Description`: MPM was created as part of the KTH rule-based system of automatic music generation [48]. MPM takes as input an encoded music score, from which it uses only the pitch and duration information of note events as specified in the encoding. It also requires harmonic information, which for testing was manually derived and provided to the model. Music punctuation was modelled using 13 rules, coded by a single expert performer. Seven rules are used to identify possible comma positions (PCPs), using a short context of maximum five contiguous notes (three notes before and two notes after the potential coma location). The seven rules provide each PCP with a score that reflects their plausibility as comma locations. Initial score values were coded by the expert performer. The remaining six rules are used to revise PCPs, by either altering their score or completely eliminating them. The criteria used to revise the rules depends on very diverse factors, which range from music theory rules, the scores awarded to preceding PCPs, to statistics of note attributes over the whole input piece of music. MPM was tested on 52 melodic excerpts (extracted from western 'art', folk, and popular music pieces). The excerpts were annotated with comma locations by an expert performer. From these, 26 excepts were used to optimise the scoring values of the rule system of MPM (the optimisation strategy was to minimise the number of insertions made by MPM, which during initial examination was found to produce an overly large amount of commas). The results over the 26 melody test set where found to be 66% hits and 33% insertions. No explicit information about misses is given and so the performance cannot be evaluated in terms of the common precision and recall measures. `Discussion`: The authors acknowledge that, while the system of rules is relatively compact, the weights assigned to each rule are hard to define and also that estimating the instances in which the revision rules need to be applied is not trivial. In respect to the testing, the rules were derived by a single expert performer (one of the authors), from the same database used for testing. Also, the comma annotations used as ground truth where marked by the same performer. This introduces a considerable bias in the evaluation. In respect to applicability of the model in information retrieval scenarios, MPM requires detailed information of the the underlying harmonisation of the melody, such as chord label (name, major/minor), and position (root or inversion), which needs to be provided manually, and hence limits the number of databases in which in can be applied. On a final note, MPM was the last model, among those reviewed in this survey, that attempted to segment melodies using only expert derived rules.

Continued on next page...

Table 14 continued from previous page

Acronym	Year		Description
SPIA	1998	[13]	`Purpose`: The *String Pattern Induction Algorithm* was developed as part of the GCTMS system. The purpose of the algorithm is to identify pattern boundaries. `Description`: The algorithm takes as input a sequence of scale-step pitch intervals (although can admit other surface representations). SPIA starts by matching patterns of two elements, and finishes once it has found the longest possible matching patterns. `Discussion`: This was the first attempt by the author to combine LBDM with a pattern analysis segmentation algorithm, later refined into the PAT algorithm. SPIA is a more simple, brute-force pattern extraction algorithm.
RPF	1999	[117]	`Purpose`: The *Representative Phrase Finder* system extracts "representative" melodic phrases to index music files for MIR. RPF consists of a phrase identification module, and a phrase "representativeness" classification module. `Description`: The idea driving the phrase boundary identification module is that phrase beginnings are often similar. The task is then to locate short melodic fragments roughly corresponding to subphrases, and then check (a) which of those melodic fragments are significantly similar, i.e. can be considered repetitions, and (b) estimate whether they correspond to the first subphrase, i.e. the aforementioned 'beginning', of a phrase. This task is implemented in three stages. First RPF estimates the locations of temporal gaps in the melody (long note durations or rests), which are used as '*sub-phrases*' boundaries; second, it uses approximate match and a automatic thresholding method to identify which of the estimated subphrases can be considered repetitions; third, it estimates which subphrases are the first subphrase in a phrase using two preference rules: (a) the subphrase starts after a long temporal gap, (b) the subphrase starts the beginning of the melody. The starting points of repetitions in which one repeated instance complies with either (a) or (b) are taken as phrase boundaries. `Discussion`: The authors of RPF acknowledge that the first step finds too many sub-phrase boundaries (over-segmentation), which affects all subsequent computation. Also, they report that, in many cases, the merging method tends to wrongly consider adjacent phrases that are too similar as a single phrase (under-segmentation).

Continued on next page...

Table 14 continued from previous page

Acronym	Year		Description
PSS	2000	[68]	`Purpose`: The *Piece-Sensitive Segmentation* Model aims to infer segment boundaries of relatively short duration ($\sim$ 3-5 notes), for monophonic music (though it claims to be generalizable to polyphonic music). The main contributions of PSS are to be able to handle a rich set of musical attributes, and moreover be able to combine the contribution of each attribute to discontinuity detection in a cognitively-meaningful way. `Description`: PSS assumes segment boundaries are cued by discontinuities in the flow of music. PSS models discontinuity by detecting "change-in-the-rate-of-change" of a given musical parameter. Musical parameters considered for analysis are grouped into 4 categories: pitch-related, rhythm-related, timbre-related, and articulation-related, with each category containing up to 4 different attributes. PSS operates by first locating discontinuities along every dimension. Subsequently, PSS computes weights for each of the attribute dimensions by examining the length of segments. Attribute dimensions where the identified boundaries produce segments of length closer to an 'ideal' segment length (4-5 notes) are given more weight. `Discussion`: In PSS all monophonic information (pitch, duration, timbre, and articulation) is assumed known a priori, which greatly restricts its applicability to real life scenarios. Moreover, PSS has not undergone systematic evaluation in large music corpora, thus its generalization capability is unknown. Also, the computed weights are given for the whole melody, rather than on a boundary-by-boundary basis. Lastly, the algorithmic description of the core process of PSS (the discontinuity detection module) is incomplete, making the implementation of the model impossible.

Continued on next page...

Table 14 continued from previous page

Acronym	Year		Description
LBDM	1996-01	[16, 14]	`Purpose`: The *Local Boundary Detection Model* was initially proposed as the segmentation module of the GCTMS music analysis system [13]. The aim of LBDM is to identify local (∼phrase-level) boundaries on melodies. `Description`: LBDM assumes segment boundaries are cued by discontinuities in the flow of a melody. LBDM computes discontinuity strength values for pitch/duration interval representations {`cp-iv, ioi, ooi`} of a melodic sequence. Peaks in the resulting sequence of strength values are taken as potential phrase boundaries. In LBDM a discontinuity strength value is assumed to represent a *local* measure of the *degree of change* between intervals. The local context taken to analyse intervallic change is three intervals, or equivalently four note events. The model computes strength values by first applying a *change rule* (CR), and subsequently a *proximity rule* (PR). The CR computes the degree of change of each interval in respect to the preceding and succeeding intervals, using a L1-based distance measure. The PR simply multiplies each interval by the computed degree of change associated to that interval. `Discussion`: The model uses a simple weighted average to combine the discontinuity strength profiles of the {cp-iv, ioi, ooi} representations, however there is evidence that the relationship among these dimensions is non-linear [124]. Also, even if a linear model is adequate, the weights are static (set at initialization), and more recent studies suggest that weights might need to change over the course of music [128].

Continued on next page...

Table 14 continued from previous page

Acronym	Year		Description
eLBDM$_1$	1999	[75, 76]	Propose minor extensions to the LBDM model, in the form of normalization factors.
Grouper	2001	[118]	`Purpose`: *Grouper* is the segmentation module of the Melisma music analysis system. Grouper was created to identify melodic phrases. `Description`: Grouper accepts as input a melody as a note list (onset/offset time, chromatic pitch), and metric information at the bar and beat level. The output is a list of binary boundary (yes/no) judgements for each note in the melody, i.e. an exhaustive partitioning of the melody into non-overlapping groups. The model assigns boundaries through the application of three preference rules: a *temporal gap rule*, placing boundaries at large IOIs and OOIs, a *phrase-length rule*, giving preference to phrase segments of about 10 notes, and a *metrical parallelism rule*, preferring that successive segments begin at parallel points in the metric grid (first beat of a bar or the first beat of a hyper-metric structure). The optimal phrase-level segmentation of a melody is computed using dynamic programming, evaluating candidate phrases according to a linear combination of the three rules. `Discussion`: Rules two and three of the model require a priori knowledge of phrase length and metric structure, respectively. Estimation of metric structure, especially at the bar level, can be problematic, since it has been shown that phrase structure influences the conception of metrical structure at that level [2]. Also, priors for phrase length can only be determined experimentally, requiring annotated corpora. Lastly, Grouper might fail to find phrases in melodies of musical styles that are characterized by low rhythmic contrast (where rule one fails to give meaningful boundary candidates).

Continued on next page...

Table 14 continued from previous page

Acronym	Year		Description
MDSM	2003	[43]	`Purpose`: The *Melodic Density Segmentation Model* aims infer 'low-level' (∼phrases) segment boundaries in melodies. To locate segment boundaries, MDSM proposes a quantization of the Gestalt rule of proximity applied to pitch. Hence, MDSM defines the task of locating boundaries as a search for large intervals ('discontinuities') in respect to surrounding intervals. However, differently from previous models of pitch discontinuity in melodies (e.g LBDM [16, 14], TPG [120]), MDSM assumes pitch intervals are perceived between all notes occurring over an interval of time (short term memory window) and not just between consecutive notes. Thus it is postulated that pitch discontinuity should be measured in respect to all intervals rather than just in respect to the previous and subsequent intervals as done by [14, 120]. Moreover, the model hypothesizes that the perceptual salience of a pitch interval's *size* is related to the 'familiarity' of that interval within a specific style (the more familiar the interval, the more stable a mental representation might be, and thus the more important as a reference to assess discontinuity). The familiarity of a pitch interval is assumed related to the frequency of appearance of that interval in a style (in the publication the major/minor model of [13] was used, but the author suggests that a better model might be collecting interval frequency counts extracted from a melodic corpus). `Description`: MDSM uses information of pitch, onset, and tempo of a melody. MDSM provides a measure of the "accumulated melodic cohesion (density)" between pitch intervals, and then identifies local boundaries as points of low melodic cohesion. MDSM computes melodic cohesion at each note event point using a sliding window approach, where the window's size is defined in respect to short-term memory restrictions, using the tempo information (4 seconds in the publication's experiments). Within the window, melodic cohesion is computed as a sum of the salience of each pitch interval. In addition, each salience value is weighted by an attenuation function, which is meant to model the effects of 'recency' (i.e. recent note events are given higher weight than distant note events). `Discussion`: In a small evaluation (7 melodies), MDSM proves more selective than LBDM (higher precision). The ground truth however, was given by boundaries that correspond to the starting points of melodic patterns (for which patterns where computed by an algorithm instead of been annotated by a human). In [43] two main problems are discussed: (1) There is a smoothing effect due to the window size and the attenuation function, which affected the 'reaction time' on the location of boundaries, i.e. in some cases MDSM would predict boundaries a couple of note events late, and (2), since MDSM uses a constant length window set at initialization, is only robust to small changes in tempo.

Continued on next page...

Table 14 continued from previous page

Acronym	Year		Description
eRPF	2004	[21]	Purpose: Extension of RPF [117]. Used to find segment boundaries and labels in melodies at the phrase and sentence level. Description: Three types of segments are defined: 'music fragments', 'phrases', 'sentences'. Music phrases and sentences comply with their music theoretic definitions in respect to their approximate length (i.e. figures/cells<phrases<sentences<sections, where < is read 'of shorter duration than'). Conversely, music fragments are more formally defined, i.e. fragments are melodic segments with boundaries indicated by 'terminative notes'. Where terminative notes are notes where the inter-onset-interval between the terminative note and the following note is excessively large. Music fragments might be either larger or shorter than phrases. The approach to segmentation of eRTF is to: (1) segment the melody into music fragments, (2) estimate phrases using the identified music fragments, and (3) estimate segments by estimating which phrases are the starting units of segments. For (1) terminative notes are identified by using LBDM model applied only to duration values. For (2) music fragments are made compliant with respect to two attribute of phrases. To this end two formal constraints are enforced in the definition of phrases: (a) phrases are longer than 6 notes and shorter or equal to 12 notes, (b) the melodic contour of the phrase must correspond to one of the four most typical arch shapes defined by Huron [60]. For (3) the starting phrases of sentences are assumed similar, and are also assumed to comply with the rules outlined in [117]. The similarity between all estimated phrases is estimated. Phrase similarity is computed as a weighted combination of the longest-common-subsequence algorithm and a heuristic algorithm. All phrase pairs sharing one phrase and exceeding a similarity threshold are given are clustered. Each cluster group is given a unique label. Finally, Using the information of possible starling phrases and the sequence of class labels, a repeated pattern finding algorithm is used to identify repeating sentences. Repeating sentences are also given a unique class label. The segmentations produced by eRPF are tested using a corpus of 50 folk melodies from Taiwan. The corpus was annotated (boundary markings and labels) at the phrase and sentence level by music experts. The automatically detected boundaries at the phrase level obtained 0.68 mean precision and 0.78 mean recall. Discussion: No performance values are given for the labelling of phrases, nor for boundary detection at the sentence level. Only mean recall performance (0.63) is provided as the evaluation of labelling at the sentence level. Overall, the analysis of the model provides little discussion in respect to the heuristics employed for determining segment boundaries, the strategies to assess similarity, and the algorithm to find repeated sentences. The eRPF model has not been included in comparative studies, hence in respect to other models and in respect to melodies of difference styles and traditions is unknown.

Continued on next page...

Table 14 continued from previous page

Acronym	Year		Description
Modus	2004	[2, 3]	`Purpose`: The Modus system is the computational counterpart of a theory of melodic structure proposed by Ahlbäck [2]. `Description`: Modus is entirely coded using cognitively inspired rules. Modus performs two types of analysis, one based on discontinuity rules similar to those found in the GTTM, and another based on melodic parallelism computed using preference rules over abstractions of pitch and duration. For the computation of melodic parallelism Modus performs an analysis of pitch (similar but not equal to pitch spelling), and metrical analysis at the tactus level. `Discussion`: The system has been mainly criticized by its inflexibility in terms of decision making (all parameters are hard-coded), and interpretability of the results.
ATTA	2004-06	[53, 54, 55]	`Purpose`: The *AutomaTic span-Tree Analyser* is a full computational implementation of the generative theory of tonal music GTTM [69]. As such, the system proposes quantifications for the segmentation rules established in the theory. The implementation is restricted to the analysis of melodies. `Description`: In ATTA GPRs 1, $2_{a,b}$, $3_{a,b,c,d,}$, 4, 5, and 6, are quantified. In addition a set of 15 weights are assigned to provide control over the different parameters of the rules, and control the way in which rules combine. The segmentation analysis module of ATTA consists of two stages. Stage one estimates the locations 'low-level' segment boundaries. The location estimates are computed using a quantification of GPRs 1, $2_{a,b}$, $3_{a,b,c,d,}$, and 6. Subsequently, relative strengths are assigned by weighting each rule (weights are to be set manually). Finally, a threshold method is applied, and only those boundary locations with values above the threshold are considered for the second stage of analysis. Stage two estimates a hierarchical organization of the boundary locations. Boundaries are re-computed in a top down fashion, using again the quantification of GPRs 1, $2_{a,b}$, $3_{a,b,c,d,}$, and 6, but this time using GPRs 4 and 5 to guide the process. Strength weights are computed automatically for each re-computed boundary. Strength weights from stages one are two are merged into a single weight by multiplying them (that is, only for those boundary locations of stages one and two that match). The hierarchy is computed by selecting the boundary locations with higher weight. `Discussion`: the ATTA system has been criticised because rules are non-adaptive, i.e. are set during system initialization and do not automatically react to changes in the input piece analysed [128]. The approach was tested in 100 classical melody extracts (up to 10 measures in length), achieving a mean-F1 performance of 0.67. It was reported in [55] that the procedure to manually optimize weights by ATTA experts took on average 10 minutes per melody excerpt. The goal of the manual optimization was to top the F1 measure obtained by a random assignation of weights.

Continued on next page...

Table 14 continued from previous page

Acronym	Year		Description
eLBDM$_2$	2004	[20]	`Purpose`: Extend de LBDM segmentation algorithm to be used as a pre-processing step to melody extraction. `Description`: The approach assumes that LBDM can be used to accurately find segment boundaries. By means of a few heuristic rules, the authors attempt to classify boundaries computed by LBDM. The classification separates boundaries in end-of-melody <EOM> and phrase-boundary <PB>. Segments separated by <PB> are merged. Melodies are identified as everything between to consecutive <EOM> boundaries. The performance of the approach is evaluated in a set of 80 pop music MIDI files. `Discussion`: The heuristic rules employed to classify boundaries are very inflexible, and moreover require knowledge of metric structure at the bar-level.
PAT	2004-06	[15, 18]	`Purpose`: The *PATtern boundary strength profile* model attempts to complement LBDM in the detection of melodic phrases by modelling the effects of identifying repetitions in boundary perception. `Description`: The model first identifies repetitions using an exact-match string pattern search algorithm. Second, the model scores the salience of identified repetitions using a heuristic function $h = \frac{\mathtt{L}^l\mathtt{F}^f}{10^{o\mathtt{TO}}}$, based on repetition length $\mathtt{L}$, frequency $\mathtt{F}$, and temporal overlap $\mathtt{TO}$, where l, f, o are user defined weights. Third, the model selects meaningful repetitions using a 'boundary strength profile'. A boundary strength profile is a vector of length equal to the input melody length. In the profile each element value encodes the strength with which a segmentation model 'perceives' a boundary at the temporal location of the element. The profile is computed by (a) assigning the salience score of a given set of repeated fragments to each instance of the set, and (b) summing the salience scores of all instances of different sets that begin and/or finish at the same time point. Peaks in the profile mark the starting and/or ending points of the most salient repetitions. `Discussion`: PAT is not quantitatively evaluated, and combination strategy between PAT and LBDM is proposed. Instead, a range of situations in which the PAT might be detecting boundaries that LBDM misses and vice-versa are presented.

Continued on next page...

Table 14 continued from previous page

Acronym	Year		Description
AMS	2008	[128]	`Purpose`: This publication presents an algorithm for *Adaptive Melodic Segmentation*, to search for phrase-level segments and subsequently motives in melodies. `Description`: AMS models boundaries as discontinuities computed in respect to intervallic representations of pitch, duration, and intensity attributes of a melody. For each note event in a melody, a weighted combination of the sizes of the attribute intervals in a local (4 note) context is used to describe the boundary strength at that point (in a way similar to the LBDM model). The novelty of the algorithm relies in that the threshold to determine the strength above which a boundary is computed adapts to data. The threshold value is updated at each time step using low-order statistics derived from past events. Essentially the threshold is set to match the standard deviation *sd* of the melodic attribute values along a specific dimension (attribute values are normalized so the *sd* is enough as a measure of dispersion). The threshold adapts to data using heuristics, raising or lowering its value by requiring that 15% - 45% of the note event attribute values are below it. The threshold is used to dynamically change the values of the attribute weights, the manner in which the change in weight is computed is not clearly explained in the publication. `Discussion`: The choice percentage range value used to adapt the threshold is not justified. This range is a rough estimate on the expected number of phrases. The influence of percentage range settings are not studied systematically. Moreover, no systematic assessment of the performance of the model has been published.

Continued on next page...

Table 14 concluded from previous page

Acronym	Year		Description
MTSSM	2010	[97]	`Purpose`: Compute the optimal segmentation of instrumental parts within a polyphonic piece by considering the segmentation of each part in respect to the 'global' segmentation of the piece (i.e. the segmentation suggested by all other instrumental parts in the piece). MTSSM aims to identify segments at multiple levels of granularity, ranging from form/section to subphrase level segments (in the publication it is mentioned that a single instrumental part can have up to five segment granularity levels). `Description`: The input to the MTSSM is a MIDI encoding of polyphonic pieces. The encoding is assumed to be composed of different tracks, one for each instrumental part. The MTSSM system is composed of two main modules. The first module computes several candidate segmentations for each track. In this module the segmentation analysis uses only information contained within the track being analysed. This is called the 'local' segmentation. To compute a local segmentation, MTSSM assumes the main cue for segment perception is repetition, and hence uses exact and approximate string matching algorithms to detect repetitions. The repetitions are detected over four different representations of the musical material: chromatic pitch values, pitch intervals, pitch contour, and event durations measured in beats. Detected non-overlapping sequences of musical material deemed 'similar enough' are taken to correspond to segments at one level of granularity (e.g. in a monophonic part, all repeated sequences of pitch intervals with a length of four beats are taken as segments with a granularity of four). The output of module one is a set of possible segmentations at multiple granularity levels. The second module of MTSSM attempts to find the local segmentations that fit best with the segmentation suggested by all other tracks in the piece. In the publication the later segmentation is referred to as the 'global' segmentation. The best fit analysis is carried out by computing a local segmentation score, a global segmentation score, and then selecting the best segmentation for each track as the segmentation that maximises the combination of local and global scores. The local score is calculated by performing a pairwise comparison of all local segmentations of a track. The global score is calculated based on the correlation between the segmentation in one track to the segmentations of other tracks. A heuristic function is used to compute the fitness between global and local segmentations. The resulting optimal segmentation of each track comprises a hierarchy of non-overlapping segments, which encompasses segment granularities ranging from subphrases to sections. `Discussion`: Each part is allowed to be polyphonic, yet the similarity assessment methods seem to be monophonic. It is unclear how the reduction of polyphony to monophony is computed within the system. Also, the heuristic function used to assess the fitness between global and local segmentations is complex (it requires the tuning of 10 parameter weights, in the publication parameter weights are tuned manually by experts). No systematic evaluation of the system is presented in the publication, hence performance and generalisation capability is unknown.

Table 15: Description of data-based segmentation models

Acronym	**Year**		**Description**
RAAM	1995	[65]	`Purpose`: To learn 'reduced representations' of melodies. In doing so RAAM estimates melodic segments at two levels of granularity: (a) longer than a bar (∼phrases), and shorter than a bar (∼subphrases). `Description`: RAAM is inspired by reductionist theories, which postulate that listeners judge the structural importance of musical events *while* forming mental representations. As a consequence, reduced memory representations can be expected retain only the 'gist' of the music listened to. Large then proposed a neural network with a Recursive Auto-Associative Memory (RAAM) architecture to test this hypothesis. The idea is to have RAAM 'learn' a melody, i.e. encode a compressed version of the melody, and then check if it is possible to accurately reconstruct the melody from the encoding. The reconstruction accuracy was tested in respect to pitch structure and segment structure. For this two simple melodies where first annotated with pitch structure (GTTM-like prolongational time spans, see Appendix A), and segment structure (approximately at the level of subphrases and phrases). Segment boundaries at the smaller constituent levels (less than a measure) were annotated to align with the locations of strong metrical beats. GTTM grouping rules (see Appendix B) were used to annotate segments boundaries at levels larger than the single measure. Each melody was first compressed using the annotations, and was then reconstructed by the decoder network. The model is evaluated using the original annotations as ground truth. `Discussion`: RAAM is supervised, i.e. needs melodies with annotated structure to learn from. The performance of RAAM in melodies not used for training has not been systematically tested, thus generalisation capacity is unknown.

Continued on next page...

Table 15 continued from previous page

Acronym	Year		Description
MPM(NN)	1998	[47]	Purpose: The aim of the *Music Punctuation Model* is to identify the boundaries of 'small structural units' (∼subphrases). It must be noted that in MPM boundary identification is seen from the perspective of music performance, i.e. the task is to provide advice for a performer (by adding 'comma' symbols to a score) rather than directly model the grouping mechanisms of human listeners. Description: A rule based version of this model was described in Table 14, here we describe a artificial neural network (NN) model developed for the same task, and so we refer to this version of the model as MPM(NN). The aim of MPM(NN) was to investigate whether the rather complex rule system hand crafted for MPM(rule) could be simplified. The NN architecture consists of three feed-forward layers, two output neurons, and two feedback neurons connecting output and input neurons. The first layer takes as input chromatic pitch intervals, harmonic information, and duration information. The second layer mixes pitch and harmonic information. The third layer mixes the combined pitch/harmonic information with duration information. The two output neurons reflect the punctuation judgement: insert comma, do not insert comma. The input and hidden layers are arranged to represent a five note context, i.e. for each time of information (pitch, harmonic, duration) five neurons are used. The NN was trained in 37 examples, each consisting of a five notes containing an annotated comma after the third note. Discussion: MPM(rule) was found to outperform the MPM(NN).

Continued on next page...

Table 15 continued from previous page

Acronym	Year		Description
DOP	2002	[8]	`Purpose`: Compute boundaries for phrase-level segments using the DOP (*Data Oriented Parsing*) method. DOP is a natural language parser that was extended to work with symbolic musical input. DOP uses a *supervised* probabilistic grammar method to parse melodic input into segments, i.e. it needs a corpus with annotated segment structure to learn from. `Description`: The idea is that human listeners with listening experience in a given musical tradition form 'phrase structure templates', and that these templates influence segment perception. To model the formation of phrase structure templates, first a phrase 'class' for each distinct phrase in a phrase annotated corpus is computed (classes are soft in that a given phrase can be a member of more than one class). The classes are computed based on absolute melodic attributes (i.e. chromatic pitch and onset-to-offset interval for duration). To segment melodic input a probabilistic grammar parsing algorithm (DOP) is used to determine the phrase class sequence with maximal probability. To compute the most probable phrase class sequence the parser uses a Markov model of phrase class sequences, i.e. it uses a corpus to estimate the conditional probability distribution of a phrase sequence given a phrase class sequence. The parser also conditions the phrase sequence on the total number of phrases in the piece. `Discussion`: The publication reports the highest segmentation performance to date on a large number of melodies (81% mean-F1 score on 1000 melodies of the Essen Collection). However, DOP uses as input absolute melodic attributes, which raises the question of how much the learning method is over-fitting to the corpus. If so, the rules learnt by DOP from the Essen collection might be expected to hold little significance for the analysis of other styles. Given that annotated corpora for symbolic segmentation are at present hardly available, the approach can not be expected to be applicable to mixed melodic corpora in the short term. Also, DOP has not been included in comparative studies, and thus its performance relative to that of segmentation models in the same corpus is unknown. (In [127] the implementation of DOP provided within the MIDI toolbox [42] was used for comparison, but we learned from the authors of this implementation that is was made only for demonstration purposes and hence can not be considered reliable.)

Continued on next page...

Table 15 continued from previous page

Acronym	Year		Description
ISSM	2002	[125]	`Purpose`: Compute boundaries for sub-phrase segments, and assess similarity between sub-phrases. `Description`: ISSM operates by generating different segmentations of a melody, and subsequently rating the interpretations looking to find the most 'perceptually preferable' one. ISSM first generates all possible segments of consecutive notes (limited to 10-note melodies due to computational complexity issues). Prior to ranking, segments are filtered if they don't comply with GPRs $2_{a,b}$, $3_{a,b,c,d}$. The non-filtered segments are rated using a neural network augmented with fuzzy logic. The network combines rules defined by the user with rules derived automatically by training the network with segment (boundary, label) annotated data. The rules to rate segments are based both on features of the individual segments (e.g. number of notes, duration of a segment), and features between segments (e.g. similarity of two segments in respect to pitch contour, pitch interval, tempo, and loudness). `Discussion`: The ranking module of ISSM requires training data annotated with both segment boundaries and labels indicating perceptually similar segments, which is expensive and time-consuming to produce. Moreover, the 10-note maximum does not allow the system access to long-term correlations (which is especially desirable for assessing the similarity between segment candidates). Finally, ISSM was trained and tested using only 15 melodies, and the results presented are only qualitative, hence generalisation capacity is unknown.

Continued on next page...

Table 15 continued from previous page

Acronym	Year		Description
E4MS	2003	[44]	`Purpose`: identify boundaries for segments in melodies using a *'entropy' based* model (hence the acronym: *Entropy for Melodic Segmentation*). It is not specified whether the model is meant to identify the boundaries of phrases, subphrases, or larger segments. `Description`: The idea behind E4MS is that the perception of segment boundaries in melodies is related to the unpredictability of melodic continuation (points of 'closure' in terms of the implication-realisation theory of Narmur, see Appendix B). E4MS uses an unsupervised machine learning technique (mixed-order Markov models) to model melodic continuation and information-theoretic entropy to measure unpredictability. The input to the model is a chromatic pitch interval and duration ratio melodic attribute representation. The Markov model parameters are: $order \in \{1, \ldots, 6\}$ and $training\ corpus = input\ melody$. Also, E4MS uses an iterative procedure common in text processing to estimate the mixing coefficients of the model. The model first processes each attribute stream separately, aiming to obtain, for each event position in the melody, a distribution of the possible attribute continuations at that position. Then it computes the entropy value at each melodic event position. This way of measuring entropy results an 'unpredictability profile' which is assumed reflects the uncertainty of melodic continuation of a modelled listener. Abrupt changes in the profile are assumed to correspond to segment boundaries. `Discussion`: *E4MS* was not evaluated quantitatively. Instead, a visual depiction the correspondence between identified boundaries and human annotated boundaries are shown for one 20^{th} Century melodic composition. Moreover, the implementation of the model is rather limited in that only the input melody is used for 'training', hence essentially modelling a listener that has only heard the input melody. Also, the preceding context to compute continuation distributions is of only 4-5 events. All in all, *E4MS* constitutes a first step in using information-theoretic measures for melody segmentation. More refined models (e.g. [87]) followed.

Continued on next page...

Table 15 continued from previous page

Acronym	Year		Description
SONNET-MAP	2003	[57]	`Purpose`: SONNET-MAP is a neural network model developed to learn melodies. The model combines two architectures, namely a *Self-Organizing Neural NETwork* and a *ARTMAP*. The SONNET-MAP combination is used to detect segment boundaries in a collection of 50 melodies composed by The Beatles. The author establishes that the boundaries detected correspond to phrases, but most segments identified are relatively brief (≤6 melodic events), and so are closer to subphrases. `Description`: The input melody representation is pitch intervals and (quantised) IOIs. These attributes are first processed independently using different SONNET modules and then integrated using a associate map fields technique. The model operates by iteratively learning a melody, after a user defined number of iterations the learning process is stopped, the pitch interval and IOI sequences the network has learned are assumed to correspond to phrases. This sequences are 'chunked out' and the learning process is reinitialised. The chucking-out/reinitialise cycle is iterated until there are no more events in the melody (heuristics are used if the network fails to learn any of the melodic material). A final SONNET module aggregates the identified subphrases forming a hierarchical memory structure that encompasses the entire input melody. `Discussion`: The author argues that the segmentations performed by SONNET-MAP are consistent with many aspects of melody perception. To validate his claims, the author analyses the decisions made by SONNET-MAP to segment 50 melodies composed by The Beatles. The decisions are shown to correspond to grouping mechanisms identified in the field of music cognition, such as: the influence of previously chunked-out segments, dimension (pitch, rhythm) which dominated the formation of the segment, temporal gap, position (beginnings found to be more important), pitch proximity, pitch-interval sequence repetition, and IOI sequence repetition. The author does not directly evaluates the segment boundaries detected by SONNET-MAP. Yet, it conducts a retrieval experiment using the automatic segmentation of the melody and which obtains a higher performance than the retrieval system operating over whole melodies, which indirectly shows that the segments learnt by SONNET-MAP might correspond to those perceived by humans.

Continued on next page...

Table 15 continued from previous page

Acronym	Year		Description
EME	2004	[63]	`Purpose`: Compute boundaries for subphrase-level segments in melodies using a *'maximum entropy' based* learning system. `Description`: The idea behind EME is that subphrase segment structure becomes intelligible to human listeners (by a large extent) due to musical enculturation. EME proposes modelling the effects enculturation on subphrase segmentation using an unsupervised machine learning approach. The approach aims to identify the sequence of non-overlapping subphrases in a given melody that best agrees with an optimality criterion. The optimality criterion is based on the notion that predictability of the next event in a sequence of musical events *within a subphrase* is *high*, while predictability of the next event *at a subphrase boundary* is *low*. To quantify this notion information theoretic entropy is used (entropy results in high values when event continuations are unpredictable). Thus, the optimality criterion is that the 'most intelligible' sequence of subphrases is that which results in the highest total entropy over the phrase. To learn which sequences of melodic events are more predictable (within a given cultural tradition), a corpus of 2323 Hungarian melodies is used. Melodies are represented as sequences of chromatic pitch intervals, and subphrases are considered to be sequences of up to 6 consecutive pitch intervals. EME computes the frequency of occurrence of pitch interval sequences containing up to 6 contiguous intervals (i.e. unigrams to 6-grams) in the corpus. It then uses these frequencies to compute the probability distribution of unigrams up to 6-grams. An adaptive gradient search algorithm is used to compute the maximum entropy sequence of subphrases, taking as input the ngrams and their associated probabilities. `Discussion`: EME is tested using only a single representation of the melody (pitch intervals). The model also requires as input boundaries of stanzas and phrases, which in the publication are assumed known a priori. Some experiments are conducted to investigate how the model decides what a clear subphrase structure is. It was found that the model prefers boundaries with intervals following the pentatonic scale (intervals of a 2^{nd}, a 4^{th}, and a 5^{th} are found to be specially important to discern subphrase segment structure). However, due to the fact that subphrase boundary annotated corpora do not exist, no systematic testing of the model's predictions was conducted, and thus it is uncertain as to whether the boundary predictions match those perceived by humans.

Continued on next page...

Table 15 continued from previous page

Acronym	Year		Description
IDyOM	2006-07	[91, 94]	`Purpose`: IDyOM (*Information Dynamics of Music*) is a machine learning model of *melodic expectation*, which has been extended to detect melodic segment boundaries. The segment granularity IDyOM is able to detect has been left open (the authors imply it can detect boundaries at various segment granularities), yet it has only been tested for phrase boundary detection. `Description`: IDyOM's segmenter is based on the idea that predictability of the next event in a sequence of musical events *within a segment* is *high*, while predictability of the next event *at a segment boundary* is *low* (in this context unpredictability can be understood as 'closure' in terms of the implication-realisation theory of Narmur, see Appendix B). To model this idea, IDyOM uses a two step process. First, it computes the probability distribution of melodic continuation at each point in the melody. Second, it processes the obtained distributions using an information-theoretic measure. The result of the later step is a real value for each event position in the melody. The sequence of real values is interpreted as a 'surprise' profile. Segment boundaries are assumed reflected as local maxima in the profile (points of high surprise). To predict the distributions of possible continuations for each melodic event, IDyOM uses a variable-length Markov model (VLMM) framework. The framework has a number of features: (a) is *unsupervised*, i.e. it does not require the training set of melodies to be annotated with segment structure. (b) it estimates continuation distributions by combining information collected both from a melodic corpus and the input melody itself. The former is called the long term model (LTM), which simulates the musical listening experience an artificial listener has prior to processing the input melody. The later is called the short term model (STM), and is meant to simulate ongoing 'real time' music listening. That is, the statistics necessary to estimate continuation distributions are collected incrementally, at run time. (c) IDyOM models the multidimensionality of melody perception by considering 'multiple viewpoints' of the melody during processing. That is, the continuation distributions are estimated by simultaneously processing (and combining the contributions of) a set of melodic attribute streams. Moreover, IDyOM simulates attention by emphasising or diminishing the contribution of the different attribute streams when estimating continuation distributions. This procedure is automatic and dynamic, i.e. contribution relevance is inferred from data and updated for the estimation of each continuation distribution. (d) IDyOM is able to estimate how much preceding context is necessary to obtain the optimal continuation distribution estimation (i.e. it defines what the optimal Markov model order is at each point in the melody at run time). `Discussion`: The model has already participated in various comparative studies. Its performance is competitive, but not state-of-the-art. Moreover, despite is sophistication, the current implementation of IDyOM also has the following limitations:

Continued on next page

Table 15 continued from previous page

Acronym	Year	Description
		(1) it has high time space complexity. This is a commonly acknowledged downside of using Markov models. The high memory requirements can be waived using compression techniques (IDyOM uses suffix trees). The high time complexity can be somehow waved by minimising the number of melodic attribute streams chosen for processing (as well as preferring attribute streams will smaller alphabets). (2) it offers no way of automatically estimating which attribute representations might be more relevant (the set of melodic attribute representations is a present set by the user). (3) despite the sophisticated mechanisms to improve probability estimation, IDyOM's event prediction (and consequently segment prediction) capacity is limited by data sparseness. That is, it was shown in [85] that the average context used for prediction in folk and classical melodic corpora is of 3-4 note events. These contexts fall short if compared to those thought to be used by humans. (4) IDyOM has no explicit model of working memory. That is, when needed, long term statistics are fetched *directly from the LTM* and combined with those of the STM to estimate continuation distributions. Moreover, the LTM is essentially a model of eidetic memory. That is, the LTM stores statistics for melodic patterns consisting of, among others, *absolute* attribute sequences (e.g. inter-onset-interval or chromatic pitch sequences) of virtually any length (IDyOM's LTM stores possess no restriction on the patterns stored). (5) IDyOM does not consider segment formation at runtime. That is, during processing of a melody, the influence of previously determined boundaries in the melody is not used to determine subsequent boundaries. (6) it has been noted in [87] that extending the model to use a metric-based representation of time rather than a event-based representation of time (as it currently does) is not trivial.

Continued on next page...

Table 15 continued from previous page

Acronym	Year		Description
IR4S	2006	[38]	`Purpose`: The focus of the paper is not on music segmentation per se, but rather on proposing extensions of a information-theory measure of statistical complexity: the *information rate* (IR) (originally introduced in [39]). In the [38] different models employing the IR measure are used to provide examples of its dynamic behaviour when processing audio/symbolic polyphonic music. By analysing of these examples it is argued that the IR measure can be used to detect time points of 'structural importance' in music (among them segment boundaries). The IR measure has been used to detect different types of structurally important points. For example, in [41] the time varying behaviour of the IR was compared to human emotional judgements when listening to music, and results suggested the IR's behaviour is related to emotional fluctuations over the course of a musical composition. Also, in [40] the IR measure was used to detect points of 'interestingness'. Finally, and perhaps most importantly for this survey, an extension of the IR measure proposed in [1] was used to determine phrase and section boundaries in minimalist musical compositions. `Description`: For this survey we focus on the description of the model used to proces symbolic input. The model takes as input a sequence of chromatic pitch values. First, the sequence is divided using windows of 40 notes (roughly three seconds), with an overlap of 30 notes between successive windows. Second, 'marginal' and a 'conditional' entropies are estimated for each window. Marginal entropies are estimated using a 0^{th}-order Markov model. Conditional entropies are estimated using a low-order Markov model. Third, IR measure is computed as the difference between marginal and conditional entropy estimates. `Discussion`: The model of symbolic sequence prediction is rudimentary, making the entropy estimates unreliable. Also, the ability of the model to predict segment boundaries was not explicitly investigated nor tested.

Continued on next page...

Table 15 continued from previous page

Acronym	Year		Description
JSD4S	2007	[130, 99]	`Purpose`: In [130] a model originally proposed by [52] to segment biological sequences is used to predict form-level boundaries in a polyphonic piece [130] In [99] the model was extended to better deal with musical data, and used to predict melodic phrase boundaries [99]. `Description`: To identify segments, the model iteratively searches for segment boundaries using a multi-branched recursion strategy. The recursive strategy employs the following algorithm: (1) scan the complete symbolic sequence in search for the boundary that bisects it into two segments of maximal contrast; (2) iterate (1) over each resulting segment until a halting criterion is met. To quantify contrast between adjacent segments, the *Jensen-Shannon divergence* is used. `Discussion`: In [99], the extensions to the original model where tested by using the model to predict phrase boundaries in a corpus of 100 folk melodies. The results show that the extensions are able to outperform the original formulation of the method. However, its performance in respect to other phrase and form level segmentation models is unknown.
PIR4S	2009	[1]	`Purpose`: The focus of the paper is to introduce a new information theory measure: the *predictive information rate* (PIR). A simple probabilistic modelling stategy is used to test the ability of the measure to aid automatic analysis in two scenarios: music segmentation and music generation. In respect to music segmentation, the measure was used to locate form level segment boundaries in two pieces of minimalist music. `Description`: The publication first presents a description and motivation for the application of information theory measures to the analysis of temporal structure in music. The description is complete and can serve as a theoretical basis for similar approaches reviewed in this appendix (E4SS, JSD4S, IDyOM, and IR4S). Abdallah defines *predictive information* (PI) in a musical context as the information a musical event carries with it about the yet unheard future, given what has already been listened to until that point. The PI is formalised as the Kullback-Leibler divergence between the predictive distribution over 'the future' before and after a musical event considered to represent 'the present' is heard. Averaging over 'all possible presents' and 'all possible pasts' results in the PIR. Two monophonic pieces of composer Philip Glass are used to analyse what type of temporal structure might the PIR reveal. The monophonies where described only as a sequence of chromatic pitch values. A simple 0^{th}-order Markov model is used to determine continuation distributions for each event in the sequences. The PI and three other variants (among them the PIR) where used to compute 'information profiles' for the two pieces. Similarly, the classic *information content* and three variants (entropy among them) where used to compute profiles for the two pieces.

Continued on next page...

Table 15 continued from previous page

Acronym	Year	Description
		Also, the LBDM model [14] and the quantification of the GTTM GPR$_{3a}$ proposed in [46] where used to compute boundary strength profiles of the pieces. Evaluation of segmentation is qualitative, corresponding to an analysis of the observable interrelationships between the information profiles, the boundary strength profiles produced by LBDM and GPR$_{3a}$, and the form level structure of the pieces (annotated by one human expert). The analysis shows that visually salient aspects of the information profiles, such as points of discontinuity or long increasing/decreasing trends are somewhat aligned to form-level segments boundaries. The analysis also shows a better fit of the information profiles than that of the boundary strength profiles. `Discussion`: The publication emphasised the 'proof of concept' status of the experimental section, as both the probabilistic model used to obtain continuation distributions as well as the number and type of music pieces used to test the model are very restricted. The extension of PI related measures to more sophisticated predictive models is not straightforward, this hinders their applicability for music segmentation in the short term.

Continued on next page...

Table 15 continued from previous page

Acronym	Year		Description
E4SS	2010	[30]	`Purpose`: Use a neural network based model to detect points of (tonal) tension and segment structure at the level of form in a polyphonic piece. `Description`: The premise in this investigation is that segment structure and 'high level' temporal structure in music is observable in an entropy profile. (Here by 'high level' temporal structure we mean time instants that mark moments of high semantic content, e.g. moments that can be described as 'triumphant', 'dramatic', or 'shocking'. Below we refer to this type of structure as 'dramatic' structure.) The main contribution of the article is that, instead of the common Markov model prediction strategy, a recurrent neural network (RNN) is used to compute continuation probabilities. Just as in IDyOM, continuation distributions are estimated as a combination of distributions computed using a short-term model (STM), representing 'veridical' knowledge, and a long-term model (LTM) representing 'schematic' knowledge. The paper focuses on the analysis of a string quartet written by Haydn. The piece is represented as a set of instrumental parts, and each part is processed individually. Basic events (∼notes) comprising each part are represented in terms of chromatic pitch and onset. The STM estimates distributions based on the quartet itself, and the LTM makes estimations using a set of 6 string quartet movements also written by Haydn. Once the continuation distributions combining STM and LTM predictions are estimated, entropy is computed at each time point, resulting in a dynamic entropy profile of the piece. The obtained entropy profile is compared to a music theoretical analysis of the piece in terms of its formal and dramatic structure. The comparison is qualitative, corresponding to an analysis of the observable interrelationships between profile and annotated structure. The analysis shows that visually salient aspects of the profile, such as points of discontinuity or long increasing/decreasing trends are somewhat aligned to both dramatic points (climaxes, cadences, or tonal tension) and form-level segments boundaries. `Discussion`: The author acknowledges that: "while RNNs enable the analysis of music that is not amenable to other modelling techniques [referring to polyphonic music], they are slow to train, limited in the size of the corpus on which they can be trained, and [. . .] cannot generalize to other ensemble types."

Continued on next page...

Table 15 concluded from previous page

Acronym	Year	Description
		Moreover, close examination of the obtained entropy profile reveals that discontinuities and rising/declining trends can also be observed at points where there is no structure annotated. This suggest that many 'false positive' boundaries would be detected using a peak selection algorithm, if the profile was to be used for form-level boundary detection.

References

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

[2] S. Ahlbäck. *Melody beyond notes: A study of melody cognition.* PhD thesis, Department of Musicology anf Film Studies Institutionen för musik-och filmvetenskap, 2004.

[3] S. Ahlbäck. Melodic similarity as a determinant of melody structure. *Musicae Scientiae*, 11(1):235–280, 2007.

[4] Michael Baker. An artificial intelligence approach to musical grouping analysis. *Contemporary Music Review*, 3(1):43–68, 1989.

[5] Michael Baker. A computational approach to modeling musical grouping structure. *Contemporary Music Review*, 4(1):311–325, 1989.

[6] Frédéric Bimbot, Emmanuel Deruty, Gabriel Sargent, Emmanuel Vincent, et al. Semiotic structure labeling of music pieces: concepts, methods and annotation conventions. In *13th International Society for Music Information Retrieval Conference (ISMIR)*, 2012.

[7] Frédéric Bimbot, Emmanuel Deruty, Gabriel Sargent, Emmanuel Vincent, et al. System & contrast: a polymorphous model of the inner organization of structural segments within music pieces (original extensive version). 2012.

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

[9] Rens Bod. Probabilistic grammars for music. In *Belgian-Dutch Conference on Artificial Intelligence (BNAIC)*, 2001.

[10] Andrew Robert Brown, Toby Gifford, and Robert Davidson. Tracking levels of closure in melodies. In *12th International Conference Music Perception and Cognition (ICMPC)*. ICMPC, 2012.

[11] Michael J Bruderer. *Perception and modeling of segment boundaries in popular music.* PhD thesis, Doctoral dissertation, JF Schouten School for User-System Interaction Research, Technische Universiteit Eindhoven, Nederlands, 2008.

[12] Michael J Bruderer, Martin F Mckinney, and Armin Kohlrausch. The perception of structural boundaries in melody lines of western popular music. *Musicae Scientiae*, 13(2):273–313, 2009.

[13] E. Cambouropoulos. *Towards a General Computational Theory of Musical Structure.* PhD thesis, Faculty of Music and Department of Artificial Intelligence, University of Edinburgh, 1998.

[14] E. Cambouropoulos. The local boundary detection model (lbdm) and its application in the study of expressive timing. In *Proceedings of the International Computer Music Conference (ICMC01)*, pages 232–235, 2001.

[15] E. Cambouropoulos. Musical parallelism and melodic segmentation. *Music Perception*, 23(3):249–268, 2006.

[16] Emilios Cambouropoulos. Musical rhythm: A formal model for determining local boundaries, accents and metre in a melodic surface. *Music, gestalt, and computing*, pages 277–293, 1997.

[17] Emilios Cambouropoulos. Melodic cue abstraction, similarity, and category formation: A formal model. *Music Perception*, 18(3):347–370, 2001.

[18] Emilios Cambouropoulos and Costas Tsougras. Influence of musical similarity on melodic segmentation: Representations and algorithms. In *Proceedings of the International Conference on Sound and Music Computing (SMC)*, 2004.

[19] Lelio Camilleri, Francesco Carreras, and Chiara Duranti. An expert system prototype for the study of musical segmentation. *Journal of New Music Research*, 19(2-3):147–154, 1990.

[20] Chuan-Wang Chang, Wen-Jie Ke, and Hewijin Christin Jiau. A heuristic approach for music segmentation. In *Innovative Computing, Information and Control, 2007. ICICIC'07. Second International Conference on*, pages 228–228. IEEE, 2007.

[21] H.C. Chen, C.H. Lin, and A.L.P. Chen. Music segmentation by rhythmic features and melodic shapes. In *IEEE International Conference on Multimedia and Expo (ICME'04)*, volume 3, pages 1643–1646. IEEE, 2004.

[22] Elaine Chew. Modeling tonality: Applications to music cognition. In *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society*, pages 206–211, 2001.

[23] Elaine Chew. Slicing it all ways: Mathematical models for tonal induction, approximation, and segmentation using the spiral array. *INFORMS Journal on Computing*, 18(3):305–320, 2006.

[24] Seung-Seok Choi, Sung-Hyuk Cha, and C Tappert. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48, 2010.

[25] Noam Chomsky. *Syntactic Structures*, volume 119. Mouton, 1957.

[26] Noam Chomsky. *Aspects of the Theory of Syntax*, volume 119. The MIT press, 1965.

[27] Eric F Clarke and Carol L Krumhansl. Perceiving musical time. *Music Perception*, pages 213–251, 1990.

[28] Foote J. Cooper, M. Summarizing popular music via structural similarity analysis. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 127–130. 2003.

[29] D. Cope. Computer modeling of musical intelligence in emi. *Computer Music Journal*, pages 69–83, 1992.

[30] G. Cox. On the relationship between entropy and meaning in music: An exploration with recurrent neural networks.

[31] E.J. Crawley, B.E. Acker-Mills, R.E. Pastore, and S. Weil. Change detection in multi-voice music: The role of musical structure, musical training, and task demands. *Journal of Experimental Psychology: Human Perception and Performance*, 28(2):367, 2002.

[32] J. de Nooijer, F. Wiering, A. Volk, and H.J.M. Tabachneck-Schijf. Cognition-based segmentation for music information retrieval systems. *Master's thesis, Utrecht University*, 2007.

[33] I. Deliège. Grouping conditions in listening to music: An approach to lerdahl & jackendoff's grouping preference rules. *Music perception*, pages 325–359, 1987.

[34] Irène Deliège. A perceptual approach to contemporary musical forms. *Contemporary Music Review*, 4(1):213–230, 1989.

[35] Irène Deliège, Marc Mélen, Diana Stammers, and Ian Cross. Musical schemata in real-time listening to a piece of music. *Music Perception*, pages 117–159, 1996.

[36] Diana Deutsch. Grouping mechanisms in music. *The psychology of music*, 2:299–348, 1999.

[37] Nelson M. Downie, S. Evaluation of a simple and effective music information retrieval method. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 73–80, 2000.

[38] S. Dubnov. Analysis of musical structure in audio and midi using information rate. In *Proceedings of International Computer Music Conference, ICMC*, 2006.

[39] Shlomo Dubnov. Spectral anticipations. *Computer Music Journal*, 30(2):63–83, 2006.

[40] Shlomo Dubnov. Unified view of prediction and repetition structure in audio signals with application to interest point detection. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(2):327–337, 2008.

[41] Shlomo Dubnov, Stephen McAdams, and Roger Reynolds. Structural and affective aspects of music from statistical audio signal analysis. *Journal of the American Society for Information Science and Technology*, 57(11):1526–1536, 2006.

[42] Tuomas Eerola and Petri Toiviainen. Mir in matlab: The midi toolbox. In *Proceedings of the International Conference on Music Information Retrieval*, pages 22–27, 2004.

[43] M. Ferrand, P. Nelson, and G. Wiggins. Memory and melodic density: A model for melody segmentation. In *Proceedings of the XIV Colloquium on Musical Informatics (XIV CIM 2003)*, pages 95–98, 2003.

[44] Miguel Ferrand, Peter Nelson, and Geraint Wiggins. Unsupervised learning of melodic segmentation: A memory-based approach. In *Proceedings of the 5th Triennial ESCOM Conference*, pages 141–144, 2003.

[45] Jonathan Foote. Visualizing music and audio using self-similarity. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 77–80. ACM, 1999.

[46] B.W. Frankland, S. McAdams, and A.J. Cohen. 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, 2004.

[47] A. Friberg, R. Bresin, L. Frydén, and J. Sundberg. Musical punctuation on the microlevel: Automatic identification and performance of small melodic units. *Journal of New Music Research*, 27(3):271–292, 1998.

[48] Anders Friberg. A quantitative rule system for musical performance. *KTH, Stockholm*, 1995.

[49] Hiromasa Fujihara, Masataka Goto, Jun Ogata, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno. Automatic synchronization between lyrics and music cd recordings based on viterbi alignment of segregated vocal signals. In *Multimedia, 2006. ISM'06. Eighth IEEE International Symposium on*, pages 257–264. IEEE, 2006.

[50] Rolf Inge Godøy. Chunking sound for musical analysis. In *Computer Music Modeling and Retrieval. Genesis of Meaning in Sound and Music*, pages 67–80. Springer, 2009.

[51] Masataka Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1783–1794, 2006.

[52] I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J. Oliver, and H.E. Stanley. Analysis of symbolic sequences using the jensen-shannon divergence. *Physical Review E*, 65(4):041905, 2002.

[53] M. Hamanaka, K. Hirata, and S. Tojo. Automatic generation of grouping structure based on the gttm. In *Proceedings of the International Computer Music Conference (ICMC)*, 2004.

[54] M. Hamanaka, K. Hirata, and S. Tojo. Atta: Automatic time-span tree analyzer based on extended gttm. In *Proceedings of the Sixth International Conference on Music Information Retrieval, ISMIR*, pages 358–365, 2005.

[55] Masatoshi Hamanaka, Keiji Hirata, and Satoshi Tojo. Implementing a generative theory of tonal music. *Journal of New Music Research*, 35(4):249–277, 2006.

[56] Dora A Hanninen. Orientations, criteria, segments: A general theory of segmentation for music analysis. *Journal of Music Theory*, pages 345–433, 2001.

[57] S. Harford. Automatic segmentation, learning and retrieval of melodies using a self-organizing neural network. In *Proceedings of International Conference on Music Information Retrieval, MD, Baltimore*, 2003.

[58] Steven Harford. *Content-Based Retrieval of Melodies using Artificial Neural Networks*. PhD thesis, School of Computing, Dublin City University, 2006.

[59] D. Huron. *Sweet anticipation: Music and the psychology of expectation*. MIT press, 2006.

[60] David Huron. The melodic arch in western folksongs. *Computing in Musicology*, 10:3–23, 1996.

[61] David Huron. What is a musical feature? fortes analysis of brahmss opus 51, no. 1, revisited. *Music Theory Online*, 7(4), 2001.

[62] E. Isaacson. What you see is what you get: On visualizing music. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 389–395, 2005.

[63] Z. Juhász. Segmentation of hungarian folk songs using an entropy-based learning system. *Journal of New Music Research*, 33(1):5–15, 2004.

[64] T.C. Justus and J.J. Bharucha. Modularity in musical processing: The automaticity of harmonic priming. *Journal of Experimental Psychology: Human Perception and Performance*, 27(4):1000, 2001.

[65] Edward W Large, Caroline Palmer, and Jordan B Pollack. Reduced memory representations for music. *Cognitive Science*, 19(1):53–93, 1995.

[66] O. Lartillot. Reflections towards a generative theory of musical parallelism. *Musicae scientiae Discussion Forum*, 5:195–229, 2010.

[67] Cremer M. Lee, K. Segmentation-based lyrics-audio alignment using dynamic programming. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, pages 395–400, 2008.

[68] David S Lefkowitz and Kristin Taavola. Segmentation in music: generalizing a piece-sensitive approach. *Journal of Music Theory*, 44(1):171–229, 2000.

[69] F. Lerdahl and R. Jackendoff. *A generative theory of tonal music*. MIT press, 1983.

[70] Micheline Lesaffre, Marc Leman, Koen Tanghe, Bernard De Baets, Hans De Meyer, and Jean-Pierre Martens. User-dependent taxonomy of musical features as a conceptual framework for musical audio-mining technology. In *Proceedings of the Stockholm Music Acoustics Conference*, pages 635–638, 2003.

[71] S. Madsen and G. Widmer. Evolutionary search for musical parallelism. *Applications of Evolutionary Computing*, pages 488–497, 2005.

[72] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.

[73] Arpi Mardirossian and Elaine Chew. Skefis–a symbolic (midi) key finding system. *1st Annual Music Information Retrieval Evaluation eXchange, ISMIR*, 2005.

[74] Elizabeth Hellmuth Margulis. Musical repetition detection across multiple exposures. *Music Perception: An Interdisciplinary Journal*, 29(4):377–385, 2012.

[75] M. Melucci and N. Orio. Musical information retrieval using melodic surface. In *Proceedings of the fourth ACM conference on Digital libraries*, pages 152–160. ACM, 1999.

[76] M. Melucci and N. Orio. A comparison of manual and automatic melody segmentation. In *Proceedings of the International Conference on Music Information Retrieval*, pages 7–14, 2002.

[77] L.B. Meyer. *Emotion and meaning in music.* University of Chicago Press, 1956.

[78] L.B. Meyer. Meaning in music and information theory. *Journal of Aesthetics and Art Criticism*, pages 412–424, 1957.

[79] Marvin Minsky. Frame-system theory. *Thinking: Readings in cognitive science*, pages 355–376, 1977.

[80] Daniel Müllensiefen, Geraint Wiggins, and David Lewis. High-level feature descriptors and corpusbased musicology: Techniques for modelling music cognition. *Systematic and Comparative Musicology: Concepts, Methods, Findings*, 24:133–155, 2008.

[81] E. Narmur. *The Analysis and Cognition of Basic Melodic Structures: The Implication–realisation Model.* University of Chicago Press, 1990.

[82] E. Narmur. *The Analysis and Cognition of Melodic Complexity: The Implication–realisation Model.* University of Chicago Press, 1992.

[83] N. Orio and G. Neve. Experiments on segmentation techniques for music documents indexing. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 104–107, 2005.

[84] Jouni Paulus, Meinard Müller, and Anssi Klapuri. State of the art report: Audio-based music structure analysis. In *Proceedings of the 11th international society for music information retrieval conference*, pages 625–36, 2010.

[85] M. Pearce. *The construction and evaluation of statistical models of melodic structure in music perception and composition.* PhD thesis, Department of Computing, City University, 2005.

[86] M. Pearce, D. Conklin, and G. Wiggins. Methods for combining statistical models of music. In *Computer Music Modeling and Retrieval*, volume 3310 of *Lecture Notes in Computer Science (LNCS)*, pages 295–312. Springer, 2005.

[87] M. Pearce, D. Müllensiefen, and G. Wiggins. A comparison of statistical and rule-based models of melodic segmentation. In *Proceedings of the Ninth International Conference on Music Information Retrieval*, pages 89–94, 2008.

[88] M. Pearce, D. Müllensiefen, and G. Wiggins. Melodic grouping in music information retrieval: New methods and applications. *Advances in music information retrieval*, pages 364–388, 2010.

[89] M. Pearce, D. Müllensiefen, and G. Wiggins. The role of expectation and probabilistic learning in auditory boundary perception: a model comparison. *Perception*, 39(10):1365, 2010.

[90] M. Pearce and G. Wiggins. Expectation in melody: The influence of context and learning. *Music Perception*, 23(5):377–405, 2006.

[91] Marcus T Pearce and Geraint A Wiggins. The information dynamics of melodic boundary detection. In *Proceedings of the Ninth International Conference on Music Perception and Cognition*, pages 860–865, 2006.

[92] Geoffroy Peeters. Deriving musical structures from signal analysis for music audio summary generation:sequence and state approach. In *Computer Music Modeling and Retrieval*, pages 143–166. Springer, 2004.

[93] Silvia Pfeiffer and Uma Srinivasan. Scene determination using auditive segmentation models of edited video. In *Media Computing*, pages 105–129. Springer, 2002.

[94] Keith Potter, Geraint A Wiggins, and Marcus T Pearce. An objective basis for music theory: Information-dynamic analysis of minimalist music.

[95] B. Rafael, S. Oertl, M. Affenzeller, and S. Wagner. Using heuristic optimization for segmentation of symbolic music. *Computer Aided Systems Theory-EUROCAST 2009*, pages 641–648, 2009.

[96] Brigitte Rafael, Michael Affenzeller, and Stefan Wagner. An adaption of the schema theorem to various crossover and mutation operators for a music segmentation problem. In *Proceedings of the Fourteenth International Conference on Genetic and Evolutionary Computation Conference Companion*, GECCO Companion '12, pages 469–476. ACM, 2012.

[97] Brigitte Rafael and Stefan M. Oertl. Mtssm - a framework for multi-track segmentation of symbolic music. 4(1):133 – 140, 2010.

[98] C. Roads. *Microsound*. MIT Press, 2001.

[99] Marcelo Rodríguez-López and Anja Volk. Melodic segmentation using the jensen-shannon divergence. In *11th International Conference on Machine Learning and Applications (ICMLA)*, volume 2, pages 351–356, 2012.

[100] Marcelo Rodríguez-López and Anja Volk. Symbolic segmentation: A corpus-based analysis of melodic phrases. In *10th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 381–388, 2013.

[101] William Nathan Rothstein. *Phrase rhythm in tonal music*. Schirmer Books New York, 1989.

[102] R. Rowe. Machine listening and composing with cypher. *Computer Music Journal*, pages 43–63, 1992.

[103] Steve Rubin, Floraine Berthouzoz, Gautham J Mysore, Wilmot Li, and Maneesh Agrawala. Content-based tools for editing audio stories. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 113–122. ACM, 2013.

[104] Frank A Russo and William Forde Thompson. An interval size illusion: The influence of timbre on the perceived size of melodic intervals. *Perception & psychophysics*, 67(4):559–568, 2005.

[105] Frank A Russo and William Forde Thompson. The subjective size of melodic intervals over a two-octave range. *Psychonomic bulletin & review*, 12(6):1068–1075, 2005.

[106] R.S. Schaefer, J.M.J. Murre, and R. Bod. Limits to universality in segmentation of simple melodies. 2004.

[107] J.A. Sloboda and D.H.H. Parker. Immediate recall of melodies. *Musical structure and cognition*, pages 143–167, 1985.

[108] Jordan BL Smith, Ching-Hua Chuan, and Elaine Chew. Audio properties of perceived boundaries in music. *Special issue on music data mining, IEEE Transactions on Multimedia*, 2013.

[109] Christian Spevak, Belinda Thom, and Karin Höthker. Evaluating melodic segmentation. In *Music and Artificial Intelligence*, pages 168–182. Springer, 2002.

[110] Neta Spiro. Footprints of musical phrase structure in listeners' responses. In *Proc. Of the Ninth International Conference on Music Perception and Cognition*, pages 1176–1183, 2006.

[111] Neta Spiro. *What contributes to the perception of musical phrases in western classical music?* PhD thesis, Institute for Logic, Language and Computation, Universiteit van Amsterdam, 2007.

[112] Neta Spiro and B Klebanov. A new method for assessing consistency of real-time identification of phrase-parts and its initial application. In *Proceedings of the Ninth International Conference on Music Perception and Cognition*, 2006.

[113] L. Stein. *Anthology of musical forms*. Summy-Birchard Company, 1962.

[114] L. Stein. *Structure and Style: The Study and Analysis of Musical Forms*. Summy-Birchard Company, 1979.

[115] Johan Sundberg, editor. *Gluing Tones: Grouping in Music, Composition, Performance, and Listening*. 72. KMA Seminars, Royal Swedish Academy of Music, 1993.

[116] Kailash Swaminathan and Venkatesh Doddihal. 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, 2007.

[117] A. Takasu, T. Yanase, T. Kanazawa, and J. Adachi. Music structure analysis and its application to theme phrase extraction. *Research and Advanced Technology for Digital Libraries*, pages 854–854, 1999.

[118] D. Temperley. *The cognition of basic musical structures*. MIT press, 2004.

[119] David Temperley. End-accented phrases: An analytical exploration. *Journal of Music Theory*, 47(1):125–154, 2003.

[120] J. Tenney and L. Polansky. Temporal gestalt perception in music. *Journal of Music Theory*, 24(2):205–241, 1980.

[121] B. Thom, C. Spevak, and K. Höthker. Melodic segmentation: Evaluating the performance of algorithms and musical experts. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 65–72, 2002.

[122] Chris Thornton. Generation of folk song melodies using bayes transforms. *Journal of New Music Research*, 40(4):293–312, 2011.

[123] Chris Thornton. Generation of folk song melodies using bayes transforms. *Journal of New Music Research*, 40(4):293–312, 2011.

[124] T. Weyde. Optimising parameter weights in models for melodic segmentation. In *Proceedings of the 5th Triennial ESCOM Conference International*, pages 130–133, 2003.

[125] Tillman Weyde. Integrating segmentation and similarity in melodic analysis. In *Proceedings of the International Conference on Music Perception and Cognition*, pages 240–243, 2002.

[126] Tillman Weyde, Jens Wissmann, and Kerstin Neubarth. An experiment on the role of pitch intervals in melodic segmentation. In *ISMIR*, pages 287–288, 2007.

[127] F. Wiering, J. de Nooijer, A. Volk, and H.J.M. Tabachneck-Schijf. Cognition-based segmentation for music information retrieval systems. *Journal of New Music Research*, 38(2):139–154, 2009.

[128] G. Wilder. Adaptive melodic segmentation and motivic identification. In *Proceedings of the International Computer Music Conference*, 2008.

[129] Jacek Michal Wolkowicz. *Application of Text-Based Methods of Analysis to Symbolic Music.* PhD thesis, Facoulty of Computing Sciences, Dalhousie University, 2013.

[130] D. H. Zanette. Segmentation and context of literary and musical sequences. *Complex Systems*, 17:279–293, 2007.